Articulatory Methods for Speech Production and Recognition

Charles Simon Blackburn

Trinity College Cambridge
&
Cambridge University Engineering Department

December 1996

Dissertation submitted to the University of Cambridge for the degree of Doctor of Philosophy

Summary

The past 15 years have seen dramatic improvements in the performance of computer algorithms which attempt to recognise human speech. The falling error rates achieved by the best speech recognition systems on limited tasks have recently enabled the development of a diverse range of applications which promise to have a significant impact on many aspects of society. Examples of these range from dictation systems for personal computers to automated over-the-telephone enquiry services and interactive voice-controlled computing and mobility aids for disabled users.

Engineering research into the recognition of acoustic signals has focused on the development of efficient, trainable models which are adapted to specific recognition tasks. While the acoustic signal parameterisations employed are usually chosen to crudely model the behaviour of the human auditory system, little or no use is typically made of knowledge regarding the mechanisms of speech production.

Physical and inertial constraints on the movement of articulators in the vocal tract cause variations in the acoustic realisations of sounds according to their phonetic contexts. The difficulty of accurately modelling these contextual variations in the frequency domain represents a fundamental limitation on the performance of existing recognition systems.

This dissertation describes the design and implementation of a self-organising articulatory speech production model which attempts to incorporate production-based knowledge into the recognition framework. By using an explicit time-domain articulatory model of the mechanisms of co-articulation, it is hoped to obtain a more accurate model of contextual effects in the acoustic signal, while using fewer parameters than traditional acoustically-driven approaches.

Separate articulatory and acoustic models are provided, and in each case the parameters of the models are automatically optimised over a training data set. A predictive statistically-based model of co-articulation is described, and found to yield improved articulatory modelling accuracy compared with X-ray articulatory traces. Parameterised acoustic vectors are synthesised by a set of artificial neural networks, and the resulting acoustic representations are used to re-score N-best recognition hypothesis lists produced by an HMM-based recogniser. The system is evaluated on two test databases, one including speaker-specific X-ray training data and the other acoustic data alone, and improvements in word recognition accuracy are obtained in each case.

Acknowledgements

There are a large number of people to thank who have helped to make the work described in this dissertation possible. Firstly I would like to thank my supervisor, Professor Steve Young for his continued encouragement, patience and insight throughout the time that I have been working on this project. He has a clarity of thought which is exceptional, and which has been of great benefit to me in our many discussions of proposed ideas and techniques during the evolution of his original proposal for a self-organising model of speech production.

Secondly I thank Professor John Westbury and his team at the University of Wisconsin for making available the X-ray data used in this dissertation, and for their patience in describing the details of their database.

My thanks go to Trinity College for funding both my research and travel expenses over the past three years. Thanks on this latter score are also due to the Cambridge University Engineering Department, the HTK travel fund, ATERB, and the Cambridge Philosophical Society.

I would also like to thank Mark Gales, Kate Knill, Gaafar Saleh, Steve Waterhouse, Jon Lawn, Beth Logan, Julian Odell and Dave Lovell for numerous useful discussions concerning various aspects of the models, and Carl Seymour, Patrick Gosling, Chris Dance and Andrew Gee for patiently explaining the intricacies of unix and related matters. I also thank Steve Waterhouse, Dave Lovell, Dan Kershaw, Beth Logan and Harriet Nock for their proofreading efforts.

Finally, special thanks are due to Vanessa King for making PhD life in Cambridge so enjoyable, and to my parents Ann Woolcock and Ruthven Blackburn for their unfailing support and encouragement over the years.

Declaration

The length of this dissertation including footnotes, tables and appendices is approximately 58,000 words.

This dissertation is entirely the result of my own work, and contains nothing which is the outcome of work done in collaboration.

Quod scripsi scripsi.
Pontius Pilate

Phonetician:	Knock knock.
Engineer:	Who's there ?
Phonetician:	Thørdjicksdörn.
Engineer:	Thordiscorn ?!
Phonetician:	No, no, "Thørdjicksdörn". The second alveolar plosive is ingressive. [*Grinning*] Try again.

The rôle of phonetics in engineering models of speech recognition.

Contents

List of Figures

List of Tables

Notation

General

V, C	Vowel and consonant, respectively
`/x/`, `/x.1/`	Phoneme "`x`" and phone "`x.1`", respectively
$F1$, $F2$	First and second formant frequencies, respectively
F_0	Fundamental frequency of voiced excitation
W, O	Sequence of words and acoustic observations, respectively

X-ray Articulators

UL	Upper lip parameter
LL	Lower lip parameter
T1 to *T4*	Tongue tip to tongue back parameters
MI	Mandibular incisor parameter
MM	Mandibular molar parameter

Synthetic Articulators

a_1, a_2	Synthetic parameters for tongue body shape
a_3, a_4	Synthetic parameters for tongue tip shape
a_5, a_6	Synthetic parameters for lip opening and protrusion, respectively
a_7	Synthetic parameter for velum opening

Articulatory Model

μ	Mean of a distribution or a set of samples
σ	Standard deviation of a distribution or a set of samples
σ^2	Variance of a distribution or a set of samples
σ_μ	Standard deviation of mean articulatory positions
σ_{av}	Average articulatory positional standard deviation
$c(x)$	Curvature of a function at the point x
$g(x)$	Gradient of a function at the point x
$\rho_{X,Y}$	Correlation coefficient for the random variables X and Y
α	Significance level in Student's T-test
t_0	Critical value for significance in Student's T-test
$\mathcal{C}(x)$	Cumulative distribution function
D	Maximum absolute difference between the cumulative distribution function of a set of samples and that of a hypothesised model
$P(A)$	Probability of A
$P(A\|B)$	Probability of A given B
$E(X)$	Expected value of the random variable X
$f_X(x)$	Probability density function of the random variable X
$f_{X\|Y}(x\|y)$	Conditional probability density function of the random variable X, given that the random variable $Y = y$
$f_{X,Y}(x,y)$	Joint probability density function of the random variables X and Y

Acoustic Model

y_i	Output of node i
b_i	Bias for node i
w_{ij}	Weight on the connection between nodes i and j
t_i	Target value for output layer node i
Δ_{ij}	Update value for weight w_{ij}
η^+, η^-	Parameter update increase and decrease factors, respectively
$\mathcal{F}()$	Activation function at hidden nodes
E	Summed squared output error
$\boldsymbol{x}$, $\hat{\boldsymbol{x}}$	Vector and corresponding estimated vector, respectively
$\boldsymbol{X}$, $\boldsymbol{X}^T$, $\boldsymbol{X}^{-1}$	Matrix, matrix transpose and matrix inverse, respectively

Area Function Generation

k_0	Scaling parameter for logarithmic quantisation curve
N	Total number of quantised area function values
Q_k	k^{th} quantised area function value
λ	Constant used in computation of Q_k
q_i	i^{th} quantised area vector
$A_{\max}$	Maximum area function value
$\boldsymbol{\mathcal{A}}_{\text{body}}$, $\boldsymbol{\mathcal{A}}_{\text{tip}}$	Area functions for tongue body and tip, respectively
α, β, γ	Constants used in computation of $\boldsymbol{\mathcal{A}}_{\text{tip}}$

Vocal Tract Model

ρ	Density of moist air at 37^oC
c	Speed of sound in moist air at 37^oC
α	Attenuation factor
F	Frequency in Hz
$g()$	Glottal waveform function
$p(x,t)$	Variation in sound pressure at position x and time t
$u(x,t)$	Variation in volume velocity flow at position x and time t
A_k	Cross-sectional area of the k^{th} vocal tract tube segment
r	Radius of an opening at the lips or nostrils
l	Length of a tube segment
τ	Time taken by a pressure wave to propagate through a tube segment
r_k	Real-valued reflection coefficient for a pressure wave at the boundary between A_{k+1} and A_k
Γ_k	Complex reflection coefficient for a pressure wave at the boundary between A_k and a complex load
Z_{0_k}	Characteristic acoustic impedance of the k^{th} vocal tract tube segment
Z_L	Complex acoustic load impedance
R, L	Acoustic resistance and inductance, respectively
G_k	Pressure gain across tube k
H	Transfer function of vocal tract network model
z	Z-transform variable

Codebook Inversion

E	Weighted cost function
$\boldsymbol{m}$, $\boldsymbol{a}$	MFCC and articulatory parameter vectors, respectively
N_m, N_a	Dimensions of MFCC and articulatory parameter vectors, respectively
K_m, K_a	Weighting constants for acoustic and articulatory errors, respectively

Linearised Kalman Filtering

$\hat{\boldsymbol{x}}_p$	Initial estimate of the articulatory vector $\boldsymbol{x}_p$ at the midpoint of phoneme p
$\hat{\boldsymbol{x}}'_p$	Re-estimated value of $\hat{\boldsymbol{x}}_p$ at the midpoint of phoneme p
$\boldsymbol{z}_p$	Target acoustic output vector at the midpoint of phoneme p
$\hat{P}_p$	Covariance matrix for $\hat{\boldsymbol{x}}_p$
$h_p()$	Mapping from articulatory to acoustic space for phoneme p
H_p, R_p	Jacobian and output error covariance matrices for $h_p()$, respectively

Transcription Re-scoring

$\boldsymbol{y}$, $\boldsymbol{t}$	Predicted and target spectral shape vectors, respectively
M	Number of transcriptions in an N-best list used for re-scoring
E_s, E_e	Energy-normalised and raw log spectral vector errors, respectively
K_s, K_e	Weighting constants for shape and energy errors, respectively
P_a, P_l	Acoustic and language model probabilities, respectively
G_s	Grammar scale factor
L	Number of labels in a correct transcription
S, I, D	Numbers of substituted, inserted and deleted labels, respectively

Abbreviations

ANN	Artificial Neural Network
CUED	Cambridge University Engineering Department
DARPA	Defense Advanced Research Projects Agency
DP	Dynamic Programming
EMA	Electromagnetic Articulography
EPG	Electropalatography
FFT	Fast Fourier Transform
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
IPA	International Phonetic Association
K-S	Kolmogorov-Smirnov
MFCC	Mel-Frequency Cepstral Coefficient
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
pdf	Probability Density Function
RM	Resource Management
RMS	Root Mean Square
RPROP	Resilient Back-Propagation
SPM	Speech Production Model
UW	University of Wisconsin
XRMB	X-ray Microbeam

Chapter 1

Introduction

The past 15 years have seen dramatic improvements in the performance of computer algorithms which attempt to recognise human speech. The falling error rates achieved by the best speech recognition systems on limited tasks have recently enabled the development of a diverse range of applications which promise to have a significant impact on many aspects of society. Examples of these range from dictation systems for personal computers to automated over-the-telephone enquiry services and interactive voice-controlled computing and mobility aids for disabled users.

These performance improvements have been due in part to the rapidly increasing speed of the microprocessors available to implement the algorithms. Despite the considerable advances made in hardware technology to date however, the prodigious ability of the human brain to perform complex visual and auditory pattern recognition tasks still far exceeds that achieved by machines. Furthermore, it is unlikely that increased computational power alone will be sufficient to close this gap, since computer-based pattern recognition algorithms are typically limited not only by their speed, but also by an inability to reliably discriminate between all of the patterns available in their input spaces.

Since a complete biological model of the human auditory and visual systems is yet to be developed, engineering research into the recognition of acoustic signals and visual scenes has focused on the development of efficient, trainable models which are adapted to specific recognition tasks. In the case of speech recognition for example, a state-of-the-art research system which uses such a task-specific data-driven approach to model continuous speaker-independent speech from a 64,000 word vocabulary has achieved a word recognition error rate of less than 8%. This is accomplished by restricting the syntax of the task grammar, recording in a noise-free environment and operating many times slower than real time [174].

Recent advances in the performance of systems such as this have resulted more from increases in the size, efficiency and robustness of existing paradigms than from the development of new recognition algorithms, and a great many problems remain unsolved [20, 34, 46, 111]. This dissertation describes the design and implementation of a self-organising articulatory speech production model (SPM), which attempts to address some of the fundamental limitations of existing recognisers [12, 14, 15, 16]. Specifically, by using an explicit time-domain articulatory model of the mechanisms of co-articulation, it is hoped to obtain a more accurate model of contextual effects in the acoustic signal, while using fewer parameters than traditional acoustically-driven approaches.

This chapter describes the motivation for this research by identifying some of the failings of existing speech recognition algorithms, before presenting a brief overview of the new model. It concludes by providing a structural outline of the remainder of the dissertation.

1.1 Computer Speech Recognition

The recognition of human speech by computers implies the correct association of a symbolic (usually textual) representation of an utterance with all or part of its acoustic signal (usually a digitised representation of a sound wave). This is separate from the task of speaker verification, which involves the establishment of a speaker's identity and not that of the utterance spoken, and from speech understanding, which involves the semantic interpretation of the transcription. A wide variety of task characteristics fall within the scope of this definition of recognition, including:

- Size and type of the unit recognised:
  Sub-word recognition; isolated word recognition; recognition of keywords embedded in continuous speech; word-level recognition of continuous speech.
- Speaker set:
  Single speaker ("speaker-dependent") recognition; multiple-speaker recognition; unlimited speaker ("speaker-independent") recognition.
- Acoustic signal:
  Noise-free speech; speech with stationary or time-varying noise; distorted speech; accented speech.
- Speech pattern:
  Read speech; spontaneous speech.
- Language:
  Speech from a single language or dialect; multi-lingual speech.
- Vocabulary and Grammar:
  Limited vocabulary size and constrained syntax; unconstrained speech.
- Modelling strategy:
  Rule-based systems; data-driven approaches.

Within any given application both high speed and high accuracy are usually desirable, and practical implementations typically involve a trade-off between these two characteristics. This dissertation is primarily concerned with data-driven word-level speaker-dependent recognition of continuous, low-noise read speech. The two databases used contain American English speech recorded from speakers with broadly identifiable regional accents, or "dialect bases". The databases have limited vocabulary sizes of approximately 500 and 1000 words respectively, and since the goal is to assess *acoustic* modelling accuracy no syntactic constraints are applied during recognition.

1.2 Acoustically-Driven Recognition Systems

In this section a brief introduction to several existing approaches to computer speech recognition is presented. These typically comprise very large statistical models—the CUED[1] large-vocabulary HTK system has more than ten million parameters [174]—which are trained on many hours of read speech, and which attempt to transcribe utterances by using data-driven learning to automatically identify discriminatory acoustic features[2] in the input speech signals[3].

In each case some of the perceived failings of the systems are identified, with particular emphasis on their strategies for modelling the different sounds, or *phonemes*[4], of a language in each of the contexts in which they occur. Since the production of human speech involves the co-ordinated movement of a set of articulators[5], each sound has a different acoustic realisation according to its context. This effect is known as co-articulation, and currently represents a significant performance-limiting factor in engineering models of both speech production and recognition.

1.2.1 Template Matching

Early attempts at data-driven isolated word recognition matched acoustic inputs against stored word templates using dynamic time warping [131]. In this approach, each word in the recognition vocabulary is converted to a parametric representation, for example by using temporal duration and formant frequency motion rules [84]. The resulting templates comprise sequences of acoustic feature vectors, and the template providing the best match to an utterance fragment is selected as the hypothesised transcription. This approach is therefore only suited to the recognition of speech consisting of sequences of words separated by silences.

Contextual Modelling

The advantage of such a model is that it accounts for word-internal contextual effects by definition, since the different sounds are *only* modelled in context. The drawback is that a separate, complex model must be learned for each word in the vocabulary, and in order to model the variations observed across multiple instances of each word uttered by a variety of speakers, an impractically large number of models must be maintained. In addition, no provision is made for modelling cross-word contextual effects.

---

[1] Cambridge University Engineering Department.

[2] Acoustic signal parameterisations are described in more detail in Section 4.2.1, and typically make use of a frequency-domain representation of a brief (eg. $25ms$) windowed segment or *frame* of speech, where the time step between frames (eg. $10ms$) is chosen to give an overlap between successive window functions.

[3] This acoustic model is typically also coupled with *a priori* syntactic constraints.

[4] The "phonemes" are the distinctive sounds of a language, in the sense that the substitution of one phoneme for another will distinguish between the identities of two words; "phones" are the minimal sounds which are acoustically self-consistent, such as each of the two voiced sounds in a diphthong; and "allophones" are the various realisations of sounds due to contextual or accentual differences. Descriptions of the phonemes and phones used in this dissertation are given in Section 5.3.2 and Appendix A.

[5] The movable structures in the vocal tract.

1.2.2 Hidden Markov Models

Recognition systems based on Hidden Markov models (HMMs) typically also employ a frequency-domain frame-based representation of the speech signal, but the acoustic and temporal models employed are usually defined at the sub-word level[6]. These models commonly represent phonemes, but may alternatively represent smaller sub-phonemic units, or else larger segments such as syllables.

Each model predicts the probability of observing a particular acoustic vector during the production of the lexical token concerned; for example, the model for the phonetic token `/u/` might specify the probability of observing a given amount of energy in each of a set of pre-defined frequency ranges at each point in time during the articulation of `/u/`. Models for larger units are then constructed by concatenating a series of sub-models representing their constituent parts—for example by combining the models for the sequence of phonemes which together comprise a word.

Recognition is performed by using a grammar to hypothesise word sequences corresponding to a particular utterance. The probability that a particular word-level transcription correctly represents the utterance is then computed by concatenating the appropriate acoustic models, and evaluating the probability that these models could have produced the observed acoustic signal[7]. All possible transcriptions need not be evaluated, as partial transcriptions with very low probabilities can be abandoned during the search. The most likely transcription of an utterance is therefore determined *indirectly*, by partially exploring the space of possible transcriptions, and evaluating the associated probabilities of observing the original acoustic input sequence.

The principal advantages of this approach stem less from the accuracy of the temporal and acoustic models employed, than from the efficiency of the algorithms which can be used to implement them. Specifically, the use of a set of sub-word models whose parameters are optimised using the Expectation-Maximisation algorithm [36], and the decoding of acoustic vector sequences in terms of these models using Viterbi search [167] have made HMMs easily the most popular approach to building automatic speech recognition systems.

Contextual Modelling

One of the drawbacks of HMMs in their basic form is that they provide rather poor contextual modelling. To begin with, each acoustic observation vector is treated not only as having a frequency-domain representation which is stationary throughout the duration of the frame, but which is also statistically independent of all other frames—an assumption which is clearly invalid.

A technique commonly employed to lessen the impact of this independence assumption is to augment the acoustic feature vector with additional parameters which crudely represent the time-varying nature of the acoustic signal in the frequency domain. This is

---

[6] In this section only a very brief introduction to a limited class of HMM speech recognition systems is given. A more detailed introduction to HMM theory can be found elsewhere [67, 130].

[7] The *a priori* probability of the transcription as computed by a language model is typically combined with this acoustic probability score.

achieved through the use of "delta" and "acceleration" parameters to capture the short-term acoustic context, by computing the differences between neighbouring static feature vectors and the differences between the resulting difference parameters, respectively.

Since the units modelled in HMM systems are typically phonemes, the *phonetic* context in which a model appears will also generally have a strong influence on the observed acoustic vectors. The duration of these effects will be both phoneme and context-dependent, since the durations of individual phonemes are variable, and the influence of the articulatory configuration required for the production of one phoneme may also spread further than the immediately neighbouring phonemes.

Contextual effects such as these cannot be directly modelled within the standard HMM framework; a less elegant indirect approach is therefore employed in which multiple context-dependent models are provided for each phoneme. In a "triphone" system for example, a single model for a phoneme such as `/u/` is replaced with separate models specifying particular left and right contexts, eg. `/b/-/u/+/t/`, denoting `/u/` preceded by `/b/` and followed by `/t/`[8]. This approach suffers from the additional drawback that it immediately leads to an explosion in the number of models to be trained, since the number of possible triphones is extremely large. As a result, an impractically large amount of training data would be required to adequately estimate the parameters of all these models. Sophisticated clustering techniques must therefore be used to ensure that sufficient training examples are available for each clustered contextual class; hence the number of distinct context-dependent models is typically considerably less than the true number of contexts observed [118, 177].

An alternative approach is that of "segmental" HMMs, which attempt to provide more accurate modelling of acoustic variability in segments of speech ranging in length from phonemes to words. Each segmental model is typically composed of a larger number of static sub-models than are the phonetic models of a standard HMM system. These models may be fixed-length probabilistic representations of phoneme-length segments [39, 120, 144], or else variable length models[9] [17, 18, 28, 35, 40, 54, 92, 104]. In each case, the algorithms attempt to model the time-varying dynamics of acoustic parameters in the frequency domain. A frame-based acoustic representation is typically used, and parameter trajectories are modelled using dynamic programming or linear dynamic systems. While promising results have been obtained using systems such as these, the development of algorithms for the identification and optimisation of a suitable set of segmental models remain significant problems, and these approaches have not yet matched the recognition performance of the best conventional HMM systems on large tasks.

1.2.3 Hybrid Recognisers

Hybrid speech recognition systems typically use some form of artificial neural network (ANN) in conjunction with an HMM in the hope of being able to take advantage of the

[8] Similar systems have been proposed using simpler "biphones" (models incorporating either left or right context), or more complex "quinphones" (which specify a five-phoneme context).

[9] Alternatively, the segmentation can be performed at multiple levels [180].

benefits offered by each of these schemes[10][21, 22, 141]. The principal advantages of using ANNs in this context are that they provide a relatively fast method for estimating classification probabilities, and a natural basis for the discriminatory estimation of parameter values[11].

ANNs are typically used to directly estimate the probability that a given acoustic observation corresponds to the production of each of a set of sound units, which are usually phonemes. While HMMs use parametric distributions to model the probability of observing a particular acoustic vector value during production of a given phoneme, ANNs are used to discriminatively predict *phoneme* probabilities given an observation, without placing strong prior assumptions on the form of the probability distribution being modelled [27].

Contextual Modelling

The short-term acoustic context can be modelled directly in ANN systems by including a number of additional speech frames either side of the frame being classified in the input vector presented to the network[12]. In practical implementations, the length of the input window used is typically similar to the number of frames used to compute the delta and acceleration parameters in HMM systems.

The provision of a model for longer-term effects due to the phonetic context is more difficult however, since the durations of these effects are variable. Whereas in HMM systems separate context-dependent models can be used for each lexical token as an approximate model of these contextual effects, ANN systems predict lexical token probabilities directly from the acoustic signal and such an approach is not immediately feasible. One solution to this difficulty is to use time-delay neural networks, in which a feed-forward structure approximates a recurrent network[13] over a finite time period. The network's weights can then be tied to make the network mapping time-shift invariant, and hence allow for the lengthening or shortening of the sounds presented at the input [168].

Robinson has approached this problem by directly implementing a recurrent network architecture [141], in which the network learns both a set of output probabilities and a mapping to an internal state representation. These state values implicitly encode information about the previous acoustic vector inputs, and are fed back as additional inputs to the network along with subsequent speech frames. This system is therefore capable of modelling longer-term acoustic contexts, but the effectiveness of the approach is limited by the time constants of the internal state vector values, which control the duration of the influence of these contextual effects. Kershaw has therefore proposed augmenting such a recurrent ANN with a set of context-dependent single layer perceptrons which explicitly model the phonetic context. This is achieved by mapping the monophone probabilities produced by the recurrent ANN onto context-dependent phoneme class probabilities, where

[10] In fact, it is possible to construct an ANN architecture whose training emulates that of an HMM-based system [26, 80, 116].

[11] Discriminatory training techniques have also been described for HMM systems [165].

[12] Delta and acceleration parameters are also commonly used, as was the case for HMM-based systems.

[13] A network in which feed-back as well as feed-forward connections are used.

a suitable set of these latter phoneme classes are determined by a clustering technique similar to that used with HMMs [83]. This technique results in improved recognition performance, yet as in the case of HMM systems it represents an inaccurate and indirect approach to contextual modelling.

1.2.4 Spectral Pre-processing

Finally, an alternative approach to modelling contextual variation is to attempt to remove its effects in the spectral domain[14] *before* extracting acoustic feature vectors. The goal of this approach is to obviate the need for accurate contextual modelling in the recognition algorithm itself.

Systems have been described in which spectral energy peaks (such as formants) are identified, and either explicit trajectory models [89] or dynamic systems [1, 2] are used to recover hypothetical spectral targets. Deviations from these target values are then postulated as being the result of co-articulatory effects, and hence the recovered target values are used in vowel or CV[15] recognition, yielding improved results. In each case the authors do not report results for the recognition of connected speech, and the wider utility of these techniques is yet to be demonstrated.

1.3 An Articulatory Approach

A problem inherent in all of the approaches described in the preceding sections—in addition to their system-specific limitations—is that they attempt to model a time-domain effect in the frequency domain.

The underlying cause of contextual variation in the acoustic signal is the co-articulation of sounds during speech production in the time domain—variations in the articulatory movements used to produce a sound according to its context in an utterance. This suggests that contextual variation itself might be better modelled in the time domain, since:

1. The non-linear relationship between articulatory positions and acoustics means that modelling contextual variation directly in the frequency domain is difficult to accomplish.

2. An explicit model of articulatory movements would provide a compact representation of co-articulatory effects, thereby avoiding the need for the many context-sensitive acoustic models used in indirect approaches.

The system described in this dissertation attempts to address these problems by incorporating a time-domain model of co-articulation into the recognition framework[16]. By using an explicit model of contextual effects on articulatory positions, it is hoped to provide accurate articulatory and acoustic modelling, while using relatively few parameters compared with standard recognition systems.

---

[14] The time-varying frequency domain.

[15] Consonant-Vowel, eg. `/k u/`.

[16] Such a combination of phonetic knowledge with existing automatic speech processing technologies has been labelled "computational phonetics" by Moore [112].

1.3.1 Articulatory System Overview

Descriptive models of articulatory movements have existed in the literature for many years, and have been increasing in sophistication as knowledge of the articulatory mechanism increases. More recently, the potential application of articulatory models to speech recognition, together with greatly improved articulatory data acquisition and computer modelling techniques, have led to renewed interest in *predictive* models of articulatory movement [62, 90, 159, 173], and articulatory speech synthesis systems [108, 145, 156].

A model of articulatory speech production can be introduced into a recognition system either by incorporating it into the existing algorithm, or else by using the production model as a secondary recognition phase. In the former case, researchers have typically employed an articulatory description in either the symbolic representation of an utterance encoded by the input state sequence of an HMM (in place of phoneme sequences), or else as the representation modelled by the output distributions (in place of acoustic vectors).

By contrast, the system described in this dissertation is an example of the latter strategy whereby the existing recognition paradigm is left intact, and the production model is used to augment its performance. A conceptual overview of the system is shown in Figure 1.1.

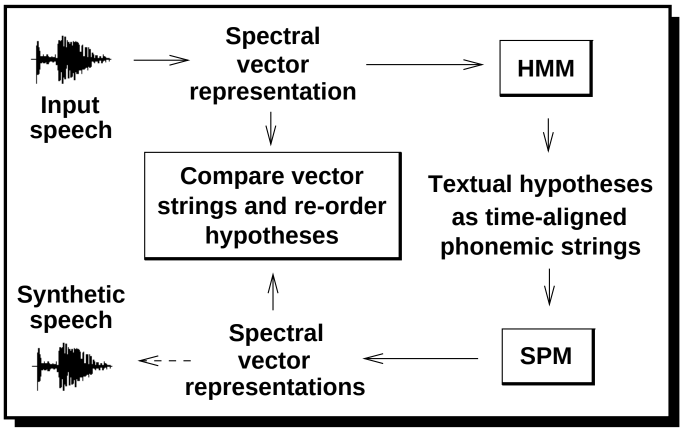

Figure 1.1: Recognition system overview. An HMM-based recognition system is used to generate a list of textual hypotheses corresponding to a parameterised input utterance. Parameterised speech vectors are re-synthesised from each transcription, and the original and re-synthesised representations are compared. The transcription corresponding to the re-synthesised vector sequence which yields the best match to that of the original speech is chosen as the system output.

As indicated in the figure, a conventional speech recogniser such as an HMM-based system provides a ranked list of hypothesised transcriptions corresponding to an utterance, along with their time-alignments to the acoustic signal. Each time-aligned transcription is then used as input to the SPM and parameterised speech is re-synthesised. By comparing the synthetic speech corresponding to each of the transcriptions with the original signal, the transcriptions are re-ranked, thus potentially offering improved recognition results.

The advantage of using an articulatory SPM in this context is that it allows an explicit time-domain representation of co-articulation to be used, since the movements of the various articulators are modelled directly. The production model is self-organising in that it learns the parameters of both an articulatory and an acoustic model from a training data set, as opposed to explicitly controlling a biomechanical model of the articulators and simulating sound wave propagation through a time-varying vocal tract. Indeed, since the primary concern is the accurate synthesis of speech spectra and not physiological plausibility, the parameters of the system may be optimised during training so that they no longer represent structures which are physiologically interpretable.

Statistics describing articulatory positions and accelerations at the midpoints of phonemes are automatically extracted from time-aligned articulatory trajectory data. The correlations between the accelerations of the articulators and their realised positions are then used to predict co-articulatory variations, and hence to generate synthetic articulatory trajectories corresponding to the time-aligned transcriptions. From these, a set of artificial neural networks predicts synthetic log spectral vector values along with their associated variances, which are compared with a spectral version of the input speech. In this way the transcription yielding the best spectral match is chosen as the most probable output[17].

The use of this co-articulation model results in a significant reduction in articulatory error by comparison with X-ray traces. In addition, the model provides a method for incorporating articulatory information into the recognition paradigm, and when combined with a conventional speech recognition system, results in a reduction in recognition error rates when evaluated on two data sets.

1.4 Dissertation Outline

This dissertation is broadly sub-divided into the following parts:

- Introduction and review of previously described speech recognition and production systems which employ articulatory representations.
- Derivation of the articulatory and acoustic models used in the new model.
- Description of the two databases to which this model has been applied.
- Evaluation of the articulatory and acoustic modelling accuracies achieved, together with the recognition performance obtained.
- Discussion of these results.

In Chapter 2, recognisers which make use of articulatory representations are reviewed, and an overview of articulatory speech synthesis techniques is presented.

Chapters 3 and 4 present the details of the articulatory speech production model. Chapter 3 concerns methods for predicting the motion of articulators from a time-aligned

[17] Transcription re-ordering algorithms have previously been described in the literature which perform re-scoring based on stochastic segmental models [119] and segmental neural networks which are specifically trained to discriminate between correct and incorrect transcription hypotheses [5, 153].

phonetic string. The co-articulation mechanism is examined, and techniques for modelling it are discussed; the methods used in the new model for both characterising and subsequently predicting articulatory variation are also described.

Chapter 4 examines mappings which can be used to predict parameterised acoustic vectors from these articulatory trajectories. Two example acoustic modelling strategies are discussed, and several self-organising approaches to approximating the mapping are presented.

The following two chapters present the data sets to which the self-organising model has been applied. An articulatory data set comprising actual articulatory traces determined from X-ray images is presented in Chapter 5. Chapter 6 provides both a review of techniques for generating synthetic articulatory data, and an example application for an acoustic speech database. In both this and the preceding chapter, techniques for automatically aligning the speech data to the corresponding phonetic transcriptions are described.

An evaluation of the production model's performance on each of these databases is presented in Chapter 7. The method for predicting articulatory variation at the midpoints of phonemes and the synthetic trajectories formed using these predictions are separately evaluated, and the results of training the artificial neural networks to map from articulatory to acoustic space are described.

Chapter 8 concerns the use of the model in augmenting an automatic speech recognition system. Both the generation of N-best transcriptions of utterances and the re-scoring of these transcriptions using the production model are described, and recognition results are presented for both data sets.

Finally, Chapter 9 provides a brief summary of the articulatory and acoustic models used in the production system, as well as a review of the recognition performance obtained. Future improvements which could be made to the system are suggested, and a concluding overview of the dissertation is provided.

Chapter 2

Articulatory Methods

2.1 Introduction

The computer speech recognition and synthesis systems which currently achieve the closest approximation to human performance are largely acoustically-driven, and typically incorporate no explicit articulatory information.

In the case of recognition systems, direct probabilistic modelling of time-varying acoustic features in the frequency domain is employed, although some syntactic-phonetic information is usually introduced through the use of statistically-based language models, and the selection of acoustic models based on their phonetic contexts [118].

Both time and frequency-domain speech synthesis techniques have been shown to produce high-quality acoustic waveforms, yet once again these methods usually rely not on articulatory modelling but on the concatenation of pre-recorded speech fragments or the direct specification of features in the frequency domain, respectively [41].

This chapter provides a review of alternative speech recognition and synthesis techniques, in which articulatory information is directly incorporated into the algorithms used. Section 2.2 describes the motivation for exploring the potential of articulatory models in automatic speech recognition systems, as well as providing a brief review of this field. Subsequently, Section 2.3 discusses techniques for implementing articulatory models of speech production, and provides a justification for the choice of a self-organising model over the use of explicit simulations of the human vocal apparatus.

2.2 Articulatory Production for Speech Recognition

The recognition of human speech by machines is achieved by computing the word sequence which is most likely to correspond to a given set of acoustic observations. Conventional speech recognition algorithms do not approach this task—as might first be thought—by directly approximating a mapping from the acoustic input space to symbolic textual representations.

HMM recognisers, for example, instead evaluate the probability that an observed acoustic vector sequence could have been produced as a realisation of a particular word-level transcription (Section 1.2.2). In mathematical terms these two quantities are related by Bayes' rule:

$$P(W|O) = \frac{P(O|W) * P(W)}{P(O)} \tag{2.1}$$

where W is the hypothesised word sequence, O is the sequence of observed acoustic vectors, $P(W)$ is the prior probability of the word sequence and $P(O)$ is that of the observation sequence. The goal of recognition is to find the word sequence W which maximises $P(W|O)$, the probability of the word sequence given the observations. The acoustic model in an HMM system computes $P(O|W)$, which is the likelihood of the observed acoustic vector sequence being produced from a hypothesised word sequence. HMMs themselves are therefore essentially specialised speech production models, which implement a form of "recognition by synthesis".

The recognition systems described in this section are similar to this in that they all employ forward models of speech production of some kind. They differ from the conventional approach however, in that they seek to incorporate *articulatory* information into the recognition algorithm, since time-domain co-articulatory effects during speech production are responsible for the contextual variations in the acoustic signal which they are attempting to model. The aim of systems such as these is therefore to improve the performance of production-based recognition systems by making explicit use of the articulatory information which characterises the production process, as opposed to inferring transcriptions from the observed acoustics alone [142].

In Sections 2.2.1 and 2.2.2, "stand-alone" systems are described, which provide not only an articulatory representation, but also a temporal model and a search algorithm, since recognition requires not only a technique for the acoustic evaluation of hypothesised transcriptions, but also an efficient method for generating likely hypotheses and their optimum time-alignments. Thus each of these systems uses either an HMM or an HMM-like state representation for temporal modelling, although the extent to which they provide explicit models of co-articulation varies considerably.

Section 2.2.3 subsequently presents systems which focus purely on articulatory production, and which therefore use a separate system[1] to provide a set of time-aligned textual hypotheses at their inputs.

2.2.1 Symbolic Articulatory State Inputs

Deng and Sun have described an approach to modelling the contextual variations in a phoneme's acoustic realisation in which utterances are represented symbolically by overlapping sequences of articulatory features[2], rather than by concatenations of lexical tokens representing (possibly context-sensitive) phonetic models [38].

Every phoneme is associated with a state-transition graph where each state represents a set of asserted or unasserted articulatory features for the context in question. This graph is implemented as an HMM chain, and recognition is performed by searching for the transcription whose articulatory feature representation is most likely to have resulted in the observed acoustic vector sequence.

[1] Typically an HMM-based recognition system.

[2] For a discussion of the temporal representation used in this model, see Section 2.3.3.

2.2.2 Probabilistic Articulatory State Outputs

Rather than decoding a sampled acoustic vector sequence in terms of articulatory labels, an explicit sampled articulatory representation can be derived and decoded in terms of phonological or lexical symbols. In this section, a number of models which take this approach are presented.

Indirect Acoustic Evaluation

Ramsay and Deng have proposed a system which uses an HMM chain whose states encode an overlapping sequence of phonological features or "gestures" in a similar manner to the articulatory features of Deng and Sun [38], but whose output distributions do not model acoustic vector probabilities but physical articulatory targets and formant frequency distributions [37, 134, 136]. Articulatory movements based on these target positions are controlled by a stochastic linear dynamic system, and the mapping to acoustic space is modelled by a series of piece-wise linear approximations to an explicit model of the vocal tract[3]. The system is then trained using Kalman filtering [9, 30] and the Expectation-Maximisation algorithm [36].

During recognition, hypothesised transcriptions of an utterance are coded as overlapping gesture sequences, and random articulatory target sequences are computed from the HMM output distributions. Smooth trajectories through articulatory space are determined by the dynamic system, and acoustic vectors are synthesised using the piece-wise linear mapping. The resulting acoustic representations for the hypothesised transcriptions are then compared with the original acoustics to determine the most likely transcription.

Articulatory-Acoustic Inversion

An alternative technique for providing an articulatory representation is to use an explicit inverse mapping from acoustic space to articulatory space, and use the resulting articulatory representation as the input signal to a recogniser[4]. Systems described in the literature taking this approach include the inversion of an articulatory-acoustic codebook to determine articulatory trajectories corresponding to an observed acoustic vector sequence[5] [137, 138], and the use of a probabilistic model to predict articulatory target features from acoustic cues [149].

The first of these systems seeks to provide an explicit time-domain description of co-articulatory effects by using a finite-state grammar of articulatory movements to perform the codebook inversion. This articulatory representation can then be used as the basis for a recognition system. By contrast, the aim of the second system is to attempt to remove the effects of co-articulation before attempting recognition, by recovering underlying

[3] Methods for modelling the human vocal tract are discussed in Section 2.3.4.

[4] Zlokarnik has shown that when actual articulatory data gathered using electromagnetic articulography (Section 2.3.2) are used to augment the acoustic signal at the input to an HMM, improved recognition performance can be obtained [179]. In a general recognition task however, such information is not readily available and must be estimated from the acoustic signal.

[5] An example of the inversion of an articulatory-acoustic mapping using this technique is presented in Section 6.4.

articulatory *targets*. Values directly estimated from the speech signal, such as formants, are used as "acoustic-articulatory" features to predict the probabilities of articulatory place and manner target variables, as well as the probabilities of transitions in place of articulation in VC and CV segments[6]. This represents an acoustic-articulatory inversion which does not attempt to predict exact articulatory movements, but rather probabilistic articulatory target feature values.

Finally, Bakis has described a third approach which explicitly incorporates both an explicit and a target-based representation [8]. Two HMMs are used, which are alternately and iteratively updated during training: one HMM controls the dynamics of co-articulation (the phonetic model), and the other predicts acoustic vector probabilities (the acoustic model). During training, the phonetic HMM takes as input a sequence of discrete target vectors in a phonetic state space. These might represent desired articulatory positions, or else some more abstract representation. The outputs describe a state space path which is chosen to minimise a cost function comprising both a phonetic error penalty—the distance between the target and actual state vectors—and an articulatory penalty, which computes the effort of moving the articulators from one state to the next.

The acoustic HMM then models the observed acoustic vector probabilities in a similar manner to a conventional HMM system, except that the input is now a continuous state-space representation rather than a discrete symbolic one. During recognition the acoustic HMM predicts smooth trajectories in articulatory space given an acoustic representation, and the phonetic HMM estimates the most probable driving phonetic target sequence which could have resulted in these smooth trajectories.

In each of the models described above the final articulatory representation is then used by a conventional recognition system to recover the most likely transcription of the utterance[7,8].

2.2.3 Combined Recognition and Synthesis Systems

An alternative approach to increasing the level of articulatory information used in the recognition process is to leave most or all of the traditional speech recognition paradigm intact, and to supplement it with production-based models or knowledge.

One unusual system along these lines uses re-synthesis of speech fragments in an attempt to increase recognition accuracy. The speech signal is first automatically segmented

[6]Vowel-Consonant and Consonant-Vowel respectively.

[7]The linear dynamic system model of Digalakis (Section 1.2.2) which uses Kalman filtering to recover underlying feature trajectories is closely related to these approaches, although it was not presented as an articulatory model.

[8]Several systems have also been described which use explicit articulatory-acoustic inversion to perform relatively simple classification tasks without attempting the recognition of continuous speech. Shirai and Kobayashi estimate articulatory parameter values using iterative optimisation, then use these for isolated vowel recognition [155]. Papcun et al. and Zacks and Thomas have described a similar vowel recognition system in which articulatory gestures are predicted from the acoustic signal using artificial neural networks trained using either standard cost functions or ones based on longer-term "shape" constraints in the output signal [122, 178]. Finally, Candille employs an articulatory-acoustic codebook to hypothesise static tract shapes for vowel sequences, and uses the corresponding predicted formant trajectories for vowel recognition [31].

into stationary and transitional regions based on changes in the frequency of the *second* formant only [121, 164]. The stationary segments are matched against stored templates to yield a list of hypothesised labels; pairs of labels are then taken in turn and corresponding speech spectra are synthesised. Based on the spectral match between the observed and synthetic transitional patterns between the stationary segments, the best hypothesised stationary label pair is chosen as the correct transcription fragment.

The system described in this dissertation is also an example of a re-synthesis technique; by contrast with the preceding system however, re-synthesis is applied to complete utterance transcriptions, as outlined in Section 1.3.1.

Finally, it is also possible to use any of the "stand-alone" systems described in Sections 2.2.1 and 2.2.2 in a combined approach. The "Waxholm" dialogue system for example, uses a conventional recognition system to provide N-best transcription hypotheses which are then re-scored using Blomberg's production model-derived templates [10, 17, 18].

2.3 Articulatory Speech Production Models

The goal of articulatory speech production is the synthesis of a continuous or sampled acoustic signal from a discrete symbolic input, via an intermediate representation in articulatory space [49]. The acoustic signal is either a raw time-domain waveform or else may be parameterised, typically in the frequency domain (Section 4.2.1). The symbolic input may be a textual representation of the utterance concerned, a (possibly overlapping) sequence of phonological labels, or a string of phonetic labels (Section 2.3.3). The potential advantages provided by an articulatory model of speech production include:

- An explicit representation of the effects of context on sound production.
- A natural integration of temporal and spatial constraints.
- An understanding of the acoustic correlates of articulatory gestures and hence their perceptual relevance.
- A framework for studying higher-level motor control mechanisms.
- A tool for the analysis of human speech defects.

This section presents a brief introduction to the articulatory mechanism, as well as a discussion of spatial and temporal models of articulatory motion. Finally, both physiological and self-organising approaches to modelling articulatory production are described.

2.3.1 The Articulatory Mechanism

The articulators are the anatomical structures involved in human speech production, as illustrated in Figure 2.1 [44, 57]. Air flow from the lungs passes through the vocal chords into the throat cavity, or pharynx, then out through the oral and/or nasal cavities.

During the production of voiced sounds such as vowels, an oscillation is set up between the vocal cords. Pressure builds up beneath the initially closed, tensed cords which forces

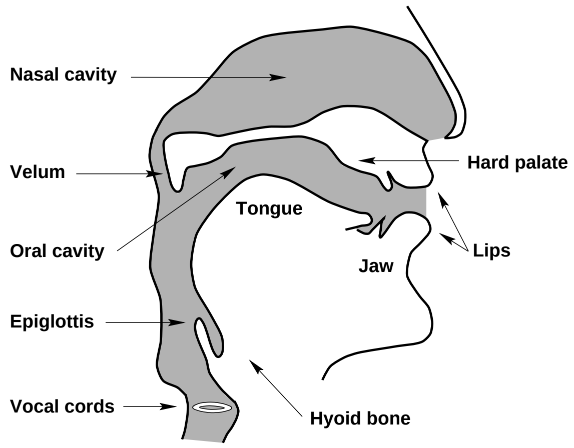

Figure 2.1: The anatomical structures used in human speech production.

them apart, leading to a flow of air through the opening, or glottis. This air flow results in a subsequent pressure reduction and the cords therefore rejoin; pressure then re-builds and the cycle is repeated, the result being a quasi-periodic pressure wave. By contrast, during unvoiced sound production the air flow through the vocal cords is unimpeded.

According to the position of the velum and the presence or absence of points of closure in the oral tract, the air flows out through the nasal cavity, the oral cavity, or both. Air flow through the oral cavity is modulated by the positions of the tongue, lips and jaw; the movements of these articulators cause deformations of the cavity which result in time-varying resonant characteristics, amplifying or attenuating the different harmonics of the excitation signal.

In addition to the excitation at the vocal cords, two other signal sources are commonly used in the vocal tract. The first of these is frication, which is produced by the turbulent flow of air through a small constriction in the vocal tract, such as between the tongue and the hard palate during the production of `/s/`. The second is a high-frequency burst, which is caused by the rapid release of air pressure built up behind a closure in the oral cavity, such as that at the lips during the production of `/p/`.

Models of speech production which mimic this human physiology can be specified in terms of the following sub-systems:

1. Definition of a spatial representation of the articulatory mechanism.

2. Conversion of a textual transcription into a temporal phonetic and/or phonological symbolic representation.

3. Control of the articulatory system to produce articulatory trajectories and hence vocal tract cavity shapes corresponding to these symbols.

4. Computation of the acoustic signal produced by propagating the appropriate excitation signals through the vocal tract cavity.

The first two model components are the topics of Sections 2.3.2 and 2.3.3, while Sections 2.3.4 and 2.3.5 present two different implementations of the latter sub-systems: explicit physiological models of the vocal tract, and statistical models which automatically extract their parameters from a training data set.

2.3.2 Spatial Articulatory Representations

The articulatory representations used when modelling human speech production fall into three categories:

- Exact specification of the position and shape of the articulators.
- Reduced-dimension parameterisations of articulatory space.
- Direct specification of vocal tract *cavity* shapes.

A complete biological model of the speech production mechanism would require an exact spatial description of the movements of the articulators. In practical systems however, it is simpler to identify a parameterised set of articulatory variables which are more easily measurable, and from which an adequate approximation to articulatory positions and shapes can be derived. Ladefoged has suggested a set of sixteen articulatory parameters which are necessary and sufficient to uniquely characterise all of the sounds in each of the known languages [91]; these are listed in Table 2.1.

Tongue front raising	Lip width
Tongue back raising	Lip protrusion
Tongue tip raising	Velic opening
Tongue tip advancing	Larynx[9] lowering
Pharynx width	Glottal aperture
Tongue bunching	Phonation tension[10]
Lateral tongue contraction	Glottal length
Lip height	Lung volume decrement

Table 2.1: Ladefoged's articulatory parameters.

Some of these parameters, such as lip width or velic opening, relate closely to the geometry of the anatomical structures from which they are derived and can be independently controlled. Others, such as the tongue parameters, refer to the overall characteristics of the vocal tract shape rather than to the position of an individual articulator. For example, raising the tip of the tongue with respect to the hard palate can be achieved by raising the jaw, raising the tongue in relation to the jaw, or a combination of these two movements. Compensatory movements such as these involving the interdependence of two or more articulators, lead to a distinction between the direct control of vocal tract shapes and the control of individual articulators.

Physiological models of speech production compute the acoustic signal radiated from the lips by solving a set of wave equations governing the propagation of an excitation

[9]The larynx is the upper part of the windpipe containing the vocal cords.

[10]The stiffness and mass of the vocal cords.

signal through the vocal tract (Appendix C). Hence they ultimately require a specification of the shape of the resonant *cavity* bounded by the articulators. This cavity shape can be specified explicitly, or else can be derived from an initial definition of (usually parameterised) articulatory positions and shapes[11].

On one side therefore are holistic models which specify vocal tract cavity shapes as goal states: an example of such a system is one in which a biomechanical model is used to explore the space of possible articulatory movements and iteratively determine an efficient strategy for the realisation of a sequence of target tract shapes [19]. Alternatively, the movements of individual parameterised articulators can be modelled explicitly, either individually or with inter-dependencies.

While the former strategy might appear to have greater biological plausibility in terms of the acquisition of articulatory control by infants [129], it is far easier to measure articulatory positions than high-level control signals. As a result, the latter strategy has formed the basis for very many more readily implementable engineering models, including that described in this dissertation [33, 45, 62, 88, 96, 101, 103, 107, 108, 145, 173].

The choice of an appropriate parameterisation of articulatory space in an explicit articulatory model represents a trade-off between the accuracy of the model and the ease of its implementation. Consider for example the articulatory parameters proposed in Table 2.1. While these parameters serve to provide a descriptive model of sound production, features such as phonation tension and pharynx width are difficult to define precisely and others such as glottal aperture and length are difficult to measure accurately during "normal" speech production.

One technique for identifying a practical set of parameters is to use a principal components analysis to identify the principal axes of movement responsible for deviations away from the neutral vocal tract position. Such a model has been proposed by Lindblom and Sundberg [96] and further developed by Maeda [101, 103].

Articulatory representations such as these are usually restricted to parameterisations of the lips, tongue, jaw and velum, coupled to simplified damped mass-spring models of the glottis or explicit models of the glottal *waveform* [45, 108, 145, 156]. Descriptions of the exact articulatory parameterisations used in this dissertation are given in Sections 5.2.3 and 6.3.1, and an example of a system for converting articulatory parameterisations into vocal tract cavity shapes is provided in the latter section.

Articulatory Data Acquisition

One of the earliest and most effective techniques used for tracking movements of the articulators is X-ray filming, or *cineradiography*. Experimental work by Houde [66] and Perkell [124] in the late 1960s provided moving X-ray pictures of subjects' heads, taken in profile while they read from set texts. The identification of soft tissue structures such as the lips and tongue can be enhanced in this approach by coating them in a barium compound, and the images thus produced can provide reasonably good estimates of tongue, jaw and lip positions. From these images, the x and y co-ordinates of key locations on

[11] Kaburagi and Honda have also proposed a system for predicting articulatory movements from vocal tract shapes [75].

the articulators can be tracked, and compared with the output of simulated models. The detailed measurement of constriction apertures and the exact dimensions of the pharyngeal cavity are usually not obtainable however, and the characteristics of the vocal fold vibrations must be measured by alternative techniques, such as high-speed video.

The greatest constraint on the use of this X-ray technique is the need to minimise the exposure of the subject's head to radiation, and only a very small amount of data can be safely recorded from any given speaker using such a system [44]. A more sophisticated technique for the collection of X-ray data greatly reduces this problem by using only a narrow computer-controlled X-ray microbeam (XRMB) to track the movements of articulatory points of interest [47]. In one such system described by Westbury, a set of 8 gold pellets are glued to points in the vocal tract, and are tracked in this manner during the synchronous recording of speech waveforms and video images [171]. This system is discussed in some detail in Chapter 5.

More recently, alternative techniques for characterising articulatory movements have been developed, such as magnetic resonance imaging (MRI), electromagnetic articulography (EMA) and electropalatography (EPG). Due to the high cost of MRI systems, only very limited use has been made of this technique, although MRI articulatory data have been described in the literature[12] [6, 175].

In an EMA system, a number of miniature coils are attached to points in the vocal tract in an analogous manner to the gold pellets used during XRMB data acquisition. The subject's head is then placed in an electromagnetic field, allowing the movement of the coils to be inferred from the corresponding induced voltages. The output of the system is a set of x and y traces for articulatory movements similar to those produced by XRMB systems [65, 125].

Finally, the EPG technique was developed as a technique for addressing the problem of discriminating between an actual contact between two articulators and the formation of a small constriction between them. Since the resolution of XRMB and EMA images is often insufficient to make this distinction, EPG systems provide a direct method for the measurement of the locations and durations of contacts between the tongue and the hard palate. This is achieved by placing a series of conducting contacts on each of these articulators and measuring the impedance between them. While palatal contact data alone are of limited use in describing the articulatory mechanism, Hardcastle has described a complementary EPG/EMA system for tracking both articulatory movements and contacts during speech production [56].

2.3.3 Temporal Symbolic Representations

Given a spatial representation of the articulatory system, a method for specifying temporal relationships during an utterance is required, to facilitate the conversion of a purely abstract textual transcription into a set of time-varying articulatory movements.

The conversion of text into a temporally-aligned representation is typically achieved

[12] Ultrasound has also been used to record images of the tongue during speech production, but this technique does not track individual flesh points, and cannot resolve hard structures in the vocal tract, such as the jaw [160].

by the use of an intermediate phonetic and/or phonological symbolic representation before articulatory movements are predicted. Regardless of the physical implementation chosen it is possible to distinguish between two different temporal strategies at this symbolic level, each of which will be discussed in this section:

- Concatenated sequences of (usually phonetic) units.
- Overlapping sequences of (usually phonological) labels.

These two techniques are often referred to in the literature as "segmental" and "non-segmental" approaches respectively [29, 97, 98, 148]. Since the term "segmental" is also used in hidden Markov modelling to refer to systems which often do not use a concatenation of conventional phonetic units (Section 1.2.2), we shall avoid using this terminology for clarity.

Concatenative Systems

Concatenative systems assume that a sequence of discrete input symbols is representable in an approximate sense at the phonological and/or phonetic level as a corresponding time-ordered sequence of elements. For example, the lexical entry "bat" might be realised phonetically as a concatenation of the individual models for `/b/`, `/ae/` and `/t/`, each having its own spatial articulatory representation.

There are two central difficulties with a model such as this. Firstly it implies the existence—at least nominally—of time-boundaries demarcating these elements, whereas speech production is a continuous process with no such hard boundaries [82]. Secondly, in its simplest form it also suggests an independence of the basic units, whereas in reality the context in which a sound occurs is known to have a strong influence on its acoustic realisation (Section 3.3).

The problem of reducing hard nominal boundaries to smooth transitions is usually addressed in articulatory synthesis by using the spatial articulatory representations derived from a series of concatenative elements as the driving functions for a dynamic system which produces smooth articulatory trajectories at its outputs, as described in Section 2.3.4.

The dependence of individual units on their context may also be approximately modelled by these system dynamics. Alternatively, these effects may be accounted for either by a secondary system which modifies the spatial representations according to their context, or else by using a large number of context-specific models for each unit, as is common in conventional speech recognition systems (Section 1.2.2) and in rule-based formant and concatenative speech synthesis systems [85].

This latter approach suffers from the requirement of very large training data sets in order to successfully model each individual unit in all of its contexts. More importantly, the fundamental advantage of using articulatory synthesis is the opportunity to explicitly model contextual effects in the time domain, so that the use of multiple context-sensitive models is somewhat counter-intuitive. The former technique—namely the use of explicit secondary models of co-articulation as an alternative to context-specific units—is discussed further in Section 3.3.

Overlaid Systems

While concatenative systems couple biologically-implausible but readily implementable representations with secondary "fixes" to render them more accurate, overlaid systems employ a more complex specification which aims to model the human production mechanism more closely throughout.

This approach typically involves an intermediate phonological representation comprising a sequence of overlapping labels or "gestures" [29, 38, 163]. These labels often span larger time intervals than single phonemes: for example the label "voiced" might be asserted during both `/b/` and `/ae/` in "bat". The advantage of this system is that it allows the description of the production mechanism to be made in terms of actual articulatory events—in this example by specifying the onset and termination of vocal cord vibration—as opposed to a description in terms of arbitrarily-defined units such as phonemes. The phonetic state of the system is then defined in terms of the set of phonological labels which are asserted at any point in time.

The YorkTalk model takes this overlaid approach one step further, and uses a directed graph rather than a sequence of phonological labels [97]. Standard tree-based parsing methods can then be used to identify the syntactic structure of an utterance. Each syntactic unit is assigned a set of phonological labels or "features", which can be shared between relevant sub-structures. This has the benefit of providing a more natural model of co-articulatory effects, since the representation of a particular part of an utterance is now *only* defined when in context.

Neither the overlaid sequential or graph-based model completely removes the need for explicit timing information, since the endpoints of the asserted labels must still be specified. These endpoints should now coincide more accurately with actual articulatory events however, and hence the problem of ensuring continuity will be reduced. Phonological labelling such as this is a complex procedure which is difficult to automate by comparison with phonetic segmentations, which can be approximately determined by the automatic alignment of phonetic transcriptions to acoustic waveforms (Sections 5.4 and 6.2). Thus while overlaid systems appear to hold great promise for articulatory modelling, a concatenative system coupled with an explicit secondary model of co-articulatory effects has been employed in this dissertation.

2.3.4 Physiological Production Models

Given a spatial description of the articulatory system and a symbolic temporal representation of an utterance, both a method for controlling the articulatory system based on this timing information, and a technique for predicting the resulting acoustic output signal are required. Physiological models of speech production achieve this latter goal by propagating a set of excitation signals through a vocal tract model defined by the articulatory specification.

Dynamic Control of Articulators

The biomechanical properties of the human speech production mechanism are extremely complex. Control of the soft and hard tissue structures of the vocal tract is achieved by the synergistic interaction of a large variety of muscle fibres. The properties of these muscles, their control signals and their inter-dependencies are very difficult to determine empirically, and hence their actions during the articulation of speech are poorly understood.

Nevertheless, even if the exact equations of motion governing the articulatory system cannot yet be derived, it is possible to make use of real articulatory data acquired as described in Section 2.3.2. By attempting to fit dynamic models to these experimental data the parameters of simplified models which approximate those governing articulatory motion can be determined.

A physiological articulatory model comprises a dynamic system and a set of forcing functions which drive it. For example, Mermelstein has proposed a model which uses exponential functions to model the closure and release of stops, and an explicit smooth function to model inter-vocalic transitions [107]; Coker's model works in a similar fashion, using a linear combination of the modes of partial differential equations or lumped-component approximations to describe articulatory movements [33].

More recently, second order dynamic systems have been used by many researchers to perform this task [128]. Critically-damped linear second-order systems have been shown to provide reasonable approximations to articulatory trajectories [147, 155], and have formed the basis for models of speech dynamics using both piece-wise parameter fitting [88] and parameter optimisation using Kalman filtering [108]. Attempts to model non-linearities in the control dynamics have included the use of recurrent artificial neural networks to predict articulatory movements [62], and the solution of simultaneous sets of non-linear second-order differential equations describing the motion of soft-tissue articulators [173].

The forcing functions used as inputs to these dynamic systems are typically articulatory target positions inferred from experimental data (Section 3.3.2). The difficulty with systems such as this however, is that it is not possible to model actual articulatory trajectories accurately using context-independent targets and fixed articulatory dynamics. This problem is illustrated in Figure 2.2, which shows simulated articulatory trajectories for lower lip height during the VC sequences `/i b/` and `/aa b/`.

In the figure the timing for both sequences is identical, but the vertical displacement of the lower lip required to achieve the labial closure in `/b/` is far greater in `/aa b/` than in `/i b/`, due to the lowering of the jaw during the production of `/aa/`. If the same target position for `/b/` and articulatory dynamics are used in both cases, and the parameters of the dynamic system are adjusted to give an appropriate transition for `/i b/` as shown, then closure will not be achieved for `/aa b/`.

Co-articulatory effects such as this may be perceptually permissible when producing voiced sounds, but in the case of stops the achievement of an occlusion in the vocal tract is essential, and thus the articulator forming the closure must reach its target position. The solution to this problem is either to provide some form of secondary non-linear trajectory warping to ensure that consonantal targets are reached [107], or else to use

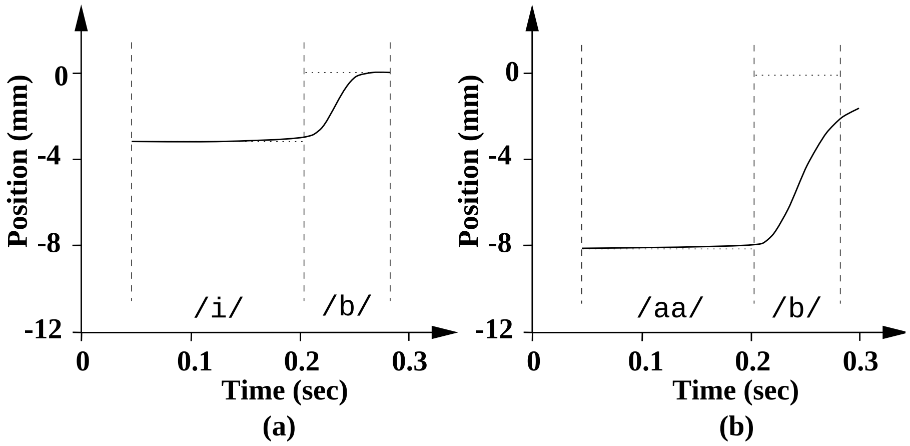

Figure 2.2: Simulated lower lip height movements during production of (a) `/i b/` and (b) `/aa b/`. Desired lower lip heights are indicated by dotted horizontal lines, and phonetic boundaries by dashed vertical lines. A positional value of $0mm$ corresponds to labial closure.

context-sensitive articulatory targets and/or dynamics[13] [33, 108].

Vocal Tract Models

The generation of an acoustic signal during human speech production is achieved by the propagation of one or more excitation signals through the vocal tract cavity. Most physiological models of the vocal tract therefore comprise a glottal model for producing voiced or quasi-periodic excitation, coupled to a model of the vocal and nasal tracts. The positions of the articulators determine both the time-varying shape of the tract and the relative lossiness of the cavity walls at varying distances from the glottis. In addition, the articulatory configuration is responsible for the noisy excitation sources created by the turbulent flow of air through any points of constriction in the tract. Voiced excitation at the glottis and/or noisy excitation at points of constriction in the tract are then propagated through the cavity and radiated as an acoustic wave at the nostrils and lips.

Models of the voicing source may either employ direct prediction of glottal waveforms, or else mechanical simulations of vocal cord vibration. Early models of glottal waveforms consisted of triangular waves or low-pass filtered impulse trains [44, 85]. Analytical forms for improved glottal pulse waveforms were later established using perceptual experiments [143], and models were developed which could produce waveforms with variable pitch period, amplitude and open quotient (the “duty cycle” of the waveform) [85]. An example of a glottal model using such a waveform is presented in Section C.5.1.

A more realistic model of the voicing source would also incorporate interactions between the quasi-periodic glottal source and the acoustic impedance of the vocal tract. This can be achieved by using a biomechanical model of the vocal cords themselves. Flanagan and Ishizaka have proposed a model for vocal-fold vibration which treats each vocal cord as a two-mass model [45, 69]. This model, or the simpler one-mass version of it, has been

[13] Context-sensitive models are discussed in further detail in Section 3.3.

used in several articulatory synthesis systems [108, 156]; more complex models involving a larger number of coupled masses have also been developed [85].

The different approaches taken to modelling the vocal tract itself include:

- Direct computation of the resonant frequencies of the specified vocal tract shape from finite difference approximations to wave equations.
- Transmission-line models of the vocal tract, in which wave propagation is described by linear and/or non-linear differential equations which are approximated by finite difference equations.
- Models of the vocal tract as a concatenation of fixed-length cylindrical tubes with variable cross-sectional area, in which wave propagation is modelled by computing reflection and transmission of the wave at section boundaries[14].

In the first approach, the shape of the tract is used to directly predict the frequencies and bandwidths of formant (or resonant) frequencies during the production of speech [33, 107, 159]. These frequency-domain models are then used to generate acoustic signals by using a formant synthesiser comprising a waveform-based voicing source coupled with a parallel set of low pass filters to produce synthetic speech spectra.

Alternatively, a time-domain approach can be used to propagate an excitation signal through the vocal tract. Early time-domain models included transmission-line analogues of the vocal tract [43, 44], which can be modelled by differential equations whose solutions are approximated by the method of finite differences [45, 102]. Such models require a great deal of computation however, and as a result many researchers have preferred to use variants of the Kelly-Lochbaum model [44, 108, 132, 145], which approximates the vocal tract by a series of uniform cross-sectional area cylindrical tubes and models the propagation of the excitation signal in terms of reflection coefficients at their junctions [81]. An example of such a vocal tract model is given in Appendix C.

Finally, Sondhi and Schroeter have proposed a model which combines a time-domain model of the glottal source similar to that of Flanagan and Ishizaka [45, 69] with a frequency-domain variant of the vocal tract approximation based on a concatenation of fixed-length cylindrical tubes [156].

2.3.5 Self-Organising Production Models

Self-organising models of speech production attempt to automatically extract appropriate values for the parameters of a set of statistical models from a set of training exemplars. The advantages provided by such an approach are that:

- The systems are automatically adaptable to new speakers, provided appropriate training data are available.
- Many of the difficulties of modelling the dynamics of the articulatory mechanism can be avoided, while still producing good approximations to physical articulatory movements.

[14] This technique is also referred to as wave digital filtering.

- The necessity for mimicking human physiology can easily be relaxed if desired.
- It is inherently possible to provide a probabilistic description of speech production which is suitable for use in the context of speech recognition systems (Section 4.2.3).

The explicit models of human speech production described in the preceding section generally do not satisfy these requirements, since procedures for automatically tuning their parameters are typically not available[15], and they produce definitive rather than probabilistic signals at their outputs.

The recognition system of Ramsay and Deng described in Section 2.2.2 is an example of an approach in which a statistical description of the production system is used as the fundamental basis for part of the model [135]. They use a stochastic HMM-based model of articulatory dynamics, in which the output distributions of the HMM states represent both articulatory positional probabilities and formant frequency distributions. The parameters of these distributions are automatically optimised over a set of articulatory and formant vector training exemplars, and are then used to decode articulatory and formant trajectories during recognition, in conjunction with an explicit linearised model of the vocal tract.

The system described in this dissertation differs from this approach in that it uses self-organising models for both the prediction of articulatory dynamics *and* for the synthesis of acoustic signals. The detailed description of these two models forms the basis of Chapters 3 and 4 respectively.

[15]Some of these explicit models do incorporate sub-systems with partially automatic data-driven optimisation procedures. For example, Meyer's model of articulatory dynamics is optimised over a training data set using Kalman filtering [108], and several systems have been described which use iterative optimisation procedures for determining vocal tract configurations [7, 19, 90, 123, 139].

Chapter 3

From Phonemes to Articulators

3.1 Introduction

This chapter describes the techniques used in the SPM to convert time-aligned phonetic strings into articulatory representations. The primary articulatory modelling task of the SPM is the provision of a *spatial* specification of articulatory movements; hence the provision of a basic description of the temporal organisation of an utterance is required *a priori*. The temporal model used in the system is a time-aligned concatenation of the phonetic units which comprise the utterance, as produced by an automatic alignment system such as HTK[1] [176].

The conversion of this symbolic sequence to a set of smooth articulatory trajectories representing the utterance is achieved by:

1. Estimation of probability distribution functions describing articulatory positions at points in time corresponding to the midpoints of phonemes.

2. Prediction of variations in these articulatory positions due to co-articulatory effects.

3. Generation of complete articulatory trajectories.

The derivation of a parametric description of variations in articulatory positions is described in Section 3.2. Section 3.3 contains a discussion of the co-articulatory mechanism and proposes a descriptive model which can also be used in the prediction of co-articulatory variation. Finally, Section 3.4 describes both a technique for implementing such a predictive model and the subsequent generation of complete articulatory trajectories.

3.2 Characterising Articulatory Variability

The spatial trajectories followed by the articulators during the production of human speech are the result of a trade-off between inertial and perceptual constraints. The former dictate that these movements be as efficient as possible, in order to minimise biomechanical effort, while the latter require that articulatory movements be sufficiently precise as to produce an acoustic signal which is intelligible to the listener.

[1] For a discussion of temporal modelling in articulatory speech production, see Section 2.3.3.

The acoustic relevance of the spatial positioning of an articulator varies considerably from sound to sound, an important observation which is discussed further in Section 3.3. All articulators show some degree of positional variability however, and this section is concerned with identifying and characterising that variability.

3.2.1 Sources of Variability

There are many sources of variability in the acoustic realisation of a sound within an utterance [126]. These include:

- Random variations due to the limited precision of muscular control of the articulators and/or limited precision in the neural control signals sent by the brain.
- Systematic variations due to the phonetic context in which the sound occurs.
- Systematic variations due to the prosodic context in which the sound occurs.
- Variations due to stationary or time-varying noise on the signal.
- Variations due to channel distortion during transmission of the signal.

The scope of this dissertation is restricted to the analysis and synthesis of low-noise high-bandwidth speech signals, and hence only the variability arising during the articulatory production of speech is of interest, and not the effects of additive noise or channel distortions.

Prosodic variation

Prosodic influences on the speech signal are typically defined over longer time intervals than the lengths of individual phonemes[2], and result from linguistic intentions such as tone, intonation and stress [170]. Variations in these three characteristics have a significant impact on the acoustic waveform, primarily affecting the pitch[3], duration and signal intensity. Standard HMM-based recognition systems however, are designed to be relatively insensitive to variations in these characteristics of the acoustic signal. For example, the acoustic parameterisations commonly used in speech processing are insensitive to variations in pitch and signal intensity. While the *relative* signal intensities at different frequencies are of great importance, changes in the overall energy content in the signal over time typically affect few if any of the acoustic parameters used. Furthermore, a feature of HMM-based systems is their use of variable-duration models to normalise the effects of temporal variations in the speech signal.

This attempt to exclude prosodic effects from the recognition process is far from coincidental, as the goal of speech recognition systems is to remove from the signal any variability which is not strongly related to the identity of the words being recognised. Prosodic information provides only weak cues as to word identities, although it is highly

[2]For this reason they are also referred to as "supersegmental" features.

[3]The fundamental frequency of oscillation of the glottis.

relevant to the *understanding* of the acoustic signal, and some researchers have proposed the use of prosodic information to distinguish between ambiguous semantic interpretations of recognised word sequences [68].

Since the durational information used by the SPM is provided by an HMM-based system, and both pitch and slowly time-varying spectral intensity features will not be used in the frequency-domain representation of the speech signal, the SPM does not incorporate a prosodic component. If the system were to be used for the synthesis of natural-sounding time-domain acoustic signals however, a technique for automatically assigning prosodic features to the transcription of an utterance to be synthesised would be required.

In terms of the observed articulatory variation therefore, prosodic influences on articulatory positions—such as increased amplitude of movement for stressed syllables—will be modelled as random variations.

Random variation

The primary source of random variation in articulatory movements is the inexact control of the muscles over multiple repetitions of the same task. Figure 3.1 shows an example of overlaid X-ray articulatory trajectories for the horizontal and vertical displacement of the tongue tip during four repetitions of the same phrase. Approximate phonetic boundaries as produced by an automatic alignment program are shown by vertical dotted lines.

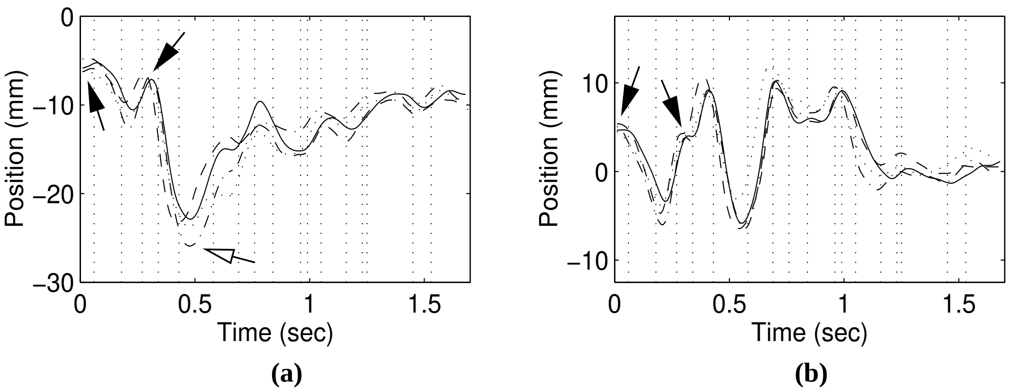

Figure 3.1: Random variations in (a) horizontal and (b) vertical position of the tongue tip taken from X-ray data during four repetitions of the phrase "The other one is too big". Solid and light arrows indicate regions of relatively low and high variability, respectively. Phonetic boundaries are indicated by dotted vertical lines.

In this example relatively little durational variation is observed, as the speaker has endeavoured to maintain a constant speaking rate. The maximum temporal displacement between the waveforms is only $\approx 0.05sec$, while there is as much as $5mm$ of positional variation. In general the vertical position of the tongue tip is more accurately reproducible than the horizontal position, although the degree of variability is time-varying during the utterance. For example, both horizontal and vertical positions are relatively tightly constrained during production of the two occurrences of the phoneme `/dh/` in "the" and "other", marked by solid arrows in the figures at the start of the utterance and after

$\approx 0.3sec$. This relative lack of variability is explained by the requirement that the tip of the tongue be placed between the teeth during production of `/dh/`, so that its position is tightly constrained.

By contrast, the horizontal position of the tongue tip during the phoneme `/w/` at the start of "one", marked by a light arrow in figure (a), shows a considerable degree of variability while still maintaining acoustic intelligibility.

Contextual Variation

Variations in achieved articulatory positions due to the phonetic context in which a sound occurs are responsible for much of the variability observed in the corresponding acoustic signals. Since these variations arise in a systematic way, it is hoped that they are predictable from the phonetic and temporal transcriptions of an utterance.

Figure 3.2 shows an example of the effects of context on the height of the *back* of the tongue, during multiple repetitions of the same phrase used in Figure 3.1, again recorded using X-ray tracking.

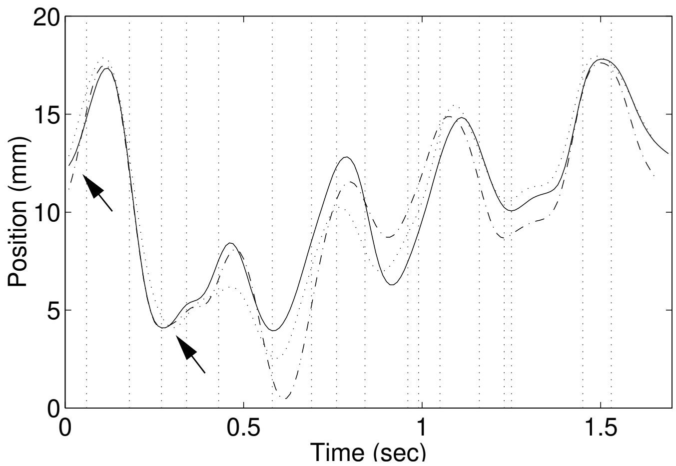

Figure 3.2: Vertical displacement of the back of the tongue taken from X-ray data during the phrase "The other one is too big". Solid arrows indicate two dissimilar but reproducible articulatory strategies used during the production of the phoneme `/dh/`. Phonetic boundaries are indicated by dotted vertical lines.

Focus again on the position of the articulator during the two occurrences of `/dh/`, at the start of the utterance and after $0.3sec$ as marked by solid arrows. As was the case in Figure 3.1, there is relatively little positional variation at these points over multiple repetitions of the phrase. By contrast with the previous figure however, a significant and consistent difference is now seen in the placement of the articulator *between* the two occurrences of `/dh/`.

This difference is explained by the context in which the sound occurs in the two examples. The initial `/dh/` (in "the") is preceded by silence and followed by `/iy/`, a vowel which requires the back of the tongue to be raised. The second `/dh/` (in "other") however, is preceded by `/ah/` and followed by `/er/`, two vowels which require a lowering of the tongue back.

Since the height of the back of the tongue can be varied during production of `/dh/` without significantly altering the acoustics, economy of movement dictates a relatively high tongue back during the initial `/dh/` in anticipation of `/iy/`, and a relatively low position for the second inter-vocalic `/dh/`. That this variation is due to contextual rather than random effects can be seen from the minimal variation observed over multiple repetitions of the utterance.

If the positional variation observed for a given articulator were characterised during the production of a particular phoneme in a large corpus of speech, both "random" variation—in this case comprising both truly random variation and prosodically-based variation—and systematic contextual variation would therefore be observed[4].

3.2.2 Positional Sampling

A relatively straightforward technique for characterising articulatory variation is to sample articulatory positions over very many examples of the production of each phoneme in a speech corpus. The samples obtained would ideally be insensitive to changes in phonetic duration, but would incorporate the effects of co-articulation due to both the right and left phonetic contexts of the phoneme. In addition, since the goal is to develop a self-organising model, this sampling process should be fully automated.

The labelling of articulatory samples according to their phonetic class implies the availability of an alignment of the acoustic signal to a phonetic transcription. The generation of these alignments is discussed in Sections 5.4 and 6.2, and involves the use of the HTK toolkit to identify initial and final time stamps demarcating each phoneme in a phonetic sequence corresponding to the utterance's transcription. These time markers bound the local regions in the speech signal which best match the HTK models for the phonemes concerned.

When modelling articulatory movements it is assumed that:

1. Each phonetic segment comprises two basic sections. In the first of these, the articulators are moving away from the positions dictated by the previous phoneme and toward those required for the phoneme in question. In the second section the articulators then begin to move to locations corresponding to the following phoneme.

2. There is a sub-section of the phonetic segment in which the articulators reach positions which are such as to produce an acoustic signal which is identifiable as the phoneme concerned. The sub-section may or may not include a region in which the positions of one or more articulators are held constant.

3. This sub-section roughly corresponds to—or else encompasses—the midpoint of the phoneme.

As an example, on one extreme are phonetic sequences such as `/s z s/` where the central phoneme is identified only by the presence of voicing. In this case the articulators are motionless throughout the production of `/z/`, so the initial and final sections of

[4] A further discussion of contextual effects on articulation ("co-articulation"), along with techniques for modelling the resulting positional variations, can be found in Section 3.3.

this phoneme are indistinguishable, and the entire phonetic segment satisfies the minimal articulatory positioning requirement.

On the other extreme are sequences such as `/ae dx ae/`, in which the tongue tip is constantly in motion. The flap `/dx/` is characterised by an initial articulatory gesture in which the tip of the tongue is moving towards the hard palate, followed by a final gesture in which it moves away from it. In between these two is a brief period around the midpoint of the phoneme when the tip briefly makes the requisite contact with the hard palate.

The requirement that the articulators satisfy positional constraints at the midpoints of phonemes is discussed further in Section 3.3, but it is observed briefly here that this is conceptually different from requiring that the articulators attain and maintain static spectral "targets". Several authors have demonstrated that only transitional information is required for the recognition of certain vowels in particular contexts, since static vowel nuclei can be deleted without impairing human recognition performance [48, 161]. This is consistent with the model presented above, since articulatory constraints at the midpoints of phonemes may be used only to define initial and final articulatory trajectories, without the need for a steady vowel nucleus region to be maintained.

The positional variability of each articulator during the production of each phoneme is therefore characterised by sampling its midpoint position as depicted in Figure 3.3. This figure shows the simulated movement of an articulator during the word "star", where phonetic boundaries are marked by solid vertical lines, and sampling points by dashed vertical lines.

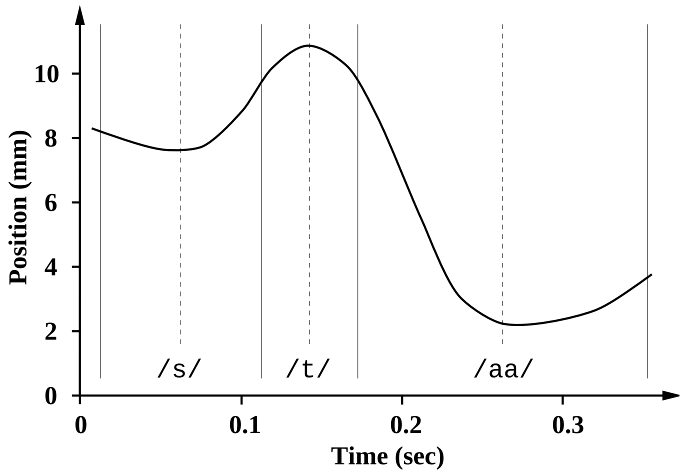

Figure 3.3: Simulated articulatory positional sampling points during the word "star". Boundaries between phonemes are indicated by solid vertical lines, and sampling points corresponding to the midpoints of phonemes are shown by dashed vertical lines.

This technique is duration-independent, and represents both an explicit and an implicit model of perceptually relevant articulatory behaviour: specific positional constraints are explicitly incorporated, while the shapes of articulatory trajectories into and out of phonemes are implicitly modelled.

3.2.3 Parametric Models

The result of this sampling process is a set of positional values for each phoneme and each articulator. The probability distribution functions of these samples are then approximated by parametric models, which efficiently encode the significant characteristics of the distributions. The articulator's positions at the midpoints of the examples of each phoneme are modelled with a single Gaussian distribution of the form:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{3.1}$$

where x is the positional value of the articulator at the midpoint of the phoneme concerned, μ is the mean of the distribution and σ^2 is its variance, which arises from both random and systematic positional variations. In making this approximation it is assumed that:

1. The distribution of articulatory positions at phonetic midpoints is unimodal.
2. In the limit of a very large number of sample points, this distribution approaches a "normal" distribution.
3. A first approximation to articulatory positions can be obtained by modelling the positional variation of each articulator independently.

The first of these assumptions implies that a single articulation strategy is used for each phoneme regardless of its context. All possible articulations will then represent different manifestations of the same basic articulatory pattern, with variations about the mean position[5] being caused by contextual effects. If a speaker maintains more than one strategy for the production of a given sound and chooses amongst these according to the phonetic context, then either a multimodal distribution must be used, or else a set of unimodal distributions maintained, and rules used for selecting between them according to the context.

The second assumption predicts the shape of the probability distribution obtained from the sample points, dictating both its rate of fall-off away from the median value, and also that this fall-off should be symmetrical about the mean position, so that the mean and median positions coincide.

When dealing with real articulatory data, a limited number of different contexts and sample points will be available, and hence the use of a "normal" distribution will by definition be approximate. In addition, the nature of the articulatory system is such that many positional distributions will not quite be symmetrical. For example, when positioning the tongue tip during the production of `/s/`, there is far greater scope for random or systematic variations due to downward rather than upward movement, due to the presence of the hard palate. This physical constraint will be manifested in the statistical data as a skew on the distribution.

Nevertheless, the single Gaussian distribution model described above serves as a useful approximation which is not only mathematically tractable, but very efficient in its use

[5]As discussed in Section 3.3 this mean value should *not* be interpreted as a "target" position.

of parameters. In practice, it proves to be a successful model of systematic articulatory variations.

Finally, each articulator's movements are modelled independently. This is a simplifying assumption which is not valid in the case of the human articulators[6], but which permits an initial approximation to articulatory movements to be relatively easily derived using the positional sampling technique[7].

3.3 Co-articulation Models

Co-articulation is the allophonic variation of a phoneme due to changes in its phonetic context [51]. Human speech does not consist of a concatenation of independent sounds, but rather of a continuous sequence of sounds which are produced by smoothly varying both the positions of the articulators and the excitation signals in the vocal tract.

Since the inertial masses of the articulators restrict their movements to finite speeds—and more importantly since humans are inclined to use the most efficient movements possible while ensuring acoustic intelligibility—the production of any particular sound is strongly influenced by the articulatory constraints imposed by its neighbours. In this section the mechanisms which give rise to co-articulatory effects are described, and two models of articulatory behaviour are presented.

3.3.1 Co-articulation Mechanisms

The study of the mechanisms of co-articulation during human speech production is an active field of research, and current models of articulatory behaviour are far from complete. This section provides a brief introduction to some of the underlying causes of these co-articulatory effects.

Causes of Co-articulation

The degree to which the phonetic context will influence the positions of the articulators during the production of a given sound varies greatly from articulator to articulator and from sound to sound. Factors which determine this variation include:

- The duration of the current sound and of its neighbours, both relative to each other and in terms of the overall speaking rate.
- The spatial separation between the desirable articulatory positions for adjacent sounds.
- The degree to which any articulatory positional constraints are perceptually relevant.

[6] An example of the correlations between such a set of articulators as measured from X-ray data is given in Section 7.2.1.

[7] The development of an articulatory model which takes into account the inter-dependence of these raw articulatory variables would represent a logical area for future research into this field, as discussed in Section 9.5.

- The physiologically-determined speed and precision with which a particular articulator's position can easily be controlled.
- The amount of articulatory effort used by the speaker.

The first two items determine the underlying geometrical framework for articulatory production. The smaller the temporal separation between the midpoints of two adjacent sounds and the larger the displacement between their desired articulatory configurations[8], the greater becomes the articulatory effort that will be required to produce them in sequence.

Once a desired articulatory position has been specified, the physiological nature of the articulator concerned strongly influences the speed and accuracy with which this position is approached. For example, the tongue tip is considerably more nimble in its movements than is the jaw, and hence is able to produce relatively rapid, accurate movements.

In "normal" conversational speech however, articulatory movements are rarely limited directly by such inertial constraints, but more often are indirectly linked to them through efficiency considerations on the part of the speaker. That this is so can be seen from the observation that human speakers are capable of clearly articulating speech at much higher rates than are typically used in conversation[9]. The movements of the articulators therefore do not in general reflect the physical limits of muscular control, but rather an economy of effort while communicating a message [95].

This does not mean that the inertial properties of the articulators are without relevance, since even if such a strategy based on minimised effort is postulated, the interrelationship between the mass of an articulator and the strength of the muscles controlling it will continue to strongly influence the co-articulation process.

Within this framework of geometrical constraints and economies of movement, the amount of co-articulation permitted *perceptually* is largely determined by the degree of specification of an articulator's position in a particular context [78]. Thus in Figures 3.1 and 3.2 in Section 3.2.1 it can be observed that during the production of `/dh/` the position of the tongue tip is highly specified, while that of the back of the tongue is underspecified—its position can be varied without greatly altering the acoustic signal. As a result, the back of the tongue is relatively free to move into a position which is most convenient for production of the sounds in either the left or the right context, or a compromise between these two. This in turn leads to an economy of movement since unnecessary displacement of the tongue back is avoided. By contrast, the position of the tip of the tongue is highly constrained during `/dh/`, since it must be in position before this phoneme can be articulated, and is not free to move to the next required position until after the relevant acoustic cues for `/dh/` have been produced.

A continuum in the degree of specification is therefore postulated, from one extreme of exact specification, through loosely specified constraints, to the other extreme of complete underspecification of articulatory positions.

---

[8] The issue of specifying desired articulatory configurations such as these is discussed in Sections 3.3.2 and 3.3.3.

[9] The 1991 Guinness Book of Records quotes the fastest achieved intelligible speech rate as 586 words per minute [106], as compared with a "standard" rate of between 100 and 200 words per minute.

Types of co-articulation

The detailed characterisation of the mechanism of co-articulation during speech production remains an active field of research. Nevertheless, it is generally agreed that there are two distinct mechanisms which give rise to co-articulatory variation:

- **Anticipatory co-articulation:** occurs from right to left and is due to timing effects, whereby the movement of an articulator toward the position dictated by the following sound commences during production of the current sound.
- **Carryover co-articulation:** occurs from left to right and is due to mechanical inertia, whereby an articulator is slow to move away from the position required by the previous sound and into that dictated by the current sound.

An example of both effects can be seen in the word "toot". When pronouncing `/t u t/`, the lip rounding gesture required for the production of `/u/` is initiated during the first `/t/`—an anticipatory effect made possible by the underspecification of lip position during production of `/t/`. This same lip underspecification then allows a further economy of movement to be achieved, as the lip rounding gesture continues during the following `/t/` via the carryover effect. The result is a rounding of the lips throughout the word, although this feature is nominally only required for the central vowel.

As is the case in this example, the largest co-articulatory movements will in general be observed for the least specified articulators—ie those which are least relevant acoustically. This does not reduce the importance of modelling co-articulatory effects however, since co-articulatory movements result from the *combination* of one sound for which an articulator's position is highly specified, with another for which it is not. A model of the articulator's behaviour during the underspecified sound is therefore required to accurately predict the trajectory that will be followed by that articulator during the transition into the acoustically relevant region.

Thus co-articulatory effects are present even for the most highly specified of articulatory configurations, since the context will determine the trajectories followed by the articulators into and out of the production of the sound concerned. The phonetic context therefore influences initial and final articulatory transitional movements to some extent for all phonemes and all articulators. The articulatory configuration reached *between* these two transitions will be dictated both by perceptual constraints (the degree of specification of an articulator) and by geometric constraints (the spatio-temporal demands placed on the articulator in relation to its ease of movement).

The presence of both anticipatory and carryover effects implies that the articulation of a sound will in general be dependent on both the preceding and following contexts. Several studies have attempted to determine whether these effects are limited to the immediately neighbouring phonemes, or whether a larger context is required. Although specific contexts have been identified in which anticipatory effects can be demonstrated over multiple phonemes, Gay has argued that in VCV utterances, both anticipatory and carryover effects do not spread further than the immediately neighbouring segments [50].

This suggests that a relatively simple, single-neighbouring-phoneme bi-directional model of co-articulation can be used as a first approximation to the co-articulation process[10].

3.3.2 Target Models

The exact mechanism of human speech production is still poorly understood. Many researchers have proposed models which attempt to describe how a discrete symbolic representation of an utterance is converted to a smooth articulatory realisation, while satisfying both perceptual and physiological constraints. A popular modelling technique is to represent an utterance by the specification of a sequence of underlying goals, and to regard articulatory movements as attempts to reach these goals while satisfying biomechanical constraints. The goals may be fixed target points, or else may specify target regions or trajectories. They may be defined in one or more of a number of different sub-spaces, hypothesised examples of which include [58]:

- **Invariant electromyographic signals:** the signals sent by the nervous system to the muscles involved in the production of a given sound are contextually invariant, and the responses of the muscles to these signals result in co-articulatory variation.
- **Articulatory targets:** articulatory movements are governed by a desire to attain articulatory targets which are maintained by both peripheral and internal feedback, the achievement of which is hampered by inertial constraints.
- **Acoustic or perceptual targets:** speech comprises a series of acoustic targets, which can be achieved (or at least approached) using a variety of articulatory movements.

These proposals differ both in the level at which invariant goals are specified (high-level motor control signals versus low-level articulatory positions), and in the degree of abstraction of the control signal (physical nerve signals and/or articulatory positions versus resultant acoustic waveform characteristics).

The disadvantages of using high-level control signals are that they are difficult to study empirically, and that even if it were possible to accurately characterise them, a detailed knowledge of the muscular system in the vocal tract would be required in order to determine their effects on the articulation of speech. By contrast, a wide variety of systems have been proposed in which some form of articulatory and/or acoustic goal sequence is used to model articulatory movements. This section provides a brief summary of several systems which take this latter approach.

Acoustic-Articulatory Targets

Several researchers have described systems which make use of either spectral energy or formant frequency targets. Akagi has attempted to extract invariant spectral targets by

[10] Experiments in large-vocabulary computer speech recognition have suggested that contextual information useful to the recognition process may also be contained in the identities of the phonemes two steps away from a given phoneme [174].

modelling spectral trajectories with dynamic systems [1, 2], and Kuwabara has proposed an explicit mapping for increasing the separability of formant frequencies for vowels, and hence reducing co-articulatory effects [89].

Techniques such as these are more applicable to modelling transitions between vowels than those involving consonants, since the identification of spectral energy trajectories is more difficult in the latter case. An alternative approach is to combine acoustic and articulatory goals, as proposed in Perkell's model of peripheral and internal feedback [127], and Stevens' system which supplements formant targets with articulatory goals to account for consonantal constrictions when controlling a formant synthesiser [159].

As discussed in Section 1.3 however, each of these systems is attempting to model a time-domain effect in the frequency domain, and a separate mapping to articulatory positions would be required for articulatory synthesis.

Articulatory Point Targets

Early articulatory models proposed specific spatial or muscular target positions for the articulators during the production of each particular sound [33, 66, 100, 107]. The movement of the articulators in response to these control signals is modelled in these systems using explicit functions of position and time. These typically comprise either single exponentials or the sum of multiple linear and exponential functions.

Later variants of this approach include the use of simple linear interpolation between target sequences [145], and the use of Kalman filtering to optimise a set of low-pass filters [172]. Shirai proposed using critically-damped second-order linear dynamic systems to approximate articulatory trajectories [155], an approach that has since been used by several researchers [88, 108].

There are several difficulties with models such as these. Firstly, it is clear that context-independent spatial targets which are invariably achieved do not exist for all articulators or all phonemes. While it is true that during `/b/` for example, lip closure is always achieved regardless of the context, articulatory positions for many other phonemes and articulators are less tightly specified, or may not be specified at all.

Even if exact spatial targets such as these did exist, if fixed dynamic properties are used to model articulatory movements independently of the context, then targets would only sometimes be reached, as demonstrated in Figure 2.2 in Section 2.3.4. A common solution is to hypothesise the existence of *virtual* targets in articulatory space which may not be physically realisable, but may instead be "undershot". In this way contextual variation would lead to a distribution of articulatory positional values, rather than a fixed spatial configuration which must be reached.

This proposal only partially accounts for the problem of contextual variation however, since there is now a difficulty with phonemes such as `/b/` which *do* require a very precise articulatory specification. As a result, these systems typically either use extreme virtual targets for phonemes such as `/b/` and constrain the lip models to stop moving once they collide, or else they apply secondary modifications to the articulatory trajectories to ensure that consonantal requirements are met [33, 107, 108, 172].

An additional problem with the use of either virtual or physical targets is the fact that quite different vocal tract shapes can be used to produce very similar acoustic output for some phonemes[11] [4, 94]. Thus multiple articulatory targets might exist for a given phoneme, with the vocal tract shape used for each individual instance being selected according to the context.

A more realistic model would therefore require the use of context-sensitive targets and/or dynamics. As discussed in Section 3.3.1, articulators typically do not move at their maximum speed, but with the minimum of effort required to ensure intelligibility. Thus since lip closure is perceptually essential during the production of `/b/`, the amount of articulatory effort used in a given context will be that which is sufficient to achieve closure *in that context.* In Meyer's system for example, more than one spectral target is used for some phonemes according to their context [108], and Kröger uses a system with fixed targets but time-varying dynamics [88]. The drawback with these approaches is that determining and specifying the context-sensitive control parameters is a difficult task, which obviates either the advantage of postulating contextually-invariant articulatory targets, or the use of simple automatic articulatory control.

Stochastic Articulatory Targets

Alternatively, it is possible to use HMM-based systems to model articulatory target positions. Bakis uses explicit articulatory positional targets, which are related to the realised articulatory positions by a set of probability distributions at the outputs of an HMM [8]. Ramsay and Deng have described a system which uses overlapping phonological gestures as the symbolic inputs to an HMM, whose output distributions represent both formant and articulatory positional distributions[12] [135]. In both cases a probabilistic distribution of possible articulations is produced, rather than a definitive sequence of articulatory movements.

Articulatory Attractors, Regions and Trajectories

Preliminary work has been described for systems which use control signals to define articulatory trajectories or movements, rather than simple point positions (although specific configurations may also be specified). For example, Vatikiotis-Bateson et al. use "via point estimation", in which several points defining an articulatory sub-trajectory are specified as a command signal [62, 166].

Several other systems falling into this category are at least in part motivated by the task dynamics model of Saltzman and Munhall [148]. This system uses an implicit target model in which articulatory goals corresponding to a phonological task are specified as attractors in articulatory space, which define constrictions in the vocal tract. Since these constrictions are usually specific to a particular articulatory configuration, the problem of discriminating between multiple acceptable vocal tract shapes during speech production is reduced.

[11] A phenomenon also known as the "ventriloquist effect", which is discussed further in Section 6.4.

[12] For more detail on these systems, see Section 2.2.2.

Bailly has described a related model of articulatory control, in which targets are regions in articulatory or acoustic space[13] which act as attractors on a biomechanical vocal tract model to produce desired articulatory trajectories [7]. A similar system developed by Honda and Kaburagi also uses target regions, and determines actual articulatory trajectories by using dynamic programming to find the path through these regions which optimises an explicit energy criterion [64, 76]. These trajectories are modelled as second-order linear systems, and hence context-sensitive system parameters are required to accurately model articulatory transitions.

Finally, Laboissiere has presented a model of articulatory movements arising from shifts in articulatory equilibrium positions [90]. A seven-muscle model of the production mechanism is used, and articulatory targets are encoded as muscle control signals which change the system's equilibrium point—in turn giving rise to the desired articulatory movements.

3.3.3 Window Models

An alternative to using specific articulatory target points or trajectories is to propose *regions* of articulatory space through which an articulator must pass in order to guarantee intelligibility. This is an inherently attractive approach, since it is both simple and flexible. It can either be used to specify "corridors" through which articulatory trajectories must pass, or else to set a range of positional values which an articulator can satisfy at any point in time. The relative degree of specification of an articulator is explicitly encoded by the size of the perceptually acceptable region, and co-articulatory effects are modelled as variations within these prescribed limits.

An early model proposed along these lines is Keating's "window" model of co-articulation [79]. This model specifies both the ranges of possible articulatory positional values and the time intervals over which these restrictions apply, as shown in Figure 3.4.

This figure shows a simulated articulatory trajectory (solid line), which is constrained to pass through the regions delineated by dashed horizontal lines. Both the duration and the degree of specification of the restriction regions are variable, and there may be intervening intervals of complete underspecification. Keating proposed that the articulatory trajectory be free to pass anywhere within these allowable ranges, with the path being chosen to fit as smoothly as possible within the constraints imposed by the context.

This is a descriptive model of articulatory movement, which differs from the point target models in that it directly specifies ranges of *achieved* articulatory motion, as opposed to indirectly controlling articulatory movements through virtual targets which are undershot by the system dynamics. Not only does this model explicitly represent relative degrees of specification, but it also emphasises the need for the rate of rise or fall of an articulator at the onset of a particular phoneme to be context-dependent, as discussed in Sections 2.3.4 and 3.3.2.

Finally, Guenther has proposed an alternative model based on target regions, which attempts to address the problem of choosing between multiple vocal tract shapes which result in similar acoustic patterns. Target regions are initially planned in formant space,

[13] The use of target regions rather than points is discussed further in Section 3.3.3.

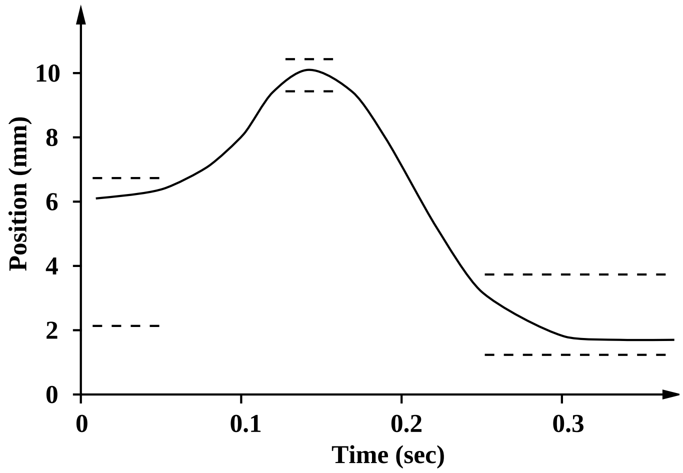

Figure 3.4: Simulated articulatory trajectory (solid line) using the window model of co-articulation. The trajectory is constrained to pass through the "windows" indicated by dashed horizontal lines.

and are then mapped onto equivalent articulatory target regions by selecting articulatory movements which best correspond to the desired formant dynamics [55].

3.3.4 A Self-Organising Probabilistic Model

A new model is now proposed which is similar to the window model in that regions of articulatory space are specified, but in which a very different description of articulatory motion is provided. A self-organising approach to modelling context-sensitive articulatory positions is used, which does not require context-dependent system parameters or rules to be inferred. The key components of this system are:

1. A descriptive probabilistic model of instantaneous articulatory positions at the mid-points of phonemes.

2. A predictive probabilistic model of articulatory positional variations at these mid-points due to co-articulatory effects.

3. A simple linear interpolation between co-articulated articulatory midpoint positions.

By contrast with Keating's model, this descriptive model does not specify time intervals during which articulators must satisfy a given spatial constraint, but instead specifies articulatory positions only at the midpoints of phonemes. Furthermore, instead of postulating a hard-limited region of equally acceptable positions, a smoothly-varying probabilistic model is used to specify articulatory positions, as demonstrated in Figure 3.5.

In this figure, the midpoints of the phonemes being articulated are indicated by dotted vertical lines, and the articulatory trajectory is shown as a dashed curve. Against each midpoint line is shown a probability density function (pdf) which represents the probability of observing various articulatory values *at that point in time*. The axes for these pdfs are the articulatory position vertically, and increasing probability in the directions

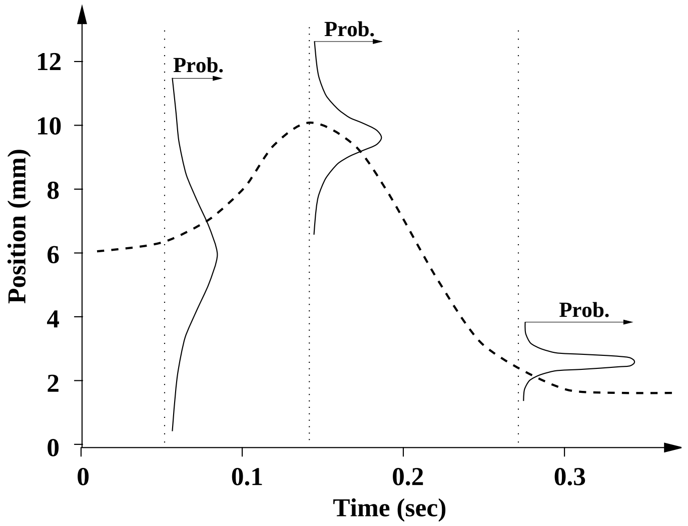

Figure 3.5: Simulated articulatory trajectory (dashed line) using a probabilistic co-articulation model. The midpoints of successive phonemes are indicated by dotted vertical lines, and associated with each midpoint is a probability distribution. These distributions define the probability that the articulatory trajectory will take particular positional values at the midpoint concerned.

shown. Thus the pdf associated with the midpoint of the leftmost phoneme indicates that the most probable (mean) articulatory position at this point in time is $\approx 6mm$, with decreasing likelihood above and below this value, and a standard deviation of $\approx 2mm$. In Keating's model these pdfs would be "square" windows rather than smooth functions, with a zero probability that an articulator will be positioned outside the window, and a flat distribution of equal probabilities within the window.

This framework for describing co-articulatory variation is neatly implemented by the technique for characterising articulatory variability described in Section 3.2. By automatically sampling a large number of articulatory trajectories at times corresponding to the midpoints of phonemes—and modelling the observed variations with unimodal Gaussian distributions—smooth probabilistic "window" functions can automatically be obtained. As discussed in Section 3.2.3, multiple distributions would be required if it were necessary to allow for the possibility of selecting between alternative articulatory strategies according to the context in which a phoneme appears. Techniques for predicting co-articulatory variation within the ranges described by these distributions, and methods for generating complete articulatory trajectories from them are described in the following sections.

3.4 Predicting Articulatory Variation

The goal of the approach to modelling co-articulatory variation taken in this dissertation is the development of an automated system for predicting variations in articulatory positions using a knowledge of the context and the relative durations of the phonetic segments. The positional variation for each articulator is directly modelled at the midpoint of each phoneme using a single Gaussian distribution. This task therefore reduces to the prediction of deviations from an articulator's mean position at these midpoints, and the generation of complete articulatory trajectories from the resulting positional values.

Articulatory movements exhibit both random and systematic positional variation[14]. While random variations are an inherent characteristic of human speech production, their perceptual effects must be negligible or else words would be randomly mis-perceived by listeners. Of greater interest is that part of the variation which is systematically reproducible, and which *may* therefore be perceptually relevant. As discussed in Section 3.3.1, consistent variations in articulatory movements due to the phonetic context are motivated by efficiency constraints, whereby articulators move early (or late) toward (or from) perceptually relevant regions, from (or toward) relatively less relevant or "unspecified" regions. To predict such movements, information which is correlated with these variations must be extracted in either explicit or rule-based form, from the time-aligned phonetic transcription.

3.4.1 Articulatory Accelerations

If it is assumed that co-articulatory effects are governed by economy of effort considerations and perceptual constraints, the greatest deviations from "ideal" articulatory behaviour are expected to be observed in regions where the muscular effort required is greatest *and* the position of the articulator concerned is relatively underspecified. While compromises in articulatory movements would be desirable wherever relatively large articulatory effort is dictated, such variations will only be permitted where the corresponding acoustic effects are perceptually acceptable.

Since the articulators have finite inertial masses, a simple measure related to the muscular effort required at a given point in time is the acceleration of the articulator concerned. In terms of articulatory trajectories, the acceleration is the second derivative, or "curvature" of the plot, so that regions of relatively high positive or negative articulatory acceleration tend to correspond to local minima or maxima in a trajectory respectively. During the production of phonemes for which an articulator's position is relatively underspecified, the required acceleration of the articulator concerned is expected to be highly correlated with deviations from the mean articulatory position, as shown in Figure 3.6.

This figure shows both a desired (solid) and an achieved (dashed) simulated articulatory trajectory around the midpoint of a phoneme, indicated by a dotted vertical line. To the right of the line marking this midpoint position is a hypothetical pdf describing the observed variation in the simulated articulator's achieved position at the midpoint, computed from many examples in a training data set as described in Section 3.2.

In regions of high acceleration, such as the large negative acceleration near the midpoint of the phoneme in this example, an economy of effort can be obtained by following a less curved trajectory, leading to an undershoot of the mean articulatory position as shown. Relatively large positive acceleration demands in turn lead to an overshoot of this mean, which therefore should not be regarded as a "target" position, but is merely descriptive of observed articulatory traces.

[14] In the model described in this dissertation systematic prosodically-based variations are also treated as "random" effects, as discussed in Section 3.2.1.

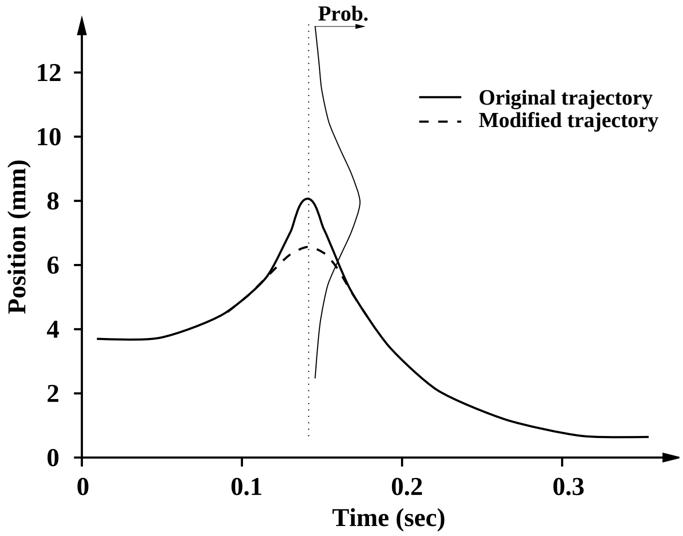

Figure 3.6: Variation in the position of a simulated articulatory trajectory at the midpoint of a phoneme (dotted vertical line). The pdf associated with the midpoint defines the probability of observing different articulatory positions. The large negative acceleration required at the midpoint of the phoneme results in an undershoot of the mean position which reduces the curvature of the resulting trajectory.

As the articulation of a phoneme comprises articulatory movements into and out of perceptually required positions, it is assumed that local minima and maxima in the trajectories usually fall close to the midpoints of phonemes as shown in this example. The width of each distribution describing the achieved articulatory positions will be determined by both the range of geometrical contexts in which the phoneme occurs—wider ranges leading to wider distributions—and the relative degree of specification of the articulator. Where an articulator's position is perceptually important, increased muscular effort will be used to achieve the desired acceleration, and less positional variation will be observed.

Estimating Curvatures

If articulatory accelerations are to be used to predict positional variations, a technique for estimating these from the curvature of time-aligned articulatory trajectories is required. The curvature c of a function f at the point x is defined as the rate of change of its gradient at that point:

$$c(x) = \frac{d^2 f(x)}{dx^2} = \lim_{\Delta x \to 0} \frac{g(x + \Delta x) - g(x)}{\Delta x} \tag{3.2}$$

where $g(x)$ is the gradient of the graph, given by:

$$g(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \tag{3.3}$$

and hence:

$$c(x) = \lim_{\Delta x \to 0} \frac{f(x) + f(x + 2\Delta x) - 2f(x + \Delta x)}{\Delta x^2} \tag{3.4}$$

Given a time-aligned phonetic transcription of an utterance, a first approximation to the shape of a corresponding articulatory trajectory can be obtained by a simple linear interpolation between that articulator's mean positions at the midpoints of each of the phonemes, as shown in Figure 3.7.

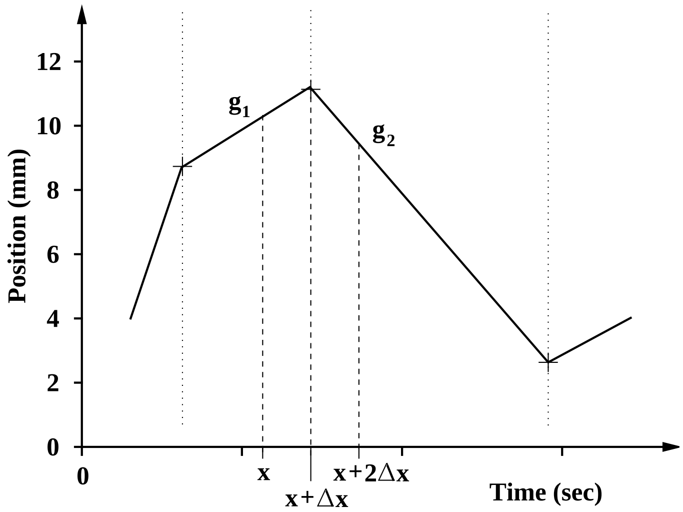

Figure 3.7: Computation of the estimated curvature measure for a simulated linearly interpolated articulatory trajectory. The midpoints of successive phonemes are shown by dotted vertical lines, and mean articulatory positions at these midpoints are indicated by crosses. The gradients of the linear segments leading into and out of the central midpoint position are denoted g_1 and g_2 respectively.

In this figure, the midpoints of phonemes are shown by vertical dotted lines, and articulatory mean positions by crosses. If the value $(x + \Delta x)$ is taken to coincide with a midpoint position as shown, then the gradients g_1 and g_2 of the initial and final articulatory trajectory segments on either side of the midpoint respectively will be given by:

$$g_1 = \frac{f(x+\Delta x) - f(x)}{\Delta x} \tag{3.5}$$

$$g_2 = \frac{f(x+2\Delta x) - f(x+\Delta x)}{\Delta x} \tag{3.6}$$

for any small value of Δx, as the segments are linear. The curvature in Equation 3.4 can then be expressed as:

$$c(x) = \lim_{\Delta x \to 0} \frac{g_2 - g_1}{\Delta x} \tag{3.7}$$

which is infinite at the midpoint of the phoneme, due to the discontinuity at the boundary between the segments. If only the *relative* curvatures of trajectories around the midpoints of phonemes in different contexts are of interest—rather than the absolute values of these curvatures—then a simple approximate measure can be obtained by setting Δx in this expression to a constant non-zero value. This simple gradient-differencing technique is therefore used to estimate relative trajectory curvatures (and hence accelerations) from a linear interpolation between successive articulatory mean positions. The resulting estimates of articulatory accelerations represent simple approximations to the amount of muscular effort dictated by the phonetic context.

Since the computation of this curvature measure requires a knowledge of articulatory mean positions, statistics describing the distribution of curvature estimates are not derived until after the articulatory positional statistics have been determined. First approximations to articulatory trajectory shapes are computed by linear interpolation between successive articulatory mean values, where these points are placed at the midpoints of phonemes using a phonetic time alignment[15] as shown in Figure 3.7. The curvature measure described above is then computed for each articulator over all examples of each phoneme, and the resulting values are modelled using a single Gaussian distribution in each case, in an analogous manner to the positional distributions described in Section 3.2.3.

3.4.2 Systematic Positional Variation Prediction

The curvature measure described in the preceding section can be used to predict systematic variations in articulatory positions at the midpoints of the phonemes in any time-aligned phonetic string. Articulatory positions and curvatures at these midpoints are modelled by the random variables P and C respectively, each of which has a pdf given by a single Gaussian distribution. From the time-aligned phonetic string, the value c of the curvature of an articulatory trajectory at the midpoint of each phoneme is estimated. Given this value, the expected value of the random variable P describing that articulator's position is required. This conditional expectation is given by:

$$E\,[P \mid C = c] \;=\; \int_{-\infty}^{\infty} p \,.\, f_{P|C}(p|c)\, dp \tag{3.8}$$

where $f_{P|C}(p|c)$ is the conditional pdf of P given $C = c$, which is defined as:

$$f_{P|C}(p|c) = \frac{f_{C,P}(c,p)}{f_C(c)} \tag{3.9}$$

where $f_C(c)$ is a Gaussian distribution with mean μ_C and standard deviation σ_C:

$$f_C(c) = \frac{1}{\sqrt{2\pi\sigma_C^2}} \exp\left(-\frac{(c-\mu_C)^2}{2\sigma_C^2}\right) \tag{3.10}$$

and $f_{C,P}(c,p)$ is the joint pdf of C and P, which in this case is just the binormal distribution:

$$f_{C,P}(c,p) = \frac{1}{2\pi\sigma_C\sigma_P\sqrt{1-\rho_{C,P}^2}} \exp\left(\frac{-1}{2(1-\rho_{C,P}^2)}\left[\left(\frac{c-\mu_C}{\sigma_C}\right)^2 - 2\rho_{C,P}\frac{(c-\mu_C)(p-\mu_P)}{\sigma_C\sigma_P} + \left(\frac{p-\mu_P}{\sigma_P}\right)^2\right]\right) \tag{3.11}$$

where $\rho_{C,P}$ is the correlation coefficient for the random variables C and P:

[15] Descriptions of the techniques used to align the acoustic signals to their transcriptions can be found in Sections 5.4 and 6.2.

$$\rho_{C,P} = \frac{E\left[(C-\mu_C)(P-\mu_P)\right]}{\sigma_C \sigma_P} \tag{3.12}$$

Substituting Equations 3.10 and 3.11 into 3.9, the following expression for the conditional pdf is obtained:

$$f_{P|C}(p|c) = \frac{1}{\sqrt{2\pi}\,\sigma_P\,\sqrt{1-\rho_{CP}^2}} \exp\left(\frac{-1}{2\sigma_P^2(1-\rho_{CP}^2)}\left[p-\mu_P-\rho_{CP}\,\frac{\sigma_P}{\sigma_C}\,(c-\mu_C)\right]^2\right) \tag{3.13}$$

which can then be substituted into Equation 3.8 to find the expected positional value. Since the conditional pdf in Equation 3.13 is itself a Gaussian however, with mean μ and variance σ^2 given by:

$$\mu = \mu_P + \rho_{CP}\,\frac{\sigma_P}{\sigma_C}\,(c-\mu_C) \tag{3.14}$$

$$\sigma^2 = \sigma_P^2\,(1-\rho_{CP}^2) \tag{3.15}$$

the expected positional value is equal to this mean:

$$E\left[P \,|\, C\!=\!c\right] = \mu_P + \rho_{CP}\,\frac{\sigma_P}{\sigma_C}\,(c-\mu_C) \tag{3.16}$$

Equation 3.16, which is known as a "regression curve" [110], allows the prediction of the most likely articulatory positional value at the midpoint of a phoneme using the estimate of the curvature, the means and standard deviations of the positional and curvature distributions, and the correlation coefficient between these two distributions.

A summary of the principal characteristics of the resulting model of co-articulatory variation is given below:

- It is entirely statistically-based and self-organising in nature.
- Articulatory positional variability at the midpoints of phonemes is modelled using a single Gaussian distribution for each articulator during the production of each phoneme.
- The required articulatory effort is estimated by computing the gradient changes ("curvatures") of linearly interpolated trajectories at the midpoints of phonemes, which are also modelled using single Gaussian distributions.
- Correlations between curvature and positional statistics are used to predict systematic positional variations from the estimated curvatures in particular contexts.

The model is symmetrical with respect to the right and left contexts of phonemes (anticipatory and carryover effects respectively), and considers only the immediate neighbours of a phoneme when computing contextual effects. The effects of context on articulatory transitions into and out of phonemes are modelled implicitly, during the construction of complete articulatory trajectories from the co-articulated values at the midpoints of the phonemes.

3.4.3 Generation of Articulatory Trajectories

Finally, a technique is required for generating continuous articulatory trajectories which are constrained to pass through the midpoint positions specified by the co-articulation model. Since these pre-specified values are not "target" positions or "attractors" for the articulators, but represent the most likely positions for the articulators at these points in time, any dynamic system model would therefore have to be such that the dynamics were varied according to the context to ensure the achievement of these positions. Alternatively, the midpoint positions could be connected using explicit linear, piece-wise linear or non-linear functions.

When choosing amongst these options it is important to consider the nature of the trajectories being approximated. On the one hand it is clear that actual articulatory movements describe smooth curves rather than linear segments, which might lead to the postulation of a dynamic system or curved function model. The caveat to using relatively complex models such as these however, is that current predictive models of articulatory position, including that described in the preceding sections, are relatively crude approximations to human physiology. As a result, it is anticipated that any errors in articulatory movements due to the choice of connecting segments will be insignificant by comparison with those due to gross errors in the articulatory positions predicted by the model. For example, Meyer et al. found in their German-language synthesiser that articulatory trajectories which were fitted to second order critically-damped transitions using Kalman filtering could be replaced with linearly interpolated trajectories, with "only small differences" being observed in the acoustic waveform [108].

The results obtained from using both linear interpolation and piece-wise (or "constrained") linear interpolation schemes to connect co-articulated values at the midpoints of successive phonemes were therefore compared. These two methods are illustrated in Figure 3.8, which shows simulated articulatory trajectories for the word "star".

In this figure, phonetic boundaries are indicated by solid vertical lines, and the midpoints of phonemes by dotted vertical lines. Co-articulated positions at these midpoints are indicated by crosses, and linear and piece-wise linear interpolated trajectories for the hypothetical articulator are shown by solid and dashed lines respectively.

While the linear trajectories are constructed by simply connecting positions at the midpoints of successive phonemes, the piece-wise linear trajectories attempt a closer approximation to actual articulatory movements by constraining the trajectories to pass through the average of two successive midpoint values at the boundary between the two phonemes concerned, as marked by circles in the figure. The resulting trajectories have higher curvatures for relatively short phonemes and lower curvatures for longer phonemes compared with the linear interpolation scheme.

When these two models were implemented and the resulting articulatory trajectories were compared with X-ray articulatory traces, no improvement in articulatory modelling accuracy was observed when using the constrained interpolation model. Standard linear interpolation was therefore retained as the model for trajectory generation in the system.

In concluding, it is observed that the use of this linear interpolation scheme implies that co-articulatory effects will be introduced even in the absence of the explicit co-articulation

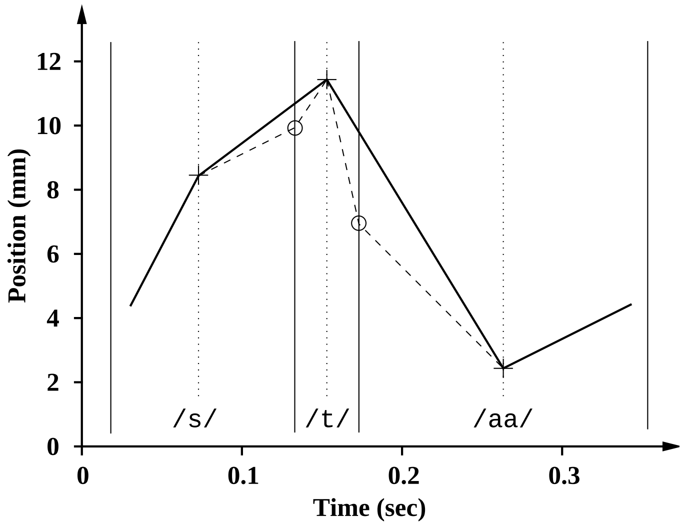

Figure 3.8: Simulated articulatory trajectories using both linear (solid) and piece-wise linear (dashed) interpolation between articulatory positions (crosses) at the midpoints of phonemes (dotted vertical lines). Phonetic boundaries are shown by solid vertical lines, and the means of successive midpoint positions at these boundaries are marked by circles.

model. If mean articulatory positions are used at the midpoints of phonemes, rather than the modified positions predicted by the technique described in Section 3.4.2, the phonetic context will still strongly influence the shape of the resulting interpolated trajectory. As demonstrated in Chapter 7 however, the use of the explicit co-articulation model leads to a significant increase in the overall articulatory modelling accuracy.

Chapter 4

From Articulators to Acoustics

4.1 Introduction

The ultimate goal of most existing models of speech production is the synthesis of a time-domain signal which approximates natural-sounding human speech as closely as possible. While formant-based and concatenative speech synthesis systems still provide the most human-like synthetic signals [42, 85], more detailed articulatory models are now being developed which may eventually surpass the performance of these systems in synthesis applications [173].

The use of articulatory production models in a computer speech recognition framework demands a different set of characteristics than those required for natural-sounding synthesis. In this chapter, a review of these desirable characteristics is provided, followed by a brief introduction to explicit vocal tract models. Subsequently, a description of an alternative approach which uses a self-organising probabilistic model of speech production to synthesise parameterised acoustic signals is presented.

4.2 Acoustic Model Selection

The features of the acoustic signal which have proved most useful for the automatic recognition of speech by machines are typically defined in the frequency, or transformed frequency domains. Since recognition typically involves the comparison of spectral waveforms to parametric spectral distributions (Section 1.2), a system such as that described in this dissertation—which seeks to augment the performance of an existing recogniser—need not synthesise a time-domain waveform at all, as demonstrated in Figure 4.1.

In this figure, the input speech to be recognised is first parameterised in the frequency domain, then passed to an HMM-based recogniser such as HTK [176], which hypothesises a number of possible transcriptions, as described in Section 8.2. Each of these transcriptions is then used as input to the SPM, which re-synthesises parameterised speech vectors. These vectors could then be converted into time-domain waveforms, but their primary purpose is to allow a comparison to be made between the re-synthesised vectors for each hypothesised transcription and those of the original speech. As detailed in Section 8.3, the transcriptions are then re-ranked in order of the likelihood that they correspond to the input utterance, according to a distance metric for comparing the parameterised vector sequences.

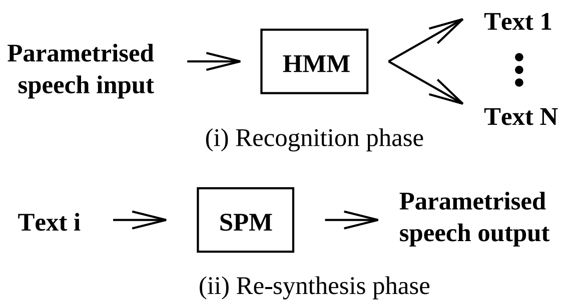

Figure 4.1: Schematic diagram of re-synthesis algorithm: (i) An HMM-based recognition system is used to generate a list of textual hypotheses corresponding to a parameterised utterance (ii) Parameterised speech is re-synthesised from each transcription.

To facilitate the computation of such a distance metric, the SPM must produce not only a prediction of the most likely spectral representation corresponding to a given transcription, but also a confidence measure for the spectral values predicted at each frequency and at each point in time. The desirable characteristics of an acoustic model in such a system can therefore be defined as follows. It should:

- Take as input a description of articulatory movements, together with a time-aligned transcription of the utterance.
- Produce as output a spectral representation of the utterance, either directly or by first synthesising a time-domain waveform and applying an explicit transform.
- Predict not only the spectral vector sequence corresponding to the utterance, but also the expected variability in these estimates.
- Automatically adapt its parameters to model data from new speakers.

As will be shown in the following sections, while explicit vocal tract models based on human physiology are well suited to meeting the first two of these requirements, implicit self-organising models may be preferable for meeting all four criteria simultaneously.

4.2.1 Acoustic Signal Parameterisations

Many different techniques have been developed for extracting features from digitally-encoded speech waveforms, including time-domain features, linear predictive coefficients and spectrally-derived representations [132]. The systems described in this dissertation use the last of these three feature types[1].

In the case of the self-organising production model, these features are directly synthesised, while the HMM-based recognition system converts the acoustic signal into a spectral representation by repeatedly applying a smooth windowing function to short sections of the time-domain speech waveform, and computing the fast Fourier transform (FFT) of the resulting values. This windowing function is chosen to have a maximum positive

[1] The possible use of alternative parameterisations is discussed in Section 9.5.2.

value at its midpoint, and falls off smoothly to zero on either side, thereby mitigating the edge effects which would otherwise be observed in the FFT of a short segment of speech. The length of the window used defines the time and frequency resolutions of the resulting spectral representation, since the signal is assumed to be spectrally stationary within the windowed region. In the HTK system for example, a Hamming window of around $25msec$ is typically used, and is stepped forward along the speech waveform by intervals, or "frames", of $10msec$ so that there is an overlap between successive windowing functions [176].

The result is a set of FFT values every $10msec$, each of which encodes a $25msec$ window of the speech signal. The frequency range of the FFT is divided into a number of frequency bins which are placed at logarithmically or Mel-spaced intervals to match the logarithmic frequency discrimination in the human cochlea. The power spectrum values in each bin, or "filterbank", are then summed to yield a quantised spectral vector in which each coefficient encodes the power in a particular frequency range during that segment of speech.

Additional transforms are often applied to these spectral vectors, the most common of which is to take the logarithm of the power values in each frequency bin, to approximate the non-linear sensitivity of the human auditory system to changes in signal intensity. Energy normalisation can also be applied, either by subtracting the mean energy in each vector from each of its coefficients, or else by computing the mean energy in each coefficient over a large number of vectors and using these values for energy normalisation over time.

Finally, a popular technique has been to compute the "cepstrum" of the signal: the *inverse* Fourier, or "cosine", transform of the log filterbank energy coefficients. This transform produces energy-normalised parameters, where the spectral information tends to be concentrated in the lower coefficients, so that by discarding the higher-order values (in effect a spectral "smoothing"), an efficient representation is obtained.

The HTK recogniser which generates hypothesis lists in the system described in this dissertation uses a cepstral encoding of the speech signal as its acoustic representation. For each speech frame, the current cepstral coefficients are computed, as are the differences between successive cepstral vectors and the differences between the differenced vectors. These "delta" and "acceleration" parameters are included as an attempt to model local dynamics in the acoustic signal, as well as the inter-dependencies between successive speech frames, as described in Section 1.2.2.

In the case of speech production systems, the computation of any of the representations described above is a straightforward matter when using a model which synthesises a time-domain acoustic signal, since the appropriate transforms can be applied explicitly to this waveform. By contrast, the self-organising model described in this dissertation generates parameterised speech vectors *directly*, and hence the choice of features must balance the complexity of the representation against its ability to capture relevant aspects of the acoustic signal.

While the cepstral representation used in the HTK recogniser provides an efficient representation of spectral characteristics, log filterbank energy coefficients were chosen as the parameterisation used in the production model, as the mapping from articulatory positions to these coefficients is less complex, and yet an accurate description of the features of the signal relevant for recognition is still possible.

4.2.2 Explicit Vocal Tract Models

A discussion of the common vocal tract models described in the literature was given in Section 2.3.4. The output of these models is either a frequency-domain or a time-domain representation of the acoustic signal, which in the latter case must be converted to a parameterised representation before use in a speech recognition framework, as described in the preceding section.

The difficulties of using explicit vocal tract models for speech recognition are twofold. Firstly, the parameters of the model must typically be tuned by hand in order to produce a signal which closely matches the speech of a particular speaker. Not only do articulatory movements vary from speaker to speaker, but the geometry and properties of the vocal tract are also variable, and there is no simple way to automate the process of adjusting these model parameters accordingly.

Secondly, a key feature of any recognition system is the ability to characterise variability in the speech signal. While an explicit vocal tract model might be able to accurately produce *an* acceptable acoustic waveform corresponding to a particular utterance, performing recognition requires the characterisation of *all* such acceptable acoustic realisations. Thus it is essential to specify which regions of the synthesised spectrum are relatively highly constrained or variable so that when computing spectral differences, errors in areas of the spectrum which are known to be highly variable will be weighted less heavily than errors in more tightly constrained regions[2].

The detailed implementation of an explicit vocal tract model based on the Kelly-Lochbaum system, which is used to generate a codebook of [articulatory vector, acoustic vector] pairs is described in Appendix C. For the reasons described above however, an implicit self-organising model of speech production has been implemented in this dissertation for the re-synthesis of parameterised speech vectors within the framework of speech recognition.

4.2.3 Self-Organising Models

By relaxing the constraint that the acoustic model should closely mimic the physiology of the human vocal tract, it is possible to formulate self-organising models of speech production which are automatically trainable, and able to characterise the variability in the acoustic signal. This is achieved by formulating either a linear or a non-linear parametric model of the mapping, along with an automatic procedure for optimising these parameters and generating confidence intervals on the outputs. Depending upon the size of the model, this learning procedure may require the use of more training data than would be needed to hand-tune the parameters of an explicit vocal tract model.

The input to such a model is a specification of either articulatory positions or vocal tract area functions, and the outputs produced are typically acoustic spectral vectors. While vocal tract areas must be supplied at the inputs of explicit vocal tract models, articulatory positions are directly used as the inputs to the SPM, since data describing X-ray articulatory traces are more readily available (Chapter 5).

[2]This issue is discussed further in Section 8.3.1.

When learning a mapping from these articulatory input parameters to an output spectral representation, there are three sources of variability in the resulting acoustic signal:

- Uncertainty in the original input values specifying articulatory or vocal tract positions.
- Uncertainty in the optimum values of the model parameters.
- Inherent "noise" or random variations in the target acoustic vector outputs themselves.

The first of these is characterised at articulatory midpoints by the variance of the conditional pdf for the articulator's position given a curvature value, which was derived in Section 3.4.2 as:

$$\sigma^2 = \sigma_P^2 \, (1 - \rho_{CP}^2) \tag{4.1}$$

and which varies from zero in the case of fully correlated positional and curvature distributions, up to the original positional variance in the case of completely uncorrelated distributions. The utility of this measure of input variability is limited by the fact that it is defined only at articulatory midpoints.

An approximation to the acoustic variability arising from both the uncertainty in the model parameters and the noise on the target values can be made by measuring the performance of the system on a training data set[3]. Once the parameters of the system have been optimised, the error variance for each of the model's outputs can be measured over the entire training set, and used as a confidence measure for the prediction of unseen acoustic output vectors, as described in Section 4.4.1.

The disadvantage of using a self-organising model to learn this acoustic mapping is that no use is made of explicit knowledge concerning wave propagation in the vocal tract. Where this knowledge is incomplete or makes erroneous approximations or assumptions however, self-organising models benefit by comparison with explicit techniques, through their ability to automatically identify useful features of the input data set. An example of a linear self-organising model is described in Section 4.3, and non-linear models employing artificial neural networks are the subject of Section 4.4.

4.3 Linear Regression

Linear regression is the process of finding a (linear) approximation to a function relating a dependent output variable y to one or more input variables $x_1, \ldots, x_p$. This exact function is approximated by a linear combination of the input variables along with an error, or "disturbance", u:

$$y = b_0 + b_1 x_1 + \cdots + b_p x_p + u \tag{4.2}$$

[3] Alternatively, an acoustic model incorporating explicit distributions on parameter values could be used [11].

where $b_0, \ldots, b_p$ are constants known as regression parameters. Using matrix notation with n data points:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{u} \tag{4.3}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_0 \\ \vdots \\ b_p \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \tag{4.4}$$

A vector of regression coefficients $\boldsymbol{b} = [b_0 \cdots b_p]^T$ is required, where $[\cdot]^T$ denotes the transpose of a vector or matrix, such that the estimated function values:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b} \tag{4.5}$$

minimise a cost function related to the error in the estimates. Many different linear system implementations are possible [105], and in this case the model is optimised so as to minimise the sum of the squares of the errors,

$$E = (\hat{\boldsymbol{y}} - \boldsymbol{y}).(\hat{\boldsymbol{y}} - \boldsymbol{y}) \tag{4.6}$$

This expression is minimised [32] by solving the system of equations known as the normal equations:

$$(\boldsymbol{X}^T\boldsymbol{X})\,\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y} \tag{4.7}$$

Provided that $(\boldsymbol{X}^T\boldsymbol{X})$ is non-singular, $\boldsymbol{b}$ can then be found as

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{4.8}$$

A separate vector of regression coefficients is estimated for each output in a multi-variable system, and in each case these values are obtained by evaluating Equation 4.8 over a set of training data points. The corresponding output error variances on the training data can then be computed, and output values for unseen input data can be predicted by using these coefficients in Equation 4.5.

The advantage of this linear predictor is that it is extremely simple to implement and fast to train. By definition however, it cannot model the non-linearities in the mapping from articulatory positions to acoustic vectors, and it may not be very robust to outliers in articulatory space, particularly in the case of sparse data sets. Since the model works by fitting hyperplanes to the training data points, test set points which lie beyond the range of those points encountered during training will be mapped onto output vectors by linear extrapolation along a direction which may have been ill-determined during training. Thus in the case of sparse or noisy data sets, even in the absence of non-linearities in the mapping, models such as these can result in large output errors.

4.4 Artificial Neural Networks

The term *artificial neural network* (ANN) encompasses a broad range of engineering models with the common feature that they make use of an interconnected network of "nodes". Each of these performs a simple computation based on its inputs, but together they may comprise a complex non-linear mapping from the network's inputs to its outputs [59, 60, 109, 146].

One of the driving forces behind the development of ANNs was the desire to build models based on biological systems. The majority of network architectures currently in use however, employ architectures and training mechanisms which are biologically implausible, but have proven to be implementable and useful. While it can be shown theoretically that a network of sufficient size is capable of learning any non-linear function of its inputs, practical implementations require networks of finite size which can be trained in reasonable time frames.

As is the case for many other statistical models—including HMMs—ANNs have the desirable ability to automatically learn a mapping between a set of input variables and a desired output representation. For HMMs, this output representation is a parametric one: the model optimises the parameters of a set of Gaussian distributions which describe the acoustic feature vectors. Some ANN models, such as radial basis function networks, are similar to this in that they use pre-determined distributions to model decision regions in classification tasks. The models described in this section however, make weaker assumptions about the nature of the mappings they approximate. While they do make implicit assumptions regarding the complexity of this mapping through the selection of network architectures, they do not use explicit parametric distributions at their outputs [113].

This lack of strong assumptions about the nature of the mapping being approximated means that ANNs can be trained without the need for *a priori* application-specific knowledge. In practice however, good performance on complex tasks is rarely achieved without exploiting such knowledge (such as the use of a frequency-domain parameterisation of an acoustic signal for example) and ANNs should be viewed as offering only a partial replacement of *a priori* parameter selection by adaptive learning procedures. The potential benefit of this approach is that ANNs are able to automatically identify useful features in the input space, and to find optimum combinations of these using a simple objective function which does not place independence constraints on individual features.

Network architectures may have either a feed-forward (no recursive connections) or a feedback (with recursive connections) structure. They are trained by the presentation of input data together with the adjustment of a set of network parameters. These input data may be either time-ordered or unordered, and in the majority of cases the network is trained on a finite number of samples, then subsequently used with fixed parameter values. The training itself may be unsupervised (fully self-organising networks), or supervised, in which case "target" output vectors are presented to the network during training. Alternatively, reinforcement learning techniques can be used, in which a simple positive or negative teaching signal is used.

Many different implementations of these architectures and training techniques are possible. A detailed review of the theory and use of ANNs in speech processing and other applications can be found in the literature [11, 23, 113, 141], and a brief discussion of the use of ANNs to predict phoneme probabilities from acoustic vectors was given in Section 1.2.3. This section presents a brief introduction to the multi-layer perceptrons and modular network architectures which were used in the SPM to predict parameterised acoustic vectors from articulatory representations.

4.4.1 Multi-Layer Perceptrons

Multi-layer perceptrons (MLPs) are one of the most popular forms of ANN, and have been applied to a wide variety of classification and function approximation tasks. In this section a brief introduction to MLP architectures and training techniques is provided, before proposing a method for using them to predict parameterised speech vectors from articulatory trajectories.

MLP Architectures

An MLP comprises several layers of interconnected perceptrons, or "nodes", through which a set of input values are propagated to obtain a corresponding set of output values. Figure 4.2 shows a typical configuration of a two-layer "feed-forward" network. Connections are only made between adjacent layers and signals are propagated uni-directionally through them. In this example the adjacent layers have been fully inter-connected.

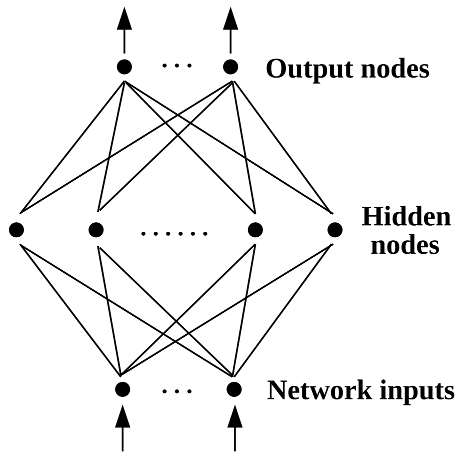

Figure 4.2: Two-layer feed-forward MLP architecture, with fully inter-connected layers.

The middle, or "hidden", layer provides an internal representation of the input space, from which the output values are derived. When the network is used for classification, this hidden space ideally represents a linearly-separable transformation of the input space. When learning a continuous function of the inputs as is the case in the SPM, these hidden units compute the moments of the input signals, for use in approximating the functions at the network outputs.

The input to each node in the hidden and output layers is computed by adding a constant "bias" value to a weighted sum of the outputs of the nodes in the previous layer, where these weights are associated with the connections between pairs of nodes. The output signal y_j of a hidden or output layer node j is computed from this input signal as follows:

$$y_j = \mathcal{F}\left(\sum_i (w_{ij} y_i) + b_j\right) \tag{4.9}$$

where y_i is the activation of node i in the previous layer, b_j is the bias for node j, and w_{ij} is the weight for the connection between nodes i and j. The activation function $\mathcal{F}$ applied to the outputs of the hidden layer nodes is usually chosen to be a differentiable function, to permit gradient-based optimisation of the network. In addition, this function typically takes the form of a thresholding non-linearity, a constraint which serves two useful purposes. Firstly, it permits non-linear functions of the network inputs to be computed. Secondly, the thresholding of the function's output for very large positive or negative values at its input means that the network mapping will be relatively insensitive to outliers in the data set.

A popular choice of hidden layer activation function, and that employed in the networks used in this dissertation, is the sigmoid function:

$$\mathcal{F}(x) = \frac{1}{1 + \exp(-x)} \tag{4.10}$$

which approaches zero for large negative values of x and one for large positive x values—referred to as the "saturation" regions of the curve. When the network is used for classification, a thresholding function such as this is typically also applied at the output layer nodes. In the case of a "regression" or smooth function-approximating network however, the activation function on the output nodes is just the identity function.

MLP Training

The supervised training of an MLP involves the adjustment of the network's parameters (weights and biases) to minimise the differences between the network's outputs and a set of target output vectors over a training data set.

The training data comprise a set of [input, output] vector pairs—in this case articulatory vectors and log spectral vectors respectively—which are repeatedly presented to the network. The network's parameters are initialised to small random values, and the input vector values are then propagated forwards through the network, to obtain a set of predicted network output values which can be compared with the supplied target vectors. The error measure commonly used for this comparison in regression networks is the summed squared error given by:

$$E = \sum_n \sum_k \frac{1}{2}(y_{n,k} - t_{n,k})^2 \tag{4.11}$$

where $y_{n,k}$ and $t_{n,k}$ are the k^{th} predicted and target coefficients respectively, for the n^{th} vector in the training set[4].

[4]If the target data distributions are Gaussians, the minimisation of this summed squared error criterion corresponds to maximising the likelihood of the network model given the training data [11].

Numerous techniques have been developed for optimising the weights and biases of such a network so as to minimise this error measure. Each of these is based to some extent on computing the partial derivatives of this error measure with respect to each network parameter, and using these derivatives to propagate the error back through the network to obtain updated parameter values ("back-propagation").

One of the simplest optimisation techniques based on these error derivatives is gradient descent, in which the parameters of the network are adjusted so as to move *in parameter space* in the direction of the negative of the local gradient of the error surface. Variants of this technique use different methods for computing the exact size and direction of the parameter updates to obtain faster or more robust convergence, and descriptions of them can be found in the literature [11, 59, 60].

The resilient back-propagation (RPROP) algorithm was used to train the networks used in this dissertation, since it gives relatively fast convergence, yet is easy to implement. This algorithm is a form of error back-propagation in which a separate update step size is maintained for each parameter in the network [25, 140]. For each input vector the sensitivity of the network's output error with respect to each of its parameters is determined by computing the partial derivatives of the error with respect to these parameters:

$$\frac{\partial E}{\partial w_{ij}}, \quad \frac{\partial E}{\partial b_j} \tag{4.12}$$

where the error E is as defined in Equation 4.11, and w_{ij} and b_j are network weight and bias parameters respectively. These partial derivatives are then evaluated by using the chain rule for differentiation, along with the expressions in Equations 4.9 and 4.10.

The weights in the network are then updated according to the rule:

$$\delta w_{ij} = \begin{cases} -\Delta_{ij} & \text{if } \frac{\partial E}{\partial w_{ij}} > 0 \\ +\Delta_{ij} & \text{if } \frac{\partial E}{\partial w_{ij}} < 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.13}$$

where the partial derivatives are summed over all vectors in the training set before updating the network's parameters[5], and a similar expression applies for updating b_j. The parameter-specific update value $\Delta_{ij}^{(n)}$ at iteration n is found from the corresponding update value for the previous iteration according to:

$$\Delta_{ij}^{(n)} = \begin{cases} \eta^{+} * \Delta_{ij}^{(n-1)} & \text{if } \frac{\partial E^{(n-1)}}{\partial w_{ij}} * \frac{\partial E^{(n)}}{\partial w_{ij}} > 0 \\ \eta^{-} * \Delta_{ij}^{(n-1)} & \text{if } \frac{\partial E^{(n-1)}}{\partial w_{ij}} * \frac{\partial E^{(n)}}{\partial w_{ij}} < 0 \\ \Delta_{ij}^{(n-1)} & \text{otherwise} \end{cases} \tag{4.14}$$

where $0 < \eta^{-} < 1 < \eta^{+}$. The effect of these update equations is to increase the size of the weight step when successive partial gradient evaluations have the same sign—since in this case the direction of the local minimum is unchanged. The step size is decreased when the gradient changes sign, since the last parameter update is then assumed to have resulted

[5] This is known as "off-line", or "batch" training.

in a jump over such a minimum. The algorithm has been found to converge rapidly compared with simple gradient descent optimisation techniques [71], and its performance is not particularly sensitive to the values of η^- and η^+, which were empirically set to 0.5 and 1.2 respectively.

Cross-Validation

When optimising each network, cross-validation is used to prevent over-fitting to the training data[6]. The parameters of the network are iteratively updated to reduce the error obtained over the training set, and the network's performance on a separate cross-validation set is monitored. The training procedure is terminated when optimum performance is obtained on the cross-validation set, which usually occurs well before the optimum performance on the training set would have been achieved.

This early-stopping procedure is intended to prevent the network from attempting to fit any noise which is present on the training data. The goal of parameter optimisation is to learn an approximation to the systematic function which is assumed to map the input values to their corresponding outputs. Such an underlying function should be common to the training and cross-validation sets, while the noise will differ between these two sets. By training on one set and using a separate set to determine the stopping criterion, it is hoped to obtain a set of network parameters which characterise the function but not the noise.

Confidence Measures

Once a network has been optimised over a training data set, ideally it would be used not only to obtain a prediction of the most probable output for an unseen input, but also some confidence measure in this predicted value.

One simple way to compute an approximate confidence measure such as this is to assess the accuracy of the network mapping on the training data[7]. Once the network has been optimised, the biases on the network outputs will be close to zero over the training data set, so an error variance on the k^{th} output can be computed as:

$$\sigma_{\text{err}_k} = \sum_{n=1}^{N} \frac{(t_{n,k} - y_{n,k})^2}{N-1} \tag{4.15}$$

where N is the number of training data vectors, $t_{n,k}$ is the target value for the k^{th} output for vector n, and $y_{n,k}$ is the corresponding value predicted by the network. The result of this process is a global error variance estimate for each network output, which can be used as an approximation to the expected error variance for unseen test data when computing spectral differences during recognition (Section 8.3.1).

[6] An alternative approach to avoiding over-fitting is to incorporate an additional term into the cost function used during training, which explicitly penalises large network parameter values—a technique known as *regularisation*. A detailed description of such algorithms can be found elsewhere [11, 99].

[7] Several alternative approaches to error estimation are also possible, such as the use of multiple networks to gauge uncertainty [99] and the estimation of error bars from explicit distributions on the network parameters [11].

MLPs for Speech Synthesis

The use of MLPs for synthesising speech is uncommon compared with rule-based formant synthesis and concatenative time-domain synthesis techniques. Sejnowski and Rosenberg have described one such system which uses an MLP to predict the values of articulatory features which are successfully used to drive an articulatory synthesiser [154]. More recently, Iso has described a model which uses an MLP to directly predict mel-spectral speech coefficients from both the previous mel-spectral vectors and a linguistic control command sequence [70]. Both the control command sequence and the parameters of the MLP which predicts the acoustic vectors are learned from a training data set, and the system is used to perform recognition of English spoken letters.

In the model described in this dissertation, MLPs are used to predict log spectral acoustic coefficients directly from articulatory positions. Unlike Iso's proposal, this system does not provide contextual inputs to the MLPs, but predicts acoustic vector outputs from the current articulatory input alone.

This simplification is made possible by the articulatory framework described in Chapter 3, which models co-articulatory effects explicitly during the generation of articulatory trajectories. The task of the networks is therefore reduced to the prediction of the acoustic signal that would result from the specified articulatory configuration at any point in time during the utterance.

4.4.2 Modular Networks

A problem commonly encountered when training ANN mappings is that of scalability: performance on a particular task usually degrades significantly as the complexity of the data being modelled increases. If the size of the network used is increased in order to provide a more powerful model, training the network's parameters often becomes impractical and, even where this is possible, performance usually does not improve in proportion to the size of the network.

An alternative solution is to use a number of smaller networks, each of which models a part of the overall problem; in this way each individual network can be kept to a manageable size as it deals with a task of reduced complexity. In applications such as phoneme classification for example, a hierarchical system which first separates speech frames into voiced and unvoiced categories could be used as a simple basis for dividing the input data into distinct subsets—each of which would then be used as input to a separate classifier. The ease with which a modelling task can be decomposed in this way varies greatly from problem to problem however, and even if a suitable subdivision is found it does not automatically follow that the outputs from the component networks can easily be recombined in a meaningful way.

One technique for achieving this is to use a hierarchical mixture of experts [72]. This approach uses a number of individual "expert" networks to solve a set of sub-problems, and provides a series of "gating" networks to recombine their outputs into an overall system response. The choice of architecture (the number of expert and gating networks) is usually made prior to commencing training, however the positioning of these networks within the input space is not specified *a priori*, but is learned in conjunction with the training of the network parameters, using maximum likelihood estimation.

The task of synthesising parameterised acoustic vectors described in this chapter provides a natural basis for sub-dividing the input space without the need for such an implicit data-driven technique. Since the acoustic vectors are synthesised from articulatory vector sequences which correspond to time-aligned phonetic labels, the identities of these phonemes can be used as the basis for such a sub-division. Thus all the articulatory vectors corresponding to the phoneme `/aa/` in an utterance can be passed to a single ANN dedicated to synthesising acoustic vectors for this phoneme, and similar networks can be used for each of the other phonemes being modelled[8]. This approach has the advantages that:

- Each component network learns only a small subset of the mapping from articulatory positions to acoustic vectors.
- Standard MLP architectures and training techniques can be used.
- Since the identities of the phonemes are known *a priori*, there is no need to provide an explicit specification of the excitation sources associated with the articulatory configurations.

The importance of this last item can be seen by considering the differences between voiced and unvoiced fricative pairs such as `/z/` and `/s/`. Since the positions of the articulators are identical during the production of these two sounds, a network which was trained to predict acoustic vectors from articulatory positions for both of these phonemes would require an additional input to indicate whether voicing was present or not. By using a separate network for the data from each phoneme this problem is avoided.

The acoustic vector outputs produced by the various networks can then be concatenated in the correct order to produce a vector sequence corresponding to the entire utterance. Alternatively, the overall number of parameters in the system could be reduced by combining the data sets for phonemes with very similar acoustic realisations but differing articulations, and training individual networks on the data from these groups of phonemes. This strategy would also be of benefit where there is insufficient data for a particularly uncommon phoneme, for which it would otherwise be difficult to reliably train an MLP mapping.

This concludes the descriptions of the articulatory and acoustic models used in the SPM. In the following chapters, the databases to which these models were applied are described, and the articulatory-acoustic modelling results and recognition performances obtained are presented and analysed.

[8]This corresponds to a mixtures-of-experts approach in which the gating networks are replaced by a hard sub-division of the input space, so that acoustic vectors are synthesised by switching between the appropriate phoneme-specific MLP "experts".

Chapter 5

X-ray Data: University of Wisconsin

5.1 Introduction

To train a self-organising speech production model, a data set comprising articulatory positional traces along with synchronously recorded speech waveforms is required. Although articulatory data such as these have been available for many years, until recently only very small amounts of data have been available for any one speaker [66, 124], and previous corpora typically provided data for only two or three speakers. The system described in this dissertation seeks to learn a statistical description of the behaviour of an individual's articulators in all of the various contexts encountered during speech production. As a result, it is important to have as much phonetically diverse speech material available as possible.

Recently, Westbury and his team have used the X-ray microbeam (XRMB) facility[1] at the University of Wisconsin (UW) to assemble a comparatively large speech production database, which not only includes recordings from many different speakers, but also a relatively lengthy and varied speech task inventory [171]. This chapter provides a brief description of the UW XRMB database in its "raw" format, and describes the techniques used to process this data in order to align the articulatory and acoustic waveforms with their phonetic transcriptions. The resulting data set can then be used to train the articulatory and acoustic models described in Chapters 3 and 4, to yield a model of the speaker's production mechanism suitable for use in automatic speech recognition, as described in Chapters 7 and 8.

5.2 Corpus Description

The UW XRMB database contains articulatory positional traces along with synchronously recorded speech waveforms for 57 speakers of American English, comprising 32 females and 25 males. Detailed information regarding the age, sex, height, weight and dental state

[1] A discussion of this and other techniques for the acquisition of articulatory data during speech production was presented in Section 2.3.2.

of each speaker is provided, along with their educational history, non-English language training, place of birth, dialect base and place of residence.

During data acquisition, a total of four signal types were recorded:

- Acoustic pressure wave.
- Neck wall vibrations.
- Horizontal and vertical positions of eight gold pellets in the vocal tract.
- Frontal and lateral video images.

For the purposes of this dissertation only the acoustic and articulatory information sources are of interest. The corpus contains both read speech and (nominally) silent oral motor tasks, and an approximate breakdown by recording time for each task type is given in Table 5.1.

Sentences:	40%
Citation words and sound sequences:	33%
Prose passages:	13%
Oral motor tasks:	8%
Counting and sequences of number names:	6%

Table 5.1: Breakdown of UW XRMB corpus.

The total recording time was approximately $19min$ per speaker, and of this only $\approx 17min$ were used, by excluding the oral motor tasks and non-word sound sequences. Approximately three quarters of the data in each of the remaining speech tasks was used for training the system (≈ 5000 phonemes per speaker), with the remainder reserved as speaker-dependent test sets. Although nominal word-level transcriptions were provided for each task, a considerable amount of pre-processing of the waveforms and transcriptions was required before a phonetic alignment could be produced.

5.2.1 Speaker Sample

The raw data described in this chapter were drawn from a pre-release compact disk containing a subset of the UW XRMB corpus. This permitted the evaluation of the SPM on data from six speakers, comprising three females and three males.

A subset of the speaker details taken from the UW handbook [171] for these six speakers is listed in Table 5.2. The subjects are of similar ages, and all except jw45 have a dialect based in Wisconsin.

5.2.2 Acoustic Data

The speech signal was recorded in the XRMB facility at a sampling period of $46\mu s$ ($\approx 21739Hz$) by using a directional microphone in the presence of machine noise. A fixed recording period was used for each task, which occasionally resulted in truncated

Speaker	*Sex*	*Dialect Base*	*Age*
jw16	F	Kiel, Wisconsin	20
jw27	F	Blair, Wisconsin	20
jw29	F	Milwaukee, Wisconsin	20
jw18	M	Hudson, Wisconsin	19
jw24	M	Jefferson, Wisconsin	19
jw45	M	Mishawaka, Indiana	21

Table 5.2: University of Wisconsin subject details.

recordings for slower speakers[2]. In addition, a short tone was played at the start of each task, and background comments such as "good" and "rep" were present at the end of many utterances[3].

A notch filter was applied to this raw acoustic signal to remove background noise at $5435Hz$, and the resulting signal was down-sampled to $16kHz$. The $16kHz$ speech was then parameterised into 12-dimensional Mel-frequency cepstral coefficients (MFCCs) and 24-dimensional Mel-frequency log spectral coefficients using HTK [176]. In both cases a Hamming window of length $25ms$ was applied to the acoustic signal before computing the Fourier transform, and a step size of $10ms$ was used between adjacent parameterised speech frames, as described in Section 4.2.1. The result of each of these parameterisations was a sequence of spectral vectors—one every $10ms$—corresponding to the acoustic signal. Since the first vector cannot be computed until $25ms$ of speech is available, the first frame will be centred at $12.5ms$, and subsequent frames at $22.5ms$, $32.5ms$, etc.

The two separate parameterisations were used for recognition and re-synthesis respectively. Since good recognition and phonetic alignment performance have been achieved using MFCC parameters in HTK-based systems [174], the MFCC vector sequences are used to provide the alignments of the training data to the corresponding transcriptions[4], and to produce transcription hypotheses and alignments for the test data sets. This also allows existing MFCC-based phonetic HMMs to be used as a starting point for training speaker-dependent models, as discussed in Section 5.4.1.

When re-synthesising a parameterised acoustic representation from a hypothesised transcription, log spectral vectors are used instead of MFCCs for the reasons outlined in Section 4.2.1. In this case 24 coefficients are used in each acoustic vector to provide a relatively detailed spectral representation of the signal, as opposed to the 12 MFCCs used when performing recognition, which provide a more complex but efficient encoding of spectral information.

[2]This problem is discussed in further detail in Section 5.3.1.

[3]A description of the techniques used to model these "spurious" sounds is given in Section 5.4.1.

[4]The details of this alignment process are presented in Section 5.4.

5.2.3 Articulatory Data

Articulatory positions in the UW system are determined by using a narrow X-ray beam to track the movements of gold pellets glued to the tongue, jaw and lips of a subject who reads from the set corpus. Three reference pellets were attached to the subject's head, and eight articulatory pellets were tracked relative to these, with the subject's head viewed in profile by the apparatus. Two of the reference pellets were placed on the bridge of the subject's nose, and the third on the maxillary (upper) incisors. The tracking pellet names and their physical positions are summarised in Table 5.3.

Upper lip:	*UL*
Lower lip:	*LL*
Ventral tongue:	*T1*
Mid-tongue:	*T2*, *T3*
Dorsal tongue:	*T4*
Mandibular incisor:	*MI*
Mandibular molar:	*MM*

Table 5.3: Locations and names of X-ray tracking pellets.

The four tongue pellets *T1* to *T4* are approximately equally spaced along the centre of the tongue from the tip (*T1*) to the back (*T4*), and the labial pellets are placed at the midline of the upper and lower lips. The mandibular pellets are attached to the teeth at the front (*MI*) and back (*MM*) of the jaw.

The x and y positions of each pellet were tracked relative to the three reference pellets attached to the subject's head, in order to remove any motion due to displacement of the head during recording. The pellet movements were recorded at sample rates which varied according to the relative accelerations of the articulators, from a minimum of $40 samples/s$ for those on the jaw and upper lip, to a maximum of $160 samples/s$ for the tip of the tongue. These trajectories were then interpolated and re-sampled at a uniform sampling period of $6.866ms$ ($\approx 146Hz$) before inclusion in the XRMB database.

Due to limitations imposed by the experimental set-up, the XRMB system occasionally mis-tracks pellets during recording, due to the loss of a pellet trace[5], or as a result of confusion between two pellet traces which pass close to one another. Tracking may be lost for a brief period or throughout an utterance, and each tracking error usually affects only one or two articulatory pellets. These tracking errors have been identified and marked by hand by Westbury's team, so that the corresponding points can be excluded from the positional sampling algorithm.

For the speaker jw29, tracking errors were sufficiently frequent in the mandibular molar pellet (parameters *MMx* and *MMy*) that articulatory statistics for these two variables could not be obtained, and hence the models for this speaker contain only the first 14 articulatory parameters. As discussed in Section 7.2.1 however, the value of the articulatory variable *MMx* does not discriminate well between phonetic identities, and *MMy* is highly correlated

[5]For example, a pellet may be lost by moving into the vicinity of a metallic dental filling which masks the characteristic pattern of the pellet usually seen by the detector.

with *MIy*, due to the placement of both the *MM* and *MI* pellets on the rigid mandible. As a result, it is anticipated that the loss of the *MM* pellet may not have a significant adverse effect on the performance of the production model for this speaker.

The pellet trajectories were re-interpolated (excluding mis-tracked points) and re-sampled at intervals of $10ms$ starting from $12.5ms$, to give values corresponding to the centres of the parameterised speech frames described in Section 5.2.2.

5.3 Phonetic Transcriptions

Nominal word-level transcriptions are provided for each of the speech tasks in the XRMB database. In order to perform a phoneme-level alignment to permit positional sampling at the midpoints of phonemes however, a transcription at the phonetic level is required.

This is achieved by defining a suitable phoneme set, and constructing a phonetic dictionary which can be used to translate word-level transcriptions into their corresponding phonetic transcriptions, as described in the following sections.

5.3.1 Word-Level Transcriptions

Although the subjects were reading from a fixed corpus during signal acquisition, they occasionally deviated from the text supplied. In addition, due to the use of a fixed recording interval for each of the tasks, the utterances for slower speakers were frequently truncated.

As a result it was necessary to hand-edit the word-level transcriptions to ensure correspondence with the recorded acoustic signal. In some cases this simply involved the correction of mis-read words or the deletion of utterance-final words; where a recording was terminated during the production of a word however, the acoustic file itself was edited to remove the signal corresponding to the resulting word fragment.

5.3.2 Phoneme and Phone Sets

The Defense Advanced Research Projects Agency (DARPA) Resource Management (RM) corpus phoneme set defined by Lee [93] was used as the basis for the phoneme-level transcriptions in the system, and is listed[6] in Table 5.4.

The "closure" phonemes in this table occur only word-finally, and represent pronunciations of words where a word-final stop is not released. This set of 47 phonemes is not exhaustive[7], but serves as a useful approximation to the discriminatory sounds of English.

In order to accurately characterise articulatory movements, a *phone-level* rather than a phoneme-level description of utterances is required. Phonemes such as `/b/` for example, require a combination of two very different articulatory gestures: the closure of the lips and their subsequent release. If articulatory positions are to be characterised by sampling trajectories at pre-determined positions, the acoustic signal must therefore be segmented into the minimal acoustically self-consistent units, or phones. While most of the phonemes

[6] The phonetic categories used in this table are only broadly indicative of the natures of the various sounds. More detailed descriptions of phonetic features can be found in introductory phonetics texts, such as [117].

[7] For example, there is no phoneme in this set representing the glottal stop at the start of "utmost".

	Vowels and Diphthongs		
Very front vowels:	`/ih/`, `/iy/`	Back diphthongs:	`/aw/`, `/ow/`, `/oy/`
Near front vowels:	`/ae/`, `/eh/`, `/ix/`	Near back vowels:	`/aa/`, `/ah/` `/ax/`, `/er/`, `/uh/`
Front diphthongs:	`/ey/`, `/ay/`	Very back vowels:	`/ao/`, `/uw/`
	Liquids and Nasals		
Liquids:	`/l/`, `/r/`, `/w/`, `/y/`	Nasals:	`/m/`, `/n/`, `/en/`, `/ng/`
	Fricatives		
Strong fricatives:	`/dh/`, `/jh/`, `/ts/`, `/v/`, `/z/`	Weak fricatives:	`/ch/`, `/f/`, `/hh/`, `/s/`, `/sh/`, `/th/`
	Stops and Closures		
Unvoiced stops:	`/k/`, `/p/`, `/t/`	Closures:	`/dd/`, `/kd/`, `/pd/`, `/td/`
Voiced stops:	`/b/`, `/d/`, `/dx/`, `/g/`		

Table 5.4: Resource Management corpus phoneme set.

in Table 5.4 already satisfy this criterion, modifications are required to the segmental units for stops, diphthongs and the strong fricative `/ts/`. This fricative is very similar to a stop in that it requires a closure and subsequent release, but differs in that the burst in this case is aspirated. For the purposes of defining phone sets and for ease of notation, it will be included with the unvoiced stops in the following discussions.

To derive such a phone set, the stops (including `/ts/`) and diphthongs are therefore sub-divided into two component phones, such that the diphthongs are composed of initial and final voiced segments, and the stops comprise a closure followed by a release, as shown in Table 5.5.

	Diphthongs		
Front initials:	`/ey.1/`, `/ay.1/`	Back initials:	`/aw.1/`, `/ow.1/`, `/oy.1/`
Front finals:	`/ey.2/`, `/ay.2/`	Back finals:	`/aw.2/`, `/ow.2/`, `/oy.2/`
	Stops		
Unvoiced closures:	`/k.1/`, `/p.1/`, `/t.1/`, `/ts.1/`	Voiced closures:	`/b.1/`, `/d.1/`, `/dx.1/`, `/g.1/`
Unvoiced releases:	`/k.2/`, `/p.2/`, `/t.2/`, `/ts.2/`	Voiced releases:	`/b.2/`, `/d.2/`, `/dx.2/`, `/g.2/`

Table 5.5: Diphthong and stop phone sets.

This modified phone set removes the need for the separate closure phonemes in Table 5.4, which can now be replaced by the initial closure phones of stops, without their associated releases. After the deletion of these phonemes a set of 56 phonemes remains[8]

[8]For ease of description, the term "phoneme" will be employed in this dissertation for collectively referring to these phonetic units, and "phone" will be used only when specifically referring to the sounds listed in Table 5.5.

which are used to characterise articulatory positions[9, 10].

5.3.3 Dictionary Construction

A phonetic dictionary comprises a list of all the words in the vocabulary of a given corpus, along with one or more phonetic pronunciations per word, and is used to convert word-level transcriptions to phonetic transcriptions. The UW corpus has a vocabulary of 440 words, for which a dictionary of *phoneme-level* pronunciations was constructed using the RM corpus phoneme set in Table 5.4, excluding closure phonemes. The subsequent conversion to phone-level transcriptions using the phones in Table 5.5 is not performed until the acoustic waveforms are aligned to their transcriptions, as described in Section 5.4.

The advantage of this technique is that it is not necessary to maintain separate pronunciation entries in the dictionary for words ending in stops, according to whether the word-final stop is released or not. For example, two dictionary entries would previously have been required using the RM phoneme set for the word "hit", representing the alternative pronunciations `/hh ih t/` and `/hh ih td/`. Now only the former entry is required, and according to the particular acoustic realisation encountered during alignment the phoneme `/t/` can be replaced with either the closure `/t.1/` or the closure and release `/t.1 t.2/`.

The dictionary was constructed using entries taken directly from the RM corpus dictionary wherever possible. Pronunciations for words in the UW corpus which did not appear in the RM dictionary were derived by "translating" entries from the much larger LIMSI-ICSI Wall Street Journal dictionary to RM phoneme-set pronunciations. Multiple pronunciations of many words were provided, where these alternative forms usually resulted from the reduction of vowels to the neutral schwa `/ax/`, such as the reduction of `/dh iy/` to `/dh ax/` in "the". Selection between these alternative pronunciations is automatically performed during the alignment process.

5.4 Generation of Alignments

The phone-level alignment of the acoustic waveforms in the training set to their word-level transcriptions involves both the selection of the most suitable pronunciation for each word from the alternatives provided in the phonetic dictionary, and the determination of the most probable locations for the corresponding phonetic boundaries. In this case the alignments were automatically determined using HTK, and in this section both the choice of the HMMs used to perform this alignment and the alignment process itself are described.

[9] Although the flap `/dx/` is assumed here to comprise a closure and subsequent release which are acoustically dissimilar, in practice this phoneme is articulated as a fricative, and the data for the phones `/dx.1/` and `/dx.2/` are recombined when training the acoustic models on the UW data, as described in Section 7.3.1.

[10] Examples of each of these sounds as they appear in English words, along with their corresponding International Phonetic Association (IPA) symbols, are given in Appendix A.

5.4.1 Model Set

To construct a set of HMMs which model the speaker-dependent acoustic data in the UW corpus, the following characteristics must be specified:

- The number of models to be used and the lexical tokens which they represent.
- The composition of these models in terms of a discrete set of "states".
- The number of parametric output distributions to be associated with each of these states.
- The set of features to be modelled by these distributions.

The process of model building can be accelerated by using an existing model set as a starting point and re-estimating the parameters of these models on the new data. This approach was adopted here by modifying a set of models which had originally been trained on the speaker-independent portion of the Resource Management (RM) corpus[11]. In this section the selection of a suitable initial RM model set is described, along with the techniques subsequently used to re-estimate its parameters on the speaker-dependent UW data.

Model Type and Number of Parameters

The choice of the number and type of HMMs to be used is constrained by the relatively small amount of data available in the UW corpus. Since the parameters of the initial RM models are to be re-estimated using the UW data, the total number of parameters used must be relatively small. In practice, this constraint prevents the use of triphone HMMs, and a set of multiple-mixture monophone models was used instead.

The use of monophone models is expected to yield a less accurate model of acoustic variations due to the phonetic context than would a triphone system trained on a larger data set. It is hoped however, that by using 5 Gaussian distributions to model the acoustic observations in each state of the monophone models, some of the variations due to the phonetic context will be characterised. This will be achieved if different Gaussian mixtures are used to model the acoustic observations according to the phonetic contexts in which the monophone model appears.

Feature Set

A total of 39 acoustic features were modelled by the Gaussian distributions in each HMM state. The first 13 of these comprised 12 MFCC values along with a normalised log energy coefficient, computed from the windowed acoustic segment being modelled. In addition, 13 parameters representing the differences between successive MFCC vectors are

[11] These models were chosen since the RM corpus phoneme set listed in Table 5.4 also forms the basis for the UW model set. A more detailed description of the RM corpus itself can be found in Chapter 6.

provided ("delta" parameters), as are the differences between successive delta parameters ("acceleration" parameters)[12]

Model Composition

As discussed in Section 5.3.2, the phone set used to transcribe the UW data is obtained from the RM phoneme set listed in Table 5.4 by deleting the closure phonemes `/dd/`, `/kd/`, `/pd/` and `/td/`, and sub-dividing the stop and diphthong phonemes into their component phones. Since the speaker-independent monophone RM HMMs used as an initial model set correspond to the phonemes listed in Table 5.4, these models must first be modified to permit the phones in Table 5.5 to be identified during the alignment process. This is achieved by an approximation which involves a simple modification to the original RM models.

Each HMM in the original RM set contains three "emitting" states[13]. If the three-state HMMs for stops and diphthongs are replaced with two-state models and the HMMs are then re-trained, the acoustic output vector distributions of these two states are observed to model the initial and final voiced sections of diphthongs, and the closure and release sections of stops respectively. By retaining information regarding the alignments of these individual HMM *states* during the alignment process, the approximate locations of the phone-level boundaries within stop and diphthong phonemes can be deduced.

Since the release of each of the stops except `/ts/` is optional, the HMM state sequences for these stops are also modified to allow transitions from the initial closure state straight through to the first state of the following phoneme, thus optionally by-passing the state modelling the release of the stop[14]. In the models for diphthongs and `/ts/` however, each of the two component states must be used to model at least one frame of acoustic data.

Background Models

As described in Section 5.2.2, a tone is played at the start of the recording of each utterance, and background comments such as "good" and "rep" are present at the end of many utterances, which would interfere with this alignment process. Separate three-state monophone HMMs corresponding to this tone and to each of the two most common background comments were therefore optimised, using parameterised speech vectors extracted from examples of each sound which were identified by hand. While the tone waveform is spectrally static and therefore well represented by a single HMM, the use of a single model for the background words "good" and "rep" is a simple approximation which is used to ensure that non-silent frames following the end of an utterance are not considered part of the utterance's transcription.

---

[12] The use of these latter 26 parameters represents an attempt to model the short-term acoustic context, as described in Section 1.2.2.

[13] Emitting states are those states in the model which have acoustic output distributions associated with them, as opposed to the "entry" and "exit" states which are used to link successive HMMs together.

[14] This model for stops is similar to one proposed by Lee [93].

Model Optimisation

The initial speaker-independent 5-mixture monophone RM model set was modified as described above, to provide two-state models for stops and diphthongs. The parameters of these models were re-estimated on the speaker-independent RM data, and combined with the 5-mixture monophone models for the background tone, and "good"/"rep" comments. The resulting models were then used to perform an initial alignment of the UW acoustic data to the corresponding hand-edited transcriptions.

This alignment was subsequently used as the basis for training a set of *speaker-dependent* 5-mixture monophone models on the UW data. In this case however, the initial model set comprised the 5 mixture background models together with *single* mixture speaker-independent monophone RM models, with the usual two-state stop and diphthong representations. While it would be simpler to start with 5-mixture speaker-independent RM models and simply re-estimate their parameters using the initial alignment described above, improved results are obtained by starting with a single-mixture model set, and then increasing the number of mixtures during the re-estimation process.

Accordingly, 5-mixture background models and single-mixture RM models were initially used, and training consisted of alternately re-aligning the data and re-training the models with an increasing number of mixtures[15], until a set of 5 mixture speaker-dependent HMMs optimised on the UW corpus were obtained.

5.4.2 Alignments

The speaker-dependent UW models were used with the word-level transcriptions and the phonetic dictionary to determine the optimum state-level alignments of the UW data using HTK. The word-level transcriptions include the specification of a mandatory tone at the start of each utterance, and optional trailing background comments. The state-level information is used to distinguish between the component phones of stops and diphthongs, but is disregarded in the case of other phonemes.

An example alignment of the spectrogram for the phrase "put these two" spoken by `jw18` is shown in Figure 5.1. The phonetic transcription is listed above the utterance, and the log energy at each frequency and time is proportional to the darkness of the plot.

The stop burst for the first occurrence of `/t/` at $\approx 0.3s$ has been omitted by the speaker, although the characteristic high-frequency noise pattern for `/t.2/` can be seen in the second example of `/t/` at $0.7s$.

While the alignments are generally quite good, errors are made for some phonemes such as `/dh/` at $\approx 0.4s$ and `/z/` at $\approx 0.6s$ in this example, each of which has a longer marked duration than is actually the case. The consequences of alignment errors such as these when generating synthetic spectral vectors for performing speech recognition are discussed further in Section 8.3.2.

---

[15] Only the number of mixtures in the RM models was increased at each stage, as the 5-mixture background models had already been optimised on the UW data.

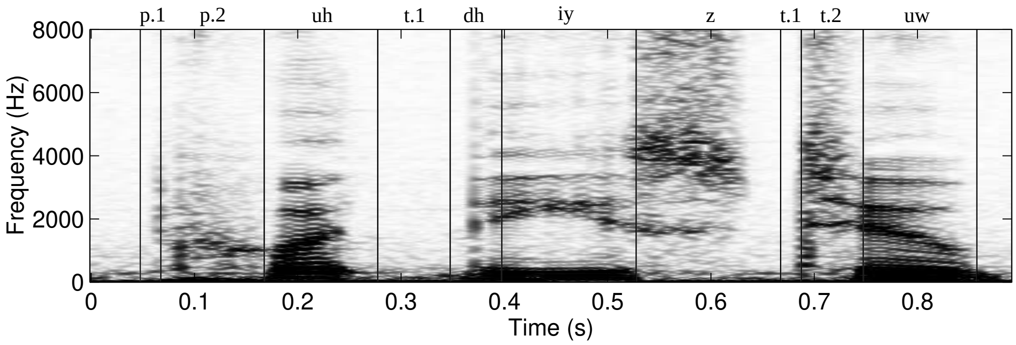

Figure 5.1: Phonetic alignment of the spectrogram for the phrase “put these two” for the speaker jw18. Automatically-generated boundaries between phonemes are indicated by solid vertical lines.

Chapter 6

Synthetic Data: Resource Management

6.1 Introduction

While articulatory databases such as that described in the preceding chapter are extremely useful for developing models of articulatory movements, in general such data are not available and acoustic training data alone must be used. One solution to this problem is to extract a set of statistics from an X-ray articulatory data set which are averaged over many speakers, and which therefore represent a speaker-independent description of articulation, which can be used to infer approximate articulatory movements for new speakers (Sections 7.2.1 and 7.4).

An alternative approach which has been used by many researchers for obtaining an articulatory representation, is to infer articulatory movements and/or vocal tract shapes directly from the acoustic signal (Section 2.2.2). This chapter describes the implementation of such a system, which uses an explicit vocal tract model to construct a codebook of [articulatory vector, acoustic vector] pairs, and which employs dynamic programming to predict articulatory trajectories given an acoustic vector sequence.

Many different approaches to vocal tract simulation have been described in the literature (Section 2.3.4), most of which require the specification of a vocal tract *cavity* shape. Since an articulatory representation is ultimately required to train the production model described in Chapters 3 and 4 however, an initial specification in terms of a set of articulatory variables is employed in this case. These are used to predict a corresponding vocal tract shape, through which the appropriate excitation signals are propagated. The components of the vocal tract model are therefore as follows:

- A set of articulatory parameters which are able to adequately describe the range of vocal tract shapes observed during the production of speech.
- A mapping from these articulatory parameters to the corresponding vocal tract shapes.
- Suitable excitation sources for the production of both voiced and unvoiced phonemes.

- A model for simulating the propagation of these excitation signal(s) through the specified vocal tract cavity.

Since the goal of implementing such a system is to provide data from which to train a self-organising production model, and not the detailed study of vocal tract simulation techniques, each of these components was implemented by adapting systems or sub-systems previously described in the literature. Section 6.2 describes the input acoustic data taken from the Resource Management database, along with the techniques used to train a set of HMMs on this data and automatically align the waveforms at the sub-phonemic level. The detailed implementation of the vocal tract model used to generate acoustic waveforms from a specification of vocal tract cavity shapes and excitation sources is developed in Appendix C.

Section 6.3 subsequently describes the construction of the codebook using this model, starting from a set of 7 synthetic articulatory parameters from which vocal tract cavity shapes are explicitly derived. Finally, Section 6.4 presents the dynamic programming algorithm used to infer synthetic articulatory trajectories from acoustic vector sequences using this codebook.

6.2 Model Set and Alignments

The acoustic data were taken from the speaker-dependent portion of the Defence Advanced Research Projects Agency[1] (DARPA) Resource Management corpus. Data from one male and one female speaker were used, whose backgrounds are detailed in Table 6.1.

Speaker	*Sex*	*Dialect*	*Age*
das1	F	Northern	27
tab0	M	Western	26

Table 6.1: Resource Management subject details.

The training and test sets for each speaker comprise 600 and 100 sentences respectively, drawn from a vocabulary of 991 words. The acoustic waveforms were recorded at a sampling frequency of $16kHz$, and were parameterised into both 12-dimensional MFCC vectors and 24-dimensional log Mel-frequency log spectral vectors in an analogous manner to the parameterisation of the University of Wisconsin (UW) data set described in Section 5.2.2.

Model Set

The aim of training a set of HMMs on this data—as was the case for the UW dataset—is to achieve the best possible acoustic modelling accuracy against which to assess the performance of the SPM. The availability of a larger training data set in the RM corpus enables a set of single mixture triphone models to be used here, in place of the

[1] Now known as the Advanced Research Projects Agency.

5-mixture monophone models described in Section 5.4.1. Following the HTK Version 2.0 RM toolkit recipe [176], a set of single mixture monophone HMMs were trained on the speaker-*independent* portion of the RM training data; these basic models were then cloned to form a set of single-mixture cross-word triphones. The phoneme set used as the basis for these HMMs was not the standard RM phoneme set, but was the same as that used for the UW data, as described in Section 5.3.2 and Appendix A.

The result of this process is a very large set of triphone models, the parameters of which cannot be reliably estimated due to a lack of sufficient training data. By applying the technique of state clustering however, in which each of the states assigned to any given cluster share a common output distribution, it is possible to avoid this problem while making efficient use of the parameters of the system. A decision tree-based clustering technique was used [118, 177], in which all triphone contexts are initially grouped together, and are then split into smaller clusters on the basis of questions regarding the phonetic context. The cluster splits are selected such that the likelihood of the training data is maximised, while ensuring that sufficient training examples are available to train each clustered set. This technique has the additional advantage of enabling the use of cross-word triphones, since models for previously unseen triphones can be synthesised using the same tree-based categorisation process.

The parameters of these speaker-independent RM cross-word tree-clustered triphones were then separately re-estimated on the 600 speaker-*dependent* training sentences for each of the two speakers `das1` and `tab0`. This yielded a set of speaker-dependent models suitable for use in aligning the training data and recognising the test data at the sub-phonemic level. In each case 39-dimensional MFCC vectors incorporating normalised log energy, delta and acceleration parameters were used as described in Section 5.4.1, and a total of 5654 physical triphone models—representing 87077 logical models—were trained for each speaker. These models were used to perform a forced state-level Viterbi alignment of the training data to the supplied transcriptions, in an analogous manner to that used for the UW data set[2].

6.3 Articulatory-Acoustic Codebook Construction

In this section, a set of synthetic articulatory parameters are presented which are used to generate oral and pharyngeal tract cavity shapes[3]. Once these shapes have been specified, along with the size of the opening at the velum and the presence or absence of voiced and/or fricated excitation, the vocal tract model described in Appendix C is used to generate the corresponding acoustic waveforms. By sampling the space of the articulatory parameters to generate a range of possible tract shapes to which different combinations of excitation signals are applied, a codebook of [articulatory vector, acoustic vector] pairs is constructed. This codebook can then be used to infer the approximate input articulatory configurations which would produce a given acoustic output.

[2]In this instance however, it was assumed that the correct transcriptions had been supplied and that the recordings were free of extraneous sounds.

[3]These two will collectively be referred to as the "oral" tract in this section for brevity.

Finally, dynamic programming is used to find the optimum paths through articulatory parameter space corresponding to acoustic vector sequences, using a cost function which incorporates both the acoustic and articulatory (or "geometric") errors. In this way a set of synthetic articulatory trajectories can be obtained for an arbitrary speaker from the acoustic signal alone, circumventing the need for X-ray data.

6.3.1 Synthetic Articulators

The articulatory model used is a modified version of that originally described by Meyer et al. [108]. Their vocal tract model consists of an oral tract whose cross-sectional area varies with time and with distance from the glottis, and a nasal tract with fixed cross-sectional shape which is coupled to the oral tract at a point approximately $7cm$ from the glottis.

The shape of the oral tract is defined by a set of articulatory variables, which are identified by parameterising area functions corresponding to X-ray images of the articulation of vowels and consonants. In all, four parameters are used to control tongue shape, two lip control variables are provided, and one parameter governs the size of the opening at the velum (the coupling point of the oral and nasal tracts).

The first two articulatory parameters control the tongue body shape, and are found by taking the eigenvalue decomposition of the X-ray area functions, and extracting the first two eigenvectors. The resulting tongue body areas are then given by:

$$\boldsymbol{\mathcal{A}}_{\text{body}} = \boldsymbol{R} + a_1 \boldsymbol{V_1} + a_2 \boldsymbol{V_2} \tag{6.1}$$

where $\boldsymbol{\mathcal{A}}_{\text{body}}$ is the piecewise constant area function approximating the cavity shape, computed by adding a linear combination of the first two eigenvectors $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ to the mean vocal tract shape, $\boldsymbol{R}$. Each of these vectors comprises nine segments corresponding to the cross-sectional areas of the vocal tract between the glottis and the lips, as shown in Figure 6.1. The weights a_1 and a_2 are the first two articulatory parameters, and are interpreted by the authors as controlling front-back and up-down movements of the tongue body respectively.

Two further tongue control parameters, a_3 and a_4, are then used to model vertical and horizontal movements of the tongue tip. This is achieved by computing a separate nine-segment function $\boldsymbol{\mathcal{A}}_{\text{tip}}$ to be added to $\boldsymbol{\mathcal{A}}_{\text{body}}$, which takes values close to zero near the glottis, but which defines constrictions of varying place and degree near the lips[4]:

$$\begin{aligned}
\boldsymbol{\mathcal{A}}_{\text{tip}}(n;\alpha,\beta) &= \frac{-2\alpha}{3(\beta-3)} \cdot \frac{1}{f^2(n)+1} && (6.2)\\
f(n) &= \exp\{\gamma(n-\beta)\} + \gamma(n-\beta) - 1, \quad \text{for } n = 1,\ldots,9 && (6.3)\\
\alpha &= 0.8(a_3 + 2.6) && (6.4)\\
\beta &= 0.03a_4 + 8.595 && (6.5)\\
\gamma &= \tfrac{2}{9}\beta - 1 && (6.6)
\end{aligned}$$

[4] On page 525 of Meyer et al. [108], Equation 6.5 is incorrectly given as $\beta = 0.03a_4$, and the caption in Figure 2 on the same page should read: "Top: $a_4 = -2.5$, bottom: $a_4 = 2.5$."

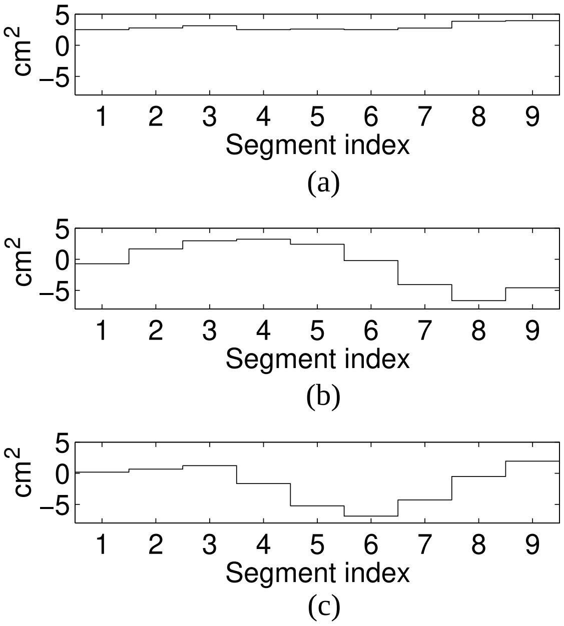

Figure 6.1: Piecewise constant vocal tract area functions from the lips (1) to the glottis (9) representing (a) Mean, or "neutral" vocal tract areas (b) First eigenvector (c) Second eigenvector.

A tenth area segment corresponding to the size of the lip opening is directly controlled by a fifth parameter, a_5, and in the original model lip protrusion and the opening at the velum are controlled by two further variables, a_6 and a_7. The authors used a fixed vocal tract length of $17.5cm$, comprising these ten area segments each of which had a uniform cross-sectional area and a length of $1.75cm$. This simplified model was chosen for its ease of implementation, but it suffers from the following disadvantages:

1. The fixed tract length means that the parameter a_6—which nominally controls lip protrusion—must actually be used to define an additional impedance at the lips to simulate the effects of such a protrusion, which is not permitted by the model.

2. The use of only ten fixed cross-sectional area segments to approximate the smoothly-varying area function of the human vocal cavity limits the ability of the model to produce waveforms with the desired spectral resolution, as the length of these segments determines the bandwidth of the signal produced[5].

These difficulties are alleviated by decreasing the length of the fixed cross-sectional area segments, and modifying the original model to permit a variable number of segments in the oral tract. In this approach, the parameter a_6 is now used to represent this overall tract length, rather than the lip protrusion in particular[6]. The seven articulatory parameters

[5]The relationship between segment length and signal bandwidth is described in Section C.4.1.

[6]These two quantities are closely related, since much of the variability in oral tract length is accounted for by lip protrusion and retraction. Other factors which influence the length include the horizontal extension of the jaw and the vertical displacement of the vocal cords.

used and their approximate interpretations are summarised in Table 6.2.

Parameter	*Interpretation*
a_1	Front-back tongue body movement
a_2	Vertical tongue body movement
a_3	Vertical tongue tip movement
a_4	Front-back tongue tip movement
a_5	Lip opening
a_6	Oral tract length
a_7	Velum opening

Table 6.2: Synthetic articulatory parameters and their interpretations.

6.3.2 Generation of Vocal Tract Area Functions

Equations 6.1 to 6.6 were used to generate nine-segment oral tract area functions (excluding lip area), using 12 different quantised values for each of the parameters a_1 to a_4. Since many parameter combinations lead to unrealistic tract shapes, area functions with a minimum area less than $0.01cm^2$ or a maximum area greater than $20cm^2$ were discarded. The maximum area requirement imposes physical limits on the cavity size, while the use of a minimum area greater than zero was chosen to exclude stop phoneme closures and nasals. This latter restriction was imposed as stop phoneme closures are not included in the codebook (Section 6.4.2), and the effects of the closed oral cavity are neglected during the synthesis of nasal waveforms.

The result was a set of 6321 oral tract shapes, which were then quantised to remove area functions which were extremely similar to one another[7], using a logarithmic quantisation similar to that described by Fant [43]. Taking k as the index of the quantised values Q, then:

$$
\begin{aligned}
Q_k &= k_0\,(e^{\lambda k} - 1), \quad \text{for } k = 1, \ldots, N && (6.7)\\
\lambda &= \frac{1}{N}\,\log\left(\frac{A_{\max}}{k_0} + 1\right) && (6.8)
\end{aligned}
$$

where $A_{\max}$ is the maximum permissible area in cm^2, N is the total number of quantisation steps, and the constant k_0 controls the shape of the curve, which is chosen to match Fant's data. The values of these parameters were:

$$A_{\max} = 16cm^2 \tag{6.9}$$

[7]Other authors, when constructing codebooks such as that described in this section, have initially retained all the area functions produced in this way, and then generated the corresponding acoustic vectors before applying a quantisation in the frequency domain [151]. Since quite different articulatory configurations can result in very similar acoustic outputs however (Section 6.4.1), a quantisation in the articulatory domain, which takes into account the non-linear relationship between articulatory positions and acoustic output is preferred.

$$N = 64 \tag{6.10}$$
$$k_0 = 0.5 \tag{6.11}$$

The resulting logarithmic quantisation curve is shown in Figure 6.2. For each quantised articulatory shape vector $\boldsymbol{q}_i$, if there are m_i synthesised articulatory shape vectors which are closer (in the logarithmic domain) to $\boldsymbol{q}_i$ than to any other quantised vector $\boldsymbol{q}_j$, only that synthesised articulatory vector which is closest to $\boldsymbol{q}_i$ is retained.

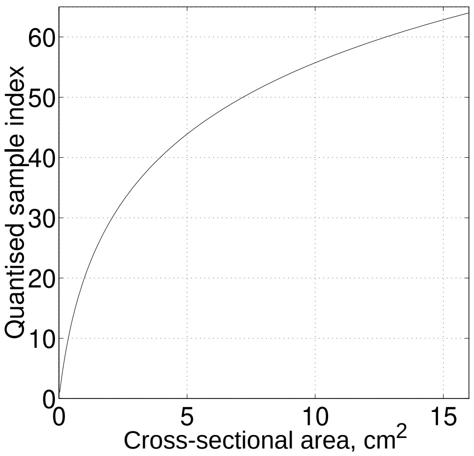

Figure 6.2: Logarithmic quantisation curve for area function data.

The non-linear quantisation used reflects the fact that changes in oral tract areas will have a greater effect on the resulting acoustics when the absolute size of the area concerned is small. For example, during the production of the fricatives `/s/` and `/sh/`, only a slight difference in tongue tip position at the point of maximum constriction (and hence minimum oral tract area) leads to sufficient spectral distortion to ensure discrimination between the two sounds. Where the area function is relatively large during the production of these sounds, eg. in the mid and back-tongue sections, considerable variation in articulatory positioning is possible without greatly affecting the spectral characteristics of the output.

Approximately half of the oral tract shapes were removed by this quantisation process, leaving 3309 distinct area vectors. To each of these a quantised set of lip opening values was added, representing a tenth oral tract segment as in Meyer's system. The logarithmic quantisation scale defined in Equations 6.7 and 6.8 was used once again, but in this case the number of distinct quantisation values, N, was set to 12 and the first two resulting values were discarded to enforce a lower bound on lip area of $\approx 0.7cm^2$, with a maximum of $16cm^2$ as before. A smaller number of quantisation values was used at the lips than for the computation of the other areas, since most lip area values lead to valid oral tract shapes, and larger values of N lead to the generation of excessive numbers of oral tract shapes. Lip area values were considered plausible if they fell within $10cm^2$ of the penultimate area function value, giving a total of 27651 basic shapes.

To implement a variable oral tract length, the resulting 10-dimensional vectors were interpolated and re-sampled twice in the logarithmic domain, to yield sets of 15 and 16-dimensional oral tract area function vectors respectively. As described in Section C.4.1, the length of each re-sampled segment was taken to be $\approx 1.1cm$, so that the corresponding oral tract lengths were $\approx 16.4cm$ and $17.5cm$ respectively, giving rise to 55302 distinct oral tract shapes.

Voiced and unvoiced fricatives were generated from those oral tract shapes which contained a constriction of less than $0.3cm^2$ situated between the second section following the glottis and the penultimate section before the lips—a restriction that reduced the number of distinct oral tract shapes for fricatives to 15942. Finally, vocal tract shapes for nasals were defined by neglecting the oral branch of the tract, and connecting the nasal tract area function[8] at the end of the pharyngeal branch, approximately $8cm$ from the glottis. These two branches were connected via an additional tube segment representing the opening at the velum, the area of which took each of the quantised values 1, 3, and $5cm^2$. Since there were 5042 different pharyngeal tube shapes, this gave a total of 15126 tract shapes for nasal sounds.

6.3.3 Codebook Construction

Using the vocal tract model described in Appendix C, acoustic waveforms for each of the vocal tract shapes in Section 6.3.2 were computed, for input excitation signals 21 pitch periods in length. In each case, the resulting waveforms were parameterised[9] into 12-dimensional Mel-frequency cepstral coefficients (MFCCs), using a Hamming window of length $25msec$ and a step size of $10msec$ between frames[10]. This parameterisation was chosen as it yields a compact representation of the speech spectrum, and hence provides a smaller search space for the dynamic programming algorithm described in Section 6.4.2. The final breakdown of the articulatory-acoustic codebook is given in Table 6.3.

Sound type	*No. of vectors*
Non-fricated voiced sounds	55302
Voiced fricatives and bursts	15924
Unvoiced fricatives and bursts	15924
Nasal sounds	15126
Total	**102276**

Table 6.3: Breakdown of vectors in articulatory-acoustic codebook.

6.4 Articulatory-Acoustic Mapping Inversion

The inversion of a mapping from either articulatory parameters or vocal tract area functions to acoustic vectors is a frequently-encountered problem in speech research. In this

[8] The area function of the nasal tract is illustrated in Figure C.4 in Section C.4.2.

[9] A more detailed description of this parameterisation process can be found in Section 4.2.1.

[10] A pitch-synchronous encoding of the acoustic waveform could be obtained instead, by setting the step size to $1/F_0$ where F_0 is the pitch period of the voiced excitation. This has very little effect on the resulting MFCCs however, provided the number of pitch periods synthesised, $N_p \gg 3$.

section, a brief review of existing inversion techniques is provided[11, 12]. The dynamic programming approach taken in this dissertation is then described, and an example of a synthetic articulatory trajectory generated in this way is presented.

6.4.1 Inversion Techniques

One of the central difficulties with any "inverse" mapping—one which takes acoustic vectors at its inputs and predicts the values of the corresponding articulatory parameters at its outputs—is the fact that quite different articulatory configurations can give rise to very similar acoustic outputs [4, 94]. This implies that attempts to model the inverse transformation using acoustic error alone [3, 61, 86, 122] are likely to produce discontinuous articulatory trajectories. A continuity constraint should therefore be applied to such trajectories, which may be implicit as in inverse filtering techniques [114, 115, 169], or explicitly imposed via a restriction to critically damped second order transitions [155] or the minimisation of geometrical distances [151, 157, 178].

In addition, the non-linearity of the inverse mapping, when combined with its non-uniqueness, can result in non-convex target regions in articulatory space [53, 73], as illustrated in Figure 6.3.

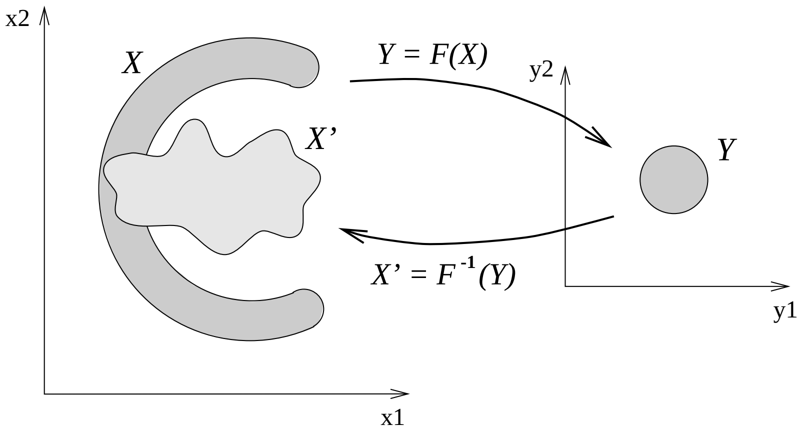

Figure 6.3: Hypothetical many-to-one non-linear mapping $F()$, illustrating the potential drawback of using an averaging technique in the inverse mapping $F^{-1}()$. If each point in Y corresponds to many points distributed throughout X, then an inverse mapping $F^{-1}()$ which averages these points may yield an inverse image X' which bears no resemblance to the original region X.

In this figure, the region X is mapped by the non-linear many-to-one function $F()$ onto the region $Y = F(X)$. An inverse mapping $F^{-1}()$ is then hypothesised, which maps Y back into $\mathcal{X}$-space, where each point in Y maps to a series of points with distributed locations in X. Then if $F^{-1}()$ performs an *averaging* function on the images of the points from Y in

[11] Many inverse mappings have been developed explicitly for the purpose of performing speech recognition, and these were discussed in Section 2.2.2 [37, 122, 134, 136, 137, 138, 155, 178].

[12] A more detailed description of the various techniques which have been proposed in this area can be found in the review articles by Strube [162] and Schroeter and Sondhi [151].

$\mathcal{X}$-space, the inverse image $X' = F^{-1}(Y)$ may bear no resemblance to the original region X, as shown.

Thus gradient-based algorithms which average over a number of training vectors, whether a single neural network [3, 122, 178], Jacobian computation [74] or unconstrained optimisation [94], may converge to an average which does not lie within the target class, resulting in an incorrect inverse model. This problem can be avoided either by subdividing the input space into regions in which the non-linear mapping is unique [133], or by jointly optimising a [forward, inverse] model pair to restrict the inverse model to a particular solution [52, 123].

6.4.2 Dynamic Programming

The inversion technique implemented in this dissertation is similar to that proposed by Schroeter and Sondhi [150, 151], and employs a dynamic programming (DP) algorithm to compute the optimum articulatory trajectory corresponding to a given acoustic vector sequence using a cost function which incorporates both acoustic and articulatory, or "geometric", error terms. Since the DP algorithm is well described in the literature[13], only a very brief outline of the technique is provided here.

DP algorithm

Suppose that an acoustic vector sequence is provided, which is to be decoded in terms of a corresponding sequence of articulatory vectors. At each point during the decoding process a set of best "paths" are maintained, where a path comprises the sequence of indices which identify those codebook vectors which have resulted in the best match to the partial input acoustic sequence seen thus far.

In the general case N_c such paths must be maintained at any given time, where N_c is the number of vectors in the codebook, since a separate path is stored for the best index sequence currently culminating in each distinct codebook vector. Although some of these partial paths will have relatively high errors associated with them, in the optimum search strategy they are all retained since a path with a high intermediate error may eventually develop into the best solution.

At each stage of the DP algorithm, the next input acoustic vector is compared with each of the acoustic vectors in the codebook. In addition, the corresponding articulatory codebook vectors are compared with the final articulatory vectors in each of the active paths. Each path is then extended by one step, and the weighted sum of these two errors is added to the previous total path errors. Whenever two paths meet at the same "node" (where a node represents a codebook vector pair) the path with the higher cost is discarded, since it can never represent the optimum solution. This process continues until the final acoustic vector has been processed, at which time the best path can be traced back from the final frame to the first.

---

[13] For a description of the dynamic programming algorithm as applied to speech recognition, see for example Holmes [63].

Implementation

In practical implementations using large codebooks such as that described in Section 6.3, it is usually necessary to reduce the computational load by relaxing the constraint that all possible paths be maintained. One technique for achieving this is to use a "beam" search, in which paths whose intermediate error exceeds that of the current best path by a certain amount are discarded, on the basis that an intermediate path with extremely high error is very unlikely to develop into the optimum solution.

An alternative sub-optimal technique is to consider only the C codebook entries yielding the best acoustic match to the incoming vector when determining the succeeding vectors for each of the currently active paths. This approach was adopted here, where C was set to 500. In addition, the speed of the algorithm was greatly enhanced by restricting the codebook search according to the four broad sound categories (excluding silence) listed in Table 6.3. As an example, when an input acoustic vector was identified by automatic alignment as a voiced fricative, only the best 500 codebook pairs from the 15924 voiced fricative entries were considered in the algorithm.

Since suitable articulatory positions corresponding to periods of silence in the input cannot be determined using codebook lookup alone, the DP algorithm was applied only to the contiguous non-silent sections of the input speech, and articulatory parameter values in the intervening silence periods were determined by linear interpolation[14].

The cost function, E, used was a weighted sum of the acoustic and articulatory errors, given by:

$$E = K_m \sqrt{\sum_{i=1}^{N_m} \left(\frac{m_i - \mu_m}{\sigma_m} \right)^2} + K_a \sqrt{\sum_{j=1}^{N_a} \left(\frac{a_j - \mu_a}{\sigma_a} \right)^2} \tag{6.12}$$

where $\boldsymbol{m}$ is an MFCC vector of dimension N_m, $\boldsymbol{a}$ is an articulatory parameter vector of dimension N_a and K_m and K_a are the weighting constants for the acoustic and geometric errors respectively. Each term in the cost function was scaled by its global mean and standard deviation (μ_m, μ_a and σ_m, σ_a respectively) computed over all of the vectors in the codebook, so that the error computation was not dominated by contributions from parameters with large absolute values. Ideally, optimum values of the ratio $K_m : K_a$ could be determined according to the identity of the acoustic vector being considered at each step. In this case a fixed value was used however, which was empirically set to 0.2.

The use of codebook look-up in the algorithm guarantees that a particular inverse solution to the articulatory-acoustic mapping is chosen at each point in time, and the use of geometric constraints in the DP search ensures that articulatory continuity is maintained, thus satisfying the requirements set out in Section 6.4.1.

[14] Since the purpose of performing the codebook inversion is to obtain a set of synthetic articulatory trajectories which can be sampled at points corresponding to the midpoints of phonemes, any inaccuracies in the positions of the articulators during periods of silence will not be significant.

Synthetic Articulatory Trajectories

The MFCC vectors corresponding to the 600 training sentences for each speaker were used as inputs to the DP algorithm, which was used to decode articulatory trajectories for the first 5 articulatory parameters in Table 6.2, which control the positions of the tongue and lips.

The parameter a_6, which controls the length of the oral tract, was excluded from the inversion process since its value is binary, merely indicating whether the tract contained 15 or 16 equal length segments. As a result, generating a sequence of values for a_6 corresponding to an acoustic input file, and sampling the resulting trajectories at the midpoints of phonemes as described in Section 3.2.2 would not lead to readily interpretable results. The exclusion of this parameter is unlikely to have a significant impact on the performance of the production model however, as the acoustic mapping used is phoneme-specific. Since a separate mapping from articulatory to acoustic space is provided for each phoneme, the synthesis of spectral values for phonemes which require a lengthening of the oral tract will automatically take this effect into account, as a longer tract will implicitly be specified for all the articulatory data for these phonemes[15].

Similarly, the parameter a_7—which controls the size of the opening in the velum—is also excluded from the inversion algorithm. While the value of this parameter at the midpoints of phonemes *can* be sampled and interpreted as a continuous variable, these values will be non-zero only for the phonemes `/en/`, `/m/`, `/n/` and `/ng/`, since the nasalisation of vowels or fricatives is not permitted by the model. As a result, the co-articulation model described in Chapter 3 will be unable to predict suitable variations in the value of this parameter during the production of nasals, and as in the case of a_6 the specification of a mean value is superfluous if this value is constant over all examples of these phonemes.

An example of a decoded articulatory trajectory corresponding to movements in the articulatory parameter a_4 for the phrase "the constellations" for the speaker `tab0` is shown in Figure 6.4. In this figure, phonetic boundaries taken from the HTK-produced label file (Section 6.2) are marked as dotted vertical lines, and the phonetic labels are derived from the DARPA-supplied word-level transcription of the utterance.

[15] This ignores any contextual effects whereby the amount of oral tract lengthening may be a function of the neighbouring phonemes. Degrees of partial lengthening such as this cannot be specified in the vocal tract model however, the length of which is quantised to two distinct values.

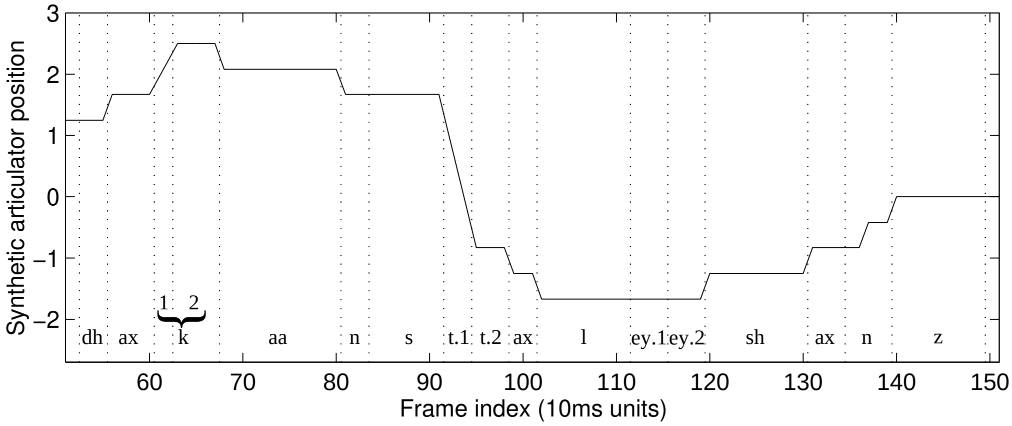

Figure 6.4: Synthetic articulatory trajectory for a_4 during the phrase "the constellations" for the speaker `tab0`. The automatically-generated boundaries between phonemes are indicated by dotted vertical lines.

Chapter 7

Production Model Evaluation

7.1 Introduction

This chapter presents an evaluation of the performance of the self-organising production model when applied to data from both the University of Wisconsin and Resource Management corpora. Sections 7.2 and 7.3 provide an analysis of the performance of the articulatory and acoustic models developed in Chapters 3 and 4 respectively, as applied to the University of Wisconsin database. In Section 7.2, the utility of the articulatory sampling technique and associated co-articulation model are assessed by comparing the predicted synthetic articulatory trajectories with the corresponding X-ray traces. In addition, a method for obtaining a set of speaker-independent articulatory statistics is presented, which is used as an alternative technique for synthesising articulatory trajectories for the acoustic data from the Resource Management corpus.

Section 7.3 describes the detailed implementation of the artificial neural network-based acoustic model of Chapter 4, and presents the results of using both this model and linear regression to predict normalised log spectral vector sequences from both synthetic and X-ray articulatory trajectories. Finally, in Section 7.4, the performances of both the articulatory and acoustic models when trained on data from the Resource Management corpus are assessed. A technique for re-estimating the synthetic articulatory statistics resulting from the codebook inversion algorithm of Section 6.4 is presented, and results are described for the use of both this re-estimated articulatory set and the speaker-independent X-ray articulatory data as a basis for synthesising acoustic vector sequences.

7.2 Articulatory Model: UW Data

In this section the results of using the articulatory positional sampling technique described in Section 3.2.2 are presented, and the relative discriminatory usefulness of the resulting statistics are examined for each of the 16 articulatory parameters. In addition to obtaining statistics describing articulatory movements for each of the six speakers studied, a set of "speaker-independent" statistics are computed from the data for all six speakers.

The validity of the use of Gaussian parametric models to represent both positional and curvature variations is subsequently assessed using the Kolmogorov-Smirnov statistic. The utility of the curvature measure in the prediction of articulatory positional variations at the midpoints of phonemes is then examined, in terms of the correlations observed between these two variables. Finally, the accuracy of the articulatory trajectories synthesised both with and without the explicit co-articulation model is evaluated by comparing them with the corresponding X-ray articulatory traces.

7.2.1 Articulatory Positional Variations

As described in Section 5.2, approximately three quarters of the data from each speaker were used as a training set to optimise the parameters of the system, and the remainder were retained as an independent test set for evaluating the models.

The result of the positional sampling algorithm is a set of values describing articulatory positions at the midpoints of phonemes for each articulator and each phoneme. Both x and y values are sampled for each of the eight articulatory pellets to yield 16 sampled articulatory variables[1], where both the absolute magnitude of the variation observed and the discriminatory usefulness of this variation vary greatly from articulator to articulator.

Articulatory Ranges

The greatest range of movement is seen in the x, or horizontal position of the tongue tip, which has a maximum range of movement of $\approx 3cm$, as compared with horizontal movements in the incisors of the lower jaw which exhibit the least displacement, at only $\approx 3mm$. In general terms, the tongue variables and those describing vertical motion in the jaw and lower lip vary more widely than do those describing horizontal lip and jaw movements, as shown in Figure 7.1 for the speaker jw18.

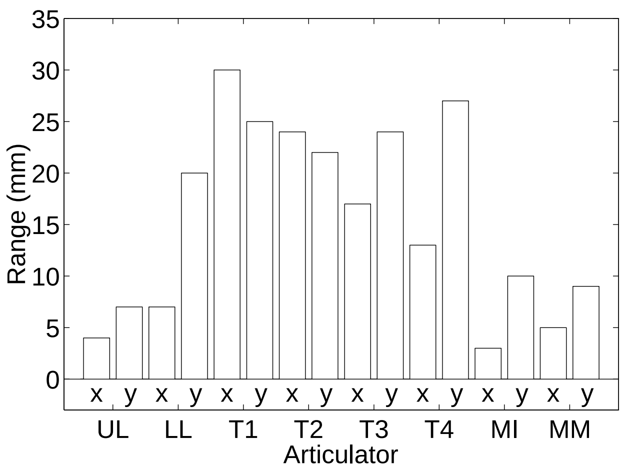

Figure 7.1: Bar graph showing approximate maximum ranges of articulatory movement for the speaker jw18.

[1] Except in the case of speaker jw29, for whom only 14 articulatory parameters could reliably be estimated, as discussed in Section 5.2.3.

Inter-Articulator Correlations

Although the model of articulatory movement treats the 16 articulatory parameters independently (Section 3.2.3), the positions of many of the articulators are actually highly correlated. Table 7.1 lists the inter-articulator correlation coefficients computed from the X-ray articulatory vectors corresponding to all of the phonemes in the training and test sets for `jw18`, excluding silences.

	ULx	***ULy***	***LLx***	***LLy***	***T1x***	***T1y***	***T2x***	***T2y***	***T3x***	***T3y***	***T4x***	***T4y***	***MIx***	***MIy***	***MMx***
ULy	**.61**														
LLx	**.41**	-.09													
LLy	.05	-.14	**.47**												
T1x	**-.27**	-.06	**-.30**	-.17											
T1y	**.24**	**.24**	.08	**.38**	**-.36**										
T2x	**-.22**	.00	**-.36**	-.16	**.95**	**-.26**									
T2y	.10	.16	-.02	-.03	-.08	**.46**	-.04								
T3x	**-.20**	.04	**-.37**	-.15	**.86**	-.10	**.94**	-.04							
T3y	-.03	.11	**-.20**	-.19	**.40**	-.03	**.48**	**.68**	**.41**						
T4x	-.15	.08	**-.33**	-.14	**.57**	**.24**	**.62**	**.36**	**.74**	**.33**					
T4y	-.01	.09	-.14	-.12	**.45**	-.12	**.54**	**.38**	**.48**	**.88**	.13				
MIx	.15	.03	**.55**	**.32**	.03	.00	.01	-.03	.02	-.07	.06	-.06			
MIy	**.22**	**.20**	**.27**	**.80**	-.05	**.51**	.02	.08	.03	-.03	.02	.03	.16		
MMx	-.08	.00	**.24**	.18	.04	-.02	.01	-.01	.06	-.06	.15	-.07	**.43**	.05	
MMy	**.28**	.14	**.27**	**.70**	-.09	**.47**	.00	.05	.01	-.05	-.03	.02	.13	**.86**	-.03

Table 7.1: Inter-articulator correlation coefficients, computed from X-ray articulatory positional vectors (excluding silence) in the training and test set files for `jw18`. Correlations with magnitudes of 0.2 or more are shown in bold.

Several trends can be observed in these correlation data. As expected, the horizontal tongue parameters *T1x* to *T4x* are highly (positively) correlated with one another, as are the corresponding vertical parameters *T1y* to *T4y*, although in the latter case the increased flexibility of the tongue tip in the vertical direction means that *T1y* is correlated only with *T2y*, and not with either of the posterior tongue height parameters. In addition, extension of the tongue tends to accompany a raising of the tongue dorsum (eg. during `/l/`), as evidenced by the positive correlations between each of *T3y* and *T4y*, and the parameters *T1x*, *T2x* and *T3x*.

The pellets *MI* and *MM* are strongly correlated in both the x and y directions due to being fixed to the rigid mandible, and in addition the height of the jaw (*MIy* and *MMy*) is correlated with the heights of the lower lip (*LLy*) and tongue tip (*T1y*). Similarly, the horizontal position of the lower incisors, *MIx*, is strongly correlated with that of the lower lip, *LLx*, as the lip is attached to the jaw near this point. In terms of the lip parameters themselves, the x and y positions of each of the upper and lower lips (*UL* and *LL*) are positively inter-correlated (each lip is elevated when extended), and their corresponding horizontal positions *ULx* and *LLx* are also positively correlated (both lips are usually extended together, eg. during `/w/`). Finally, these horizontal lip positions are *negatively* correlated with the horizontal position of the tongue (*T1x* to *T4x*), so that extension of the lips tends to coincide with retraction of the tongue (eg. during `/u/`).

The existence of these correlations indicates that alternative articulatory parameter sets could be derived from the X-ray data which might provide more efficient articulatory representations. Transformations of parameter space could also potentially provide a more natural framework for modelling compensatory articulations, in which movement in one articulator (eg. jaw height) compensates for inaccuracy in the position of a second articulator (eg. tongue height). Finally, modelling the movements of articulatory structures along axes other than the fixed horizontal and vertical directions used in this study should lead to a more accurate model of co-articulatory behaviour[2].

Discriminatory Usefulness

The degree to which an articulator's position is likely to discriminate between two different sounds varies greatly, with the articulators exhibiting the greatest range of movement also proving the most discriminatory. For example, the range of horizontal movement observed at the back of the jaw, *MMx*, is almost identical across all phonemes, with similar mean positions and a standard deviation of $\approx 1.2mm$ being observed in each case[3]. In fact, a value of $3.15mm$ for *MMx* in the data for speaker `jw18` falls within one standard deviation of the mean *MMx* position for *every* phoneme sampled, and hence the discriminatory power of this variable is extremely low[4].

By contrast, both the mean and standard deviation of the vertical positional samples for the tongue tip vary greatly from phoneme to phoneme, so that this articulatory variable is extremely useful for discriminating between phonetic identities. This is as expected, since the position of the tongue tip has a very strong influence on the shape—and hence the acoustic properties—of the vocal tract, as compared with the slight extension and retraction of the jaw. These differences in discriminatory ability are illustrated in Figure 7.2, which plots two standard deviations for each articulatory variable for the speaker `jw18`.

The first of these is the mean standard deviation for the variable concerned, measured at the midpoints of each phoneme p:

$$\sigma_{av} = \sqrt{\frac{1}{N_p} \sum_{p=1}^{N_p} \sigma_p^2} \tag{7.1}$$

where N_p is the total number of phonemes, and σ_p is the standard deviation of the articulator's positional samples at the midpoint of the phoneme p. The standard deviation of the distribution of the *mean* values of each variable at these midpoints is then computed as:

---

[2]The potential for future research in this area is discussed in Section 9.5.1.

[3]Although on average, the jaw extends slightly further during `/ch/`, `/f/` and `/sh/`, and retracts slightly for `/ey.1/` and `/ey.2/`.

[4]Hence it is expected that the absence of this articulatory parameter for `jw29` may not have a significant adverse effect on the performance of the model for this speaker.

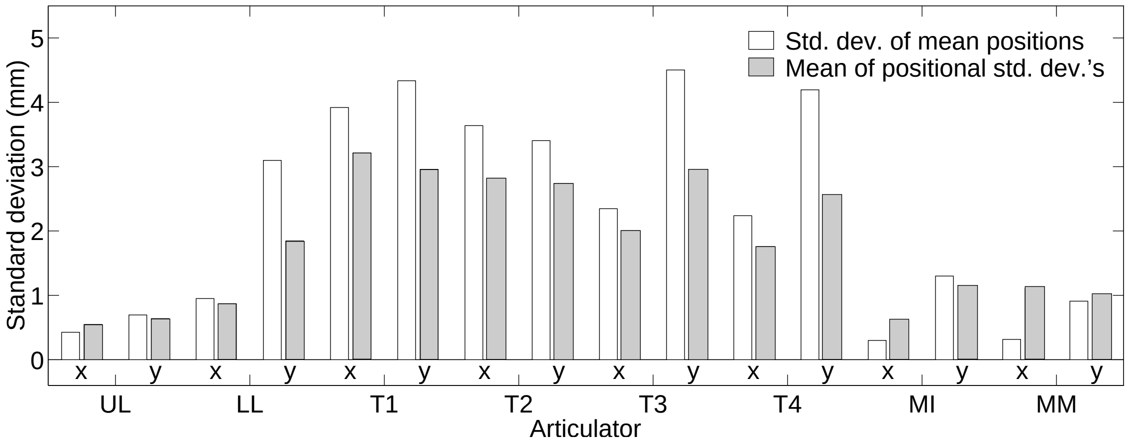

Figure 7.2: Bar graph showing comparisons of standard deviations of mean articulatory positions σ_μ (unshaded) against average articulatory positional standard deviations σ_{av} (shaded), as measured at the midpoints of phonemes for the speaker jw18.

$$\sigma_\mu = \sqrt{\frac{1}{N_p - 1}\sum_{p=1}^{N_p}(\mu_p - \mu)^2} \tag{7.2}$$

where μ_p is the mean value of the articulator's positional samples at the midpoint of phoneme p, and μ is the mean of the N_p mean positions:

$$\mu = \frac{1}{N_p}\sum_{p=1}^{N_p}\mu_p \tag{7.3}$$

Relatively large values of σ_μ (unshaded) indicate that the mean position of the articulator varies significantly from phoneme to phoneme. When the value of σ_{av} (shaded) is less than σ_μ as is the case for the tongue parameters, the articulator's positions are also relatively tightly constrained during the production of each individual phoneme, compared with the variations in mean position observed between different phonemes. This in turn means that on average, the positions of these articulators are relatively highly specified, and hence discriminatory between phonemes.

Where σ_{av} is relatively large compared with σ_μ however, the position of the articulatory variable concerned is less likely to be strongly correlated with the identity of the phoneme, as is the case for *MMx* as previously described. From Figure 7.2, the articulatory trajectories corresponding to the tongue, lower lip, *ULy* and *MIy* parameters are expected to be of greater relevance (on average) to the computation of acoustic outputs than those corresponding to the remaining parameters.

Speaker-Independent Statistics

Although a central goal of this dissertation is the development of a set of speaker-dependent models of speech production from X-ray articulatory data, the ability to extend the technique to speakers for whom such data are not available is highly desirable, as discussed in Chapter 6. One approach to providing synthetic articulatory data for such speakers, is to develop a set of *speaker-independent* statistics from the available X-ray data, and use these as a set of bootstrap parameters from which to train models for the new speakers.

This can be achieved by combining all of the articulatory data from each of the six UW speakers[5] sampled at the midpoints of each phoneme p. Since the geometry of each speaker's vocal apparatus is variable, as is the exact placement of the reference pellets on the subject's head during recording (Section 5.2.3), the articulatory samples taken from each speaker must be normalised before they can be meaningfully combined in this way.

Each articulatory parameter value a was therefore scaled by its mean μ_a and standard deviation σ_a, computed over the entire data set for the speaker concerned:

$$a_{\text{scaled}} = \frac{a - \mu_a}{\sigma_a} \tag{7.4}$$

Since different speakers may use significantly different articulatory strategies during speech production, this scaling procedure does not guarantee that the combined articulatory statistics describe plausible articulations. Some measure of confidence in the resulting values can be gained however, by comparing the scaled parameter distributions for the various speakers. Figure 7.3 shows a comparison of the means and standard deviations of the normalised positions of the articulatory parameter *T2y*, during production of the first seven phonemes[6] by the six speakers in the UW database[7].

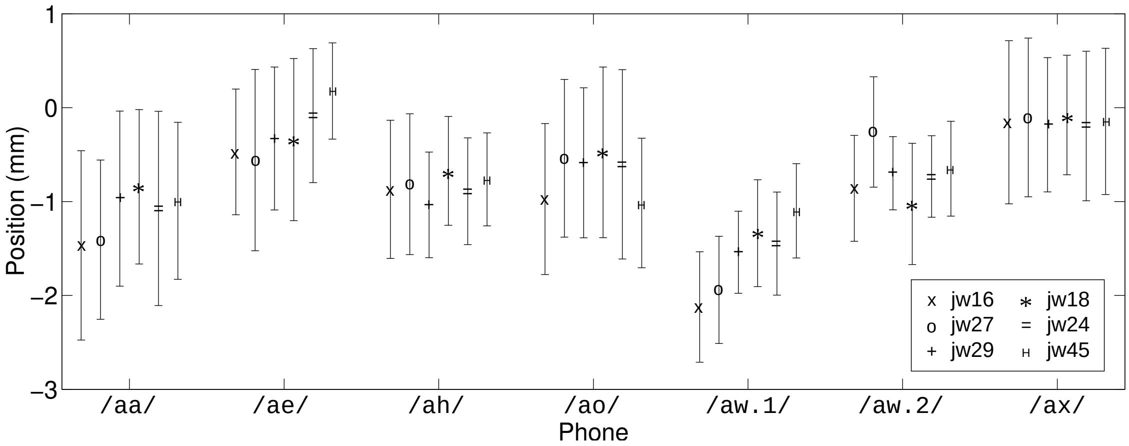

Figure 7.3: Plot showing normalised mean and standard deviation statistics for the articulatory parameter *T2y* during production of the first seven phonemes.

As can be seen from the figure, there is reasonably good agreement between the normalised parameters across the six speakers for this articulator and phoneme subset. While

[5]With the exception of the articulatory parameters *MMx* and *MMy*, for which the data from the 5 speakers excluding jw29 were combined.

[6]Strictly speaking the results are shown for the first six phonemes, where one of these is composed of two distinct phones.

[7]Only a small fraction of the total number of these statistics is shown, since in all there are 5264 such means and standard deviations, which far exceeds what can practically be represented.

the more detailed study of possible differences in the articulations used by various speakers is a topic of great interest, such a study lies outside the scope of this dissertation. It is therefore assumed here that the normalised data set combination technique described above can be applied to the data for each phoneme and articulator as a first approximation to a speaker-independent articulatory model.

7.2.2 Parametric Positional and Curvature Models

Given a set of positional statistics, the corresponding articulatory trajectory curvatures at the midpoints of the phonemes are estimated as described in Section 3.4.1. The distributions of both positional and curvature values are then modelled using a single Gaussian distribution for each [articulator, phoneme] pair, thus characterising the range of variation seen in each case by just two parameters, namely the mean and variance of the Gaussian concerned.

The quality of the match between the probability density function of the sampled data and that of the corresponding parametric model can be assessed using a variety of statistical tests. The Kolmogorov-Smirnov, or K-S, test was employed here [158], which predicts the probability that the observed data could have been drawn from a hypothesised distribution, by measuring the maximum absolute difference D between the cumulative distribution function of the data, $\mathcal{C}_D(x)$ and that of the hypothesised model, $\mathcal{C}_M(x)$:

$$D = \max_{-\infty<x<\infty} |\,\mathcal{C}_D(x) - \mathcal{C}_M(x)| \tag{7.5}$$

Since the distribution of this statistic can be computed for multiple random data sets drawn from the *same* distribution, the significance of an observed value of D can be computed by comparing it with this distribution of D values for like data sets. The probability that a measured value of D is less than that which would be obtained from a random data set generated by the hypothesised model, D_M, is then approximately given by:

$$P(D < D_M) = 2 \sum_{k=1}^{\infty} (-1)^{k-1} \, \exp(-2k^2[D\,(0.12 + \sqrt{N} + 0.11/\sqrt{N})]^2) \tag{7.6}$$

where N is the number of data points. High values of P indicate a high probability that the data could have been drawn from the hypothesised distribution.

This expression was computed for each of the positional and curvature distributions of the 16 articulatory variables[8] during the production of each of the 56 phonemes, to give a set of 1792 K-S probabilities per speaker. The resulting percentage proportions of positional and curvature statistics which have greater than 90%, 50% and 10% probability respectively of being drawn from single Gaussian distributions are given in Table 7.2. Statistics corresponding to the speaker-independent data are also provided, for the speaker jwAvg[9]. In general terms, the positional variables have a much better fit to Gaussian distributions than do the curvature measures, and although relatively few distributions

[8] 14 variables in the case of jw29.

[9] Or "Joe Average".

have greater than 90% probability of having been drawn from single Gaussians, a minority have extremely poor fits.

	Positional models (%) with probabilities greater than:			*Curvature models (%) with probabilities greater than:*		
Speaker	*90%*	*50%*	*10%*	*90%*	*50%*	*10%*
jw16	17	56	87	3	18	52
jw27	17	59	89	2	17	52
jw29	17	52	89	2	20	56
jw18	19	56	86	2	17	54
jw24	17	56	88	3	18	55
jw45	19	56	84	4	18	53
jwAvg	8	36	68	0	2	11

Table 7.2: Percentage proportions of articulatory positional and curvature distributions with greater than 90%, 50% and 10% probability respectively of being drawn from a single Gaussian distribution.

By plotting histograms for the articulatory distributions which yielded poor matches, it was observed that these were typically either caused by skews on the distributions, or distribution tails which fell away at non-Gaussian rates. Example plots showing the positional sample values, as well as the computation of the K-S probability from cumulative distributions, and comparisons of the corresponding Gaussian and histogram shapes are given in Figures 7.4 and 7.5 for two articulatory positional distributions for the speaker jw18.

Figure 7.4 shows the positional distribution for *MIy* during the phoneme `/s/`, in which the fit to a Gaussian model is relatively good. Figure 7.5 shows the distribution of *ULy* for `/s/`, which yields a very low K-S probability due to the very different shapes of the two tails of the histogram above and below the median position, where a "harder" positional limit is imposed in the positive direction. As discussed in Section 3.2.3, skews such as this are due to both the limited variety of contexts available in the UW data, and to the physiological limitations imposed on articulatory positions in the vocal tract.

Since all six speakers were reading from the same text during data acquisition and used similar articulatory strategies, these difficulties are not reduced when the data are combined in the speaker-independent case. In fact, since the number of points in each distribution is now much larger, the skews observed on the data are considered less likely to have been produced by random effects, and hence the K-S probabilities for jwAvg are significantly lower than those for the individual speakers.

Finally, for some data sets poor fits to Gaussian distributions may result indirectly from errors in the automatic alignment of the training data by HTK and/or mis-tracked pellet samples which were not identified in the database. Errors such as these will typically lead to outliers in the articulatory positional sample distributions for the phoneme concerned, which are poorly modelled by the single Gaussian assumption. In these cases however, a poor fit to the articulatory data is *desirable* in order to exclude erroneous data points

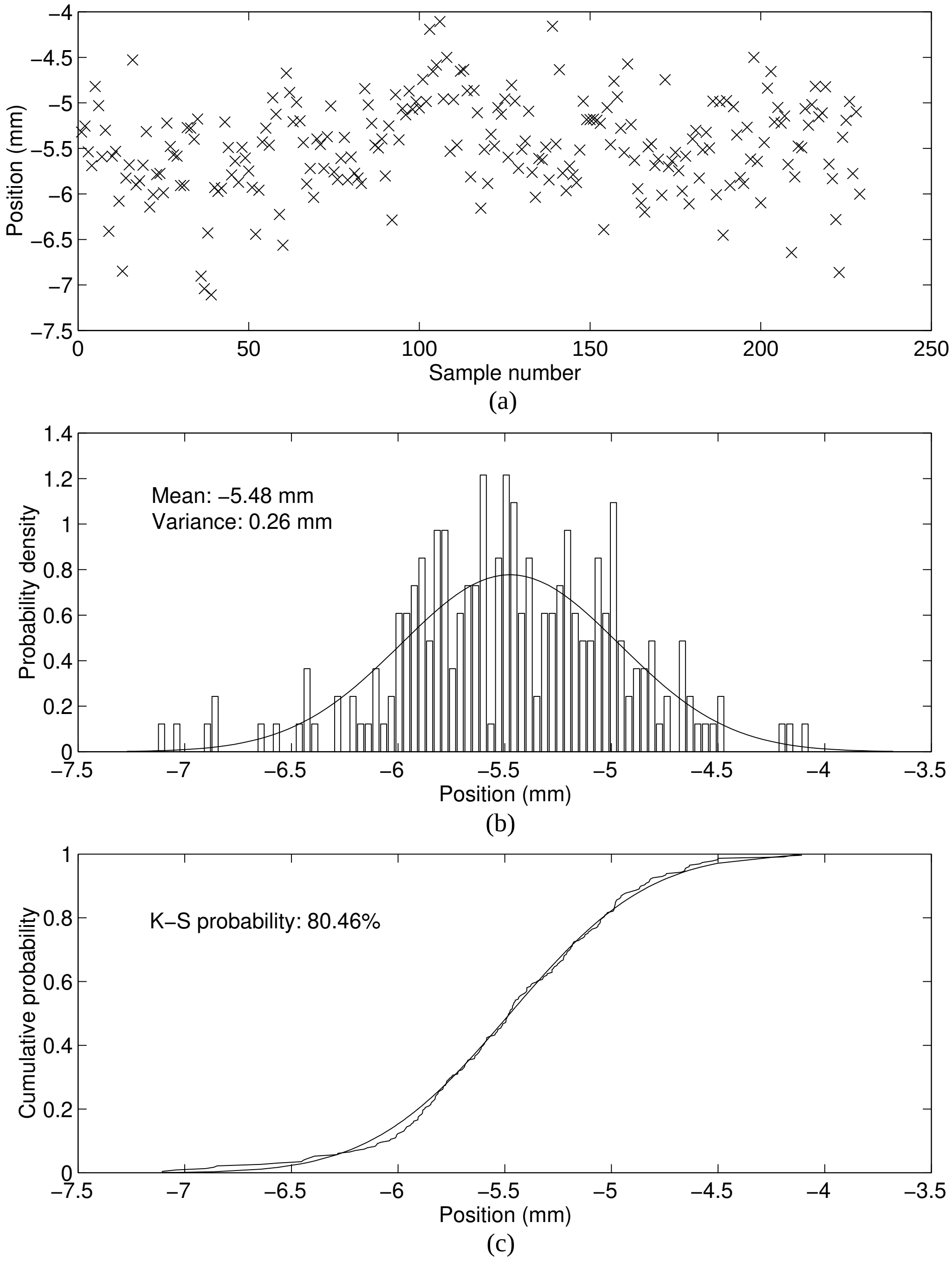

Figure 7.4: Plots of (a) Positional samples (b) Comparison of data histogram with Gaussian distribution and (c) Kolmogorov-Smirnov probability computation from the cumulative distribution functions of the data and the hypothesised Gaussian model, for the positional distribution of MIy during production of `/s/` by `jw18`.

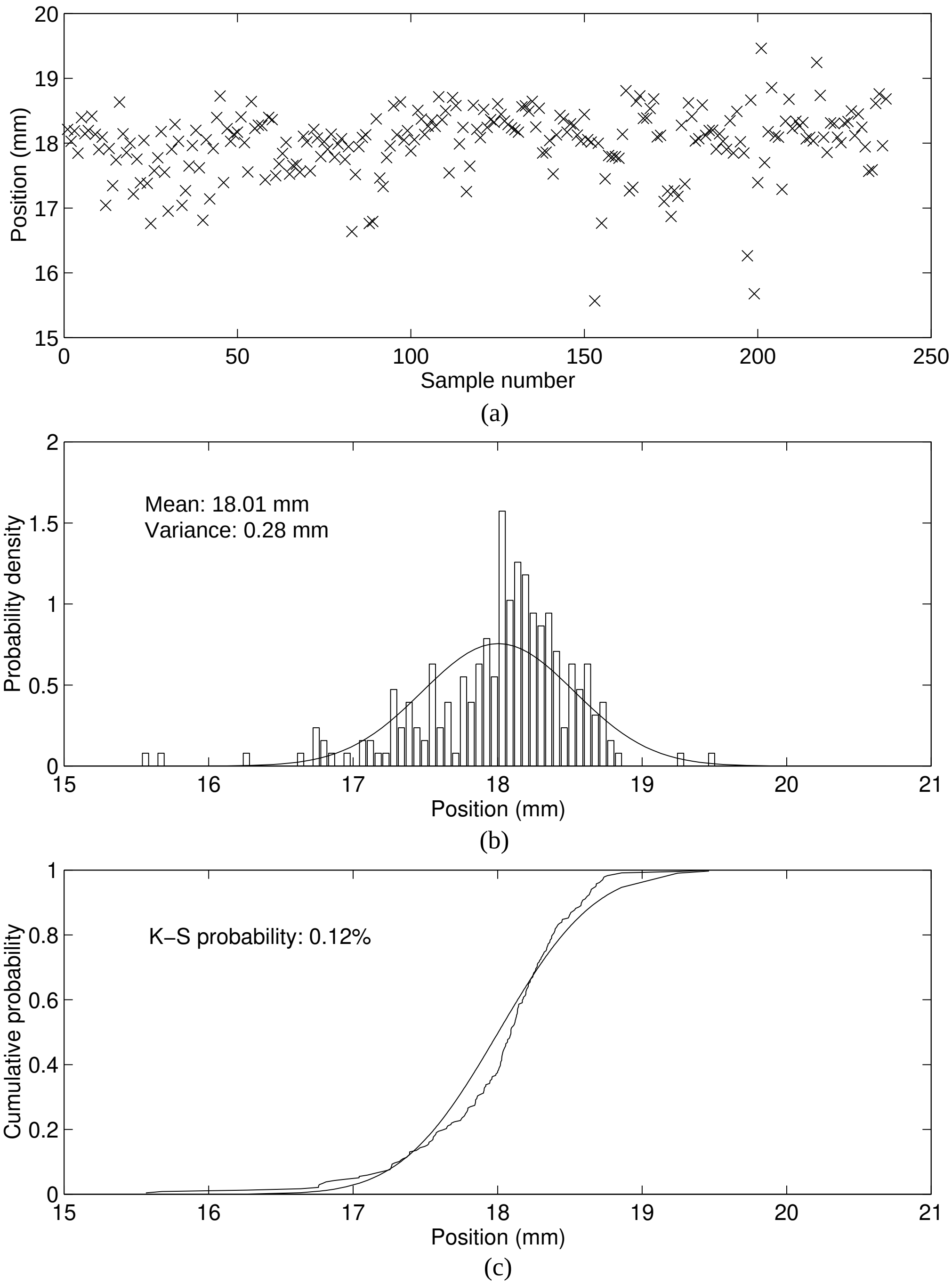

Figure 7.5: Plots of (a) Positional samples (b) Comparison of data histogram with Gaussian distribution and (c) Kolmogorov-Smirnov probability computation from the cumulative distribution functions of the data and the hypothesised Gaussian model, for the positional distribution of ULy during production of `/s/` by `jw18`.

from the statistical models[10].

7.2.3 Correlation Coefficients

Correlation coefficients for pairs of curvature and positional statistics were computed as described in Section 3.4.2. Before using these to predict the positional variations at the midpoints of phonemes from the curvature estimates however, their statistical significance was first assessed, since relatively small data set sizes were available for the less frequent phonemes.

Significance Testing

Student's T-test [87] can be used to compute the significance of a correlation coefficient ρ computed over N data points from two Gaussian distributions. The constant t_0 is first computed as:

$$t_0 = \rho \sqrt{\frac{N-2}{1-\rho^2}} \tag{7.7}$$

Then to test the hypothesis $\rho = 0$ against the possibilities that $\rho > 0$ or $\rho < 0$ at the significance level α (eg. $\alpha = 0.05$ at the 5% significance level), Student's cumulative T-distribution with $N-2$ degrees of freedom is used to compute the probability P of Student's T variable taking a value less than or equal to t_0. Significance at the level α requires:

$$P(T \leq t_0) \ \geq \ 1 - \frac{\alpha}{2} \tag{7.8}$$

Figure 7.6 shows a plot of the regions of acceptance for the hypothesis $\rho = 0$ against $\rho \neq 0$ for $\alpha = 0.05$, as the number of data points N and the absolute magnitude of the correlation coefficient ρ are varied.

Correlation coefficients which are not significant at this level are therefore set to zero, and the remainder are retained. Table 7.3 shows the number of correlation coefficients excluding those with absolute values less than 0.1 which were set to zero when this significance test was applied to both the speaker-dependent and speaker-independent data sets. As expected, due to the larger data set sizes in the speaker-independent case, far fewer coefficients are judged insignificant compared with the speaker-dependent data sets.

Correlation Magnitudes

Strong positive correlations were observed between the positional and curvature estimates for most articulators during the production of the different phonemes. A large positive curvature value—corresponding to a local minimum in an articulatory trajectory, where the articulator's direction of motion is changing rapidly—will therefore tend to result in a positional value greater than the mean position and *vice-versa*, as discussed in Section 3.4.

[10] An example of this effect is described in Section 7.3.3.

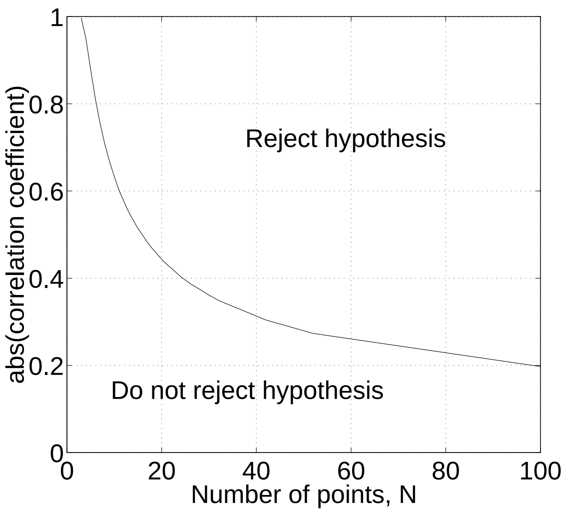

Figure 7.6: Regions of acceptance for the hypothesis $\rho = 0$ against $\rho \neq 0$ at a significance level of 5%.

Speaker	*Insignificant values (%)*
jw16	17.4
jw27	18.8
jw29	19.3
jw18	20.2
jw24	22.8
jw45	18.3
jwAvg	6.0

Table 7.3: Percentages of correlation coefficients set to zero by Student's significance test with $\alpha = 0.05$, excluding those with absolute values less than 0.1.

The mean significant correlation coefficients obtained for each articulator and averaged over all phonemes are illustrated in Figure 7.7 for the speaker jw18, and in Figure 7.8 for the speaker-independent data.

Once again there is good agreement between the statistics for the speaker-dependent and speaker-independent data sets, reflecting the consistent results obtained across all six speakers. The x and y positions of the four tongue pellets show the greatest degree of correlation between curvature and position, with the horizontal positions of the two jaw pellets, MIx and MMx showing considerably less correlation than the other articulators.

This is as expected, since as described in Section 7.2.1 the amount of variation in the degree of protrusion of the jaw is not only small in magnitude with respect to the other articulatory variables, but is also rarely discriminatory between phonemes. Much of the variation for these two variables will therefore be random rather than systematic, and hence is not correlated with the curvature measure which is assumed to be predictive only of systematic positional variations. This lack of correlation for randomly varying

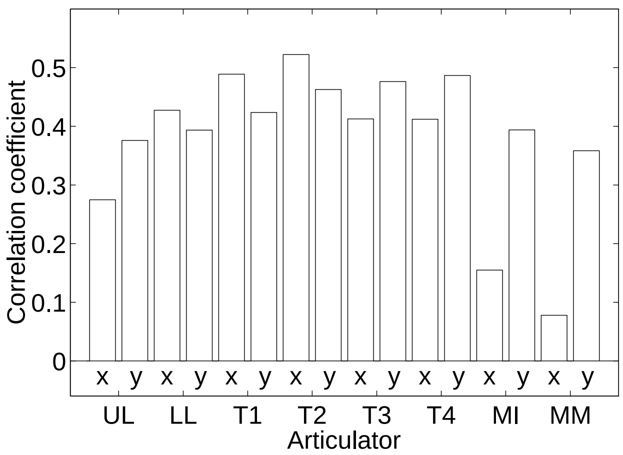

Figure 7.7: Bar graph of mean significant correlation coefficients for each articulator over all phonemes, for the speaker jw18.

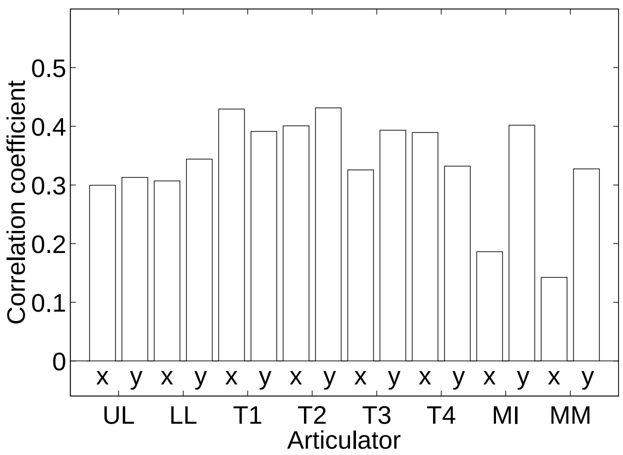

Figure 7.8: Bar graph of mean significant correlation coefficients for each articulator over all phonemes, for speaker-independent data.

parameters allows the removal of a great deal of the random movement seen in actual X-ray articulatory traces when synthesising articulatory trajectories, as will be discussed in Section 7.2.4.

Figure 7.9 shows an example of the correlation coefficients for the x and y positional values of the 8 articulatory pellets for jw18, sampled at the midpoints of all examples of the phoneme /s/, where the correlation coefficients for *T3x*, *MIx* and *MMx* have been set to zero as their values were insufficiently large to be considered significant.

The relatively low correlation coefficients for lower lip, jaw and tongue tip positions in this figure reflect the fact that these articulators are highly constrained in position for the production of /s/. By contrast, the tongue back and upper lip are relatively free to move into positions dictated by neighbouring phonemes, as evidenced by the larger correlations for *UL* and *T2* to *T4*.

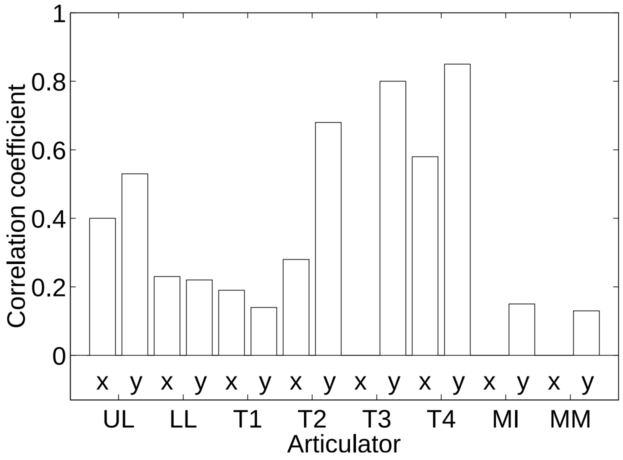

Figure 7.9: Correlation coefficients for the phoneme `/s/` for `jw18`.

7.2.4 Articulatory Trajectories

As described in Section 3.4, variations in articulatory positions at the midpoints of phonemes due to co-articulatory effects can be predicted by computing the estimated curvature c of the articulatory trajectory from a linear interpolation between successive mean articulatory positions, and then using this curvature value to determine the most probable positional value, given by:

$$E\left[P \mid C\!=\!c\right] \;=\; \mu_P + \rho_{CP}\,\frac{\sigma_P}{\sigma_C}\,(c - \mu_C) \tag{7.9}$$

where μ_P, σ_P and μ_C, σ_C are the means and standard deviations of the positional and curvature distributions respectively, and ρ_{CP} is the correlation coefficient between these two distributions. Since these correlations arise from systematic variations in articulatory positions, the values at the midpoints of phonemes predicted by this model will contain less random variation than would be seen in actual X-ray traces.

Complete articulatory trajectories were generated by linear interpolation between the co-articulated time-aligned midpoint positions of successive phonemes, as described in Section 3.4.3. To enhance the system's robustness to unusual contexts given the small size of the training data set, a low-order low pass filter was applied to the resulting trajectories. This filter's effect on standard articulatory trajectories is very slight, and its purpose is to remove very sharp articulatory movements which are otherwise observed in approximately 0.3% of phonetic contexts.

The training and test set data in the UW corpus were aligned to their corresponding transcriptions using HTK, to yield a time-aligned phonetic string for each utterance. Synthetic articulatory trajectories were then constructed from these, both with and without modifying positions at the midpoints of the phonemes using the explicit co-articulation model[11]. The errors between synthetic training and test set trajectories were computed

[11] The trajectories generated without using the explicit co-articulation model nevertheless incorporate strong co-articulatory effects, as discussed in Section 3.4.3.

at all points with respect to the corresponding X-ray traces, and in each case the co-articulation model gave a reduction in the articulatory errors computed over entire utterances. Figure 7.10 shows six bar graphs representing a breakdown by articulator of the mean error obtained over the 51 utterances in the test sets for each of the UW speakers, both with and without the explicit co-articulation model[12].

The error for each parameter has been scaled by that articulator's standard deviation over the entire training set to give a more meaningful comparison, since the absolute magnitudes of the positional variations of the articulators vary greatly (Figure 7.1). As shown in Figure 7.10, each individual articulator's positional error decreases when the co-articulation model is used. The errors in tongue position are generally less than those for lip and jaw position, with the x position of the front and back of the jaw (MIx and MMx) being most poorly modelled.

The poor performance obtained for these latter articulators is not surprising, since horizontal movements of the jaw have relatively little effect on the acoustic signal, as previously described. As discussed in Section 3.4, this implies that much of the variation seen in these variables will be random movement, as evidenced by the correspondingly low mean correlation coefficients in Figures 7.7 and 7.8. Since these random variations cannot be modelled, the co-articulation model has relatively little effect on the errors obtained for these articulators, as demonstrated in Figure 7.10. This is a desirable result, since it implies that the co-articulation model has the greatest beneficial effect on the positions of the articulators which are most significant acoustically.

An example of the effects of the explicit co-articulation model when synthesising an articulatory trajectory is given in Figure 7.11. This figure shows both synthesised and X-ray articulatory traces for the articulator $T4y$, corresponding to part of a test set utterance for the speaker `jw18`.

Synthetic trajectories generated both with and without using explicit co-articulation at the midpoints of the phonemes are shown, and a phoneme-level transcription is provided in which phonetic boundaries determined using HTK are shown by vertical dotted lines. The principal effects of using the explicit co-articulation model when synthesising the articulatory trajectory are as follows:

- The tongue back is raised to form an occlusion at the rear of the oral tract during the production of `/n/` and `/g/`, as evidenced by the relatively large tongue back heights around frames 200 and 260 respectively. The use of the explicit co-articulation model to modify articulatory positions at the midpoints of these phonemes results in a more accurate approximation to these positional extrema.

- An economy of articulatory movement is achieved by the tongue back during the first syllable of "program". The *mean* positions of $T4y$ at the midpoints of the phonemes `/r/` and `/ow/` can be deduced from the synthetic articulatory trace produced with-

[12] These trajectories were synthesised using speaker-dependent articulatory statistics. The speaker-independent statistics derived from the UW data are used in Section 7.4 to generate articulatory trajectories for the RM speakers.

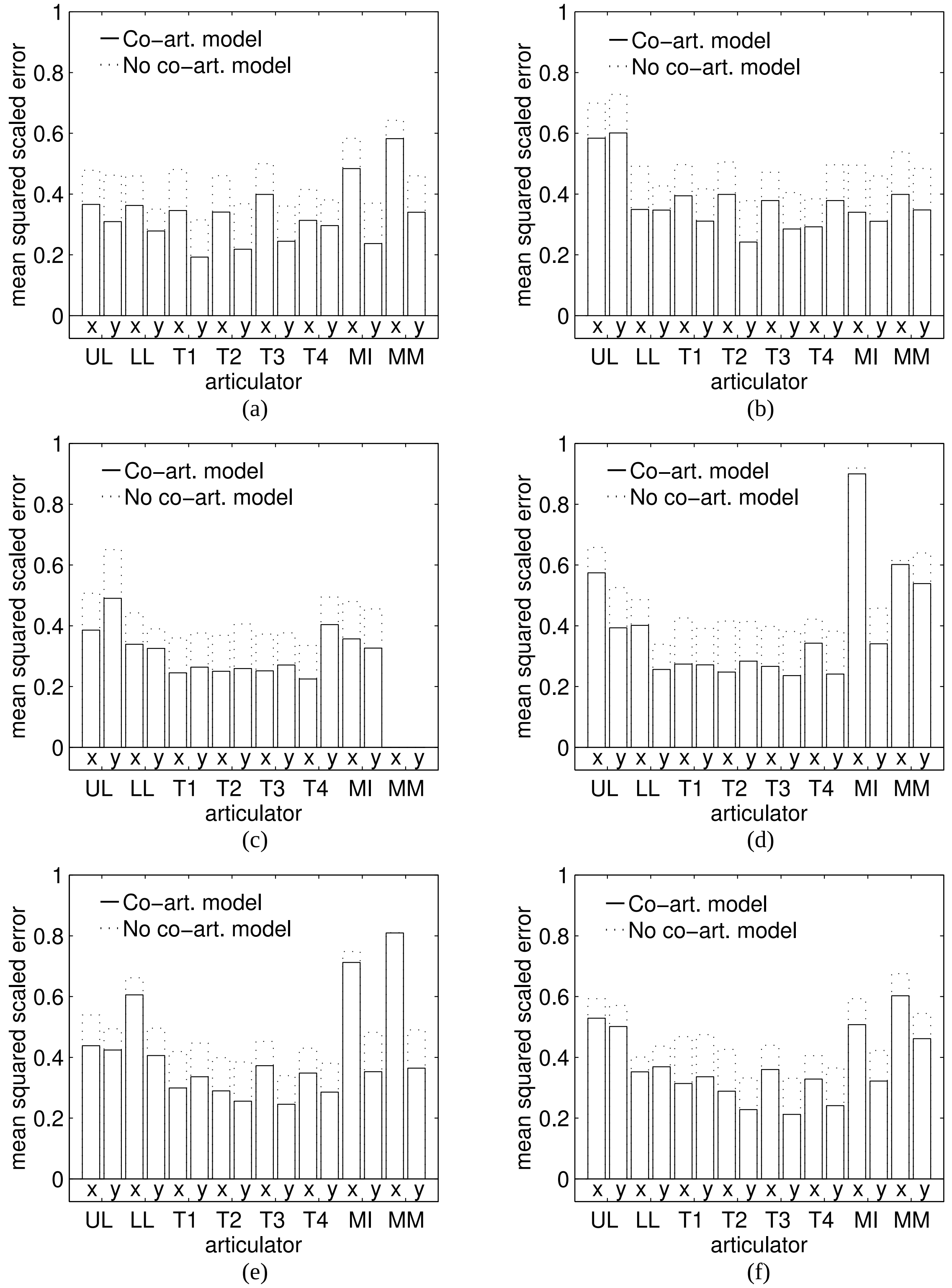

Figure 7.10: Normalised test set errors by articulator, with and without co-articulation model: (a) jw16 (b) jw27 (c) jw29 (d) jw18 (e) jw24 (f) jw45.

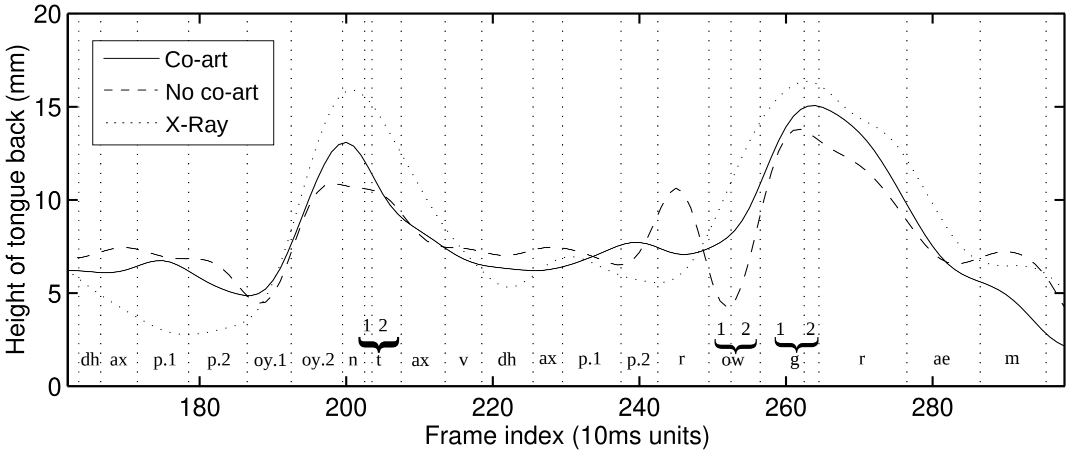

Figure 7.11: Movements in the height of the tongue back ($T4y$) for the speaker jw18 during the test set utterance: "the point of the program". Articulatory trajectories synthesised both with and without the explicit model of co-articulation are shown, as well as the corresponding X-ray trace. Automatically-generated phonetic boundary positions are shown by vertical dotted lines.

out co-articulation[13]. The achievement of these mean positions requires relatively rapid changes in the direction of motion of $T4y$—and hence relatively large muscular effort—between frames 240 and 260, as evidenced by the dashed curve. Since these phonemes are largely articulated using movements in the tongue tip however, the height of the back of the tongue is relatively underspecified, and hence is free to follow a trajectory which requires less articulatory effort, as seen in both the X-ray trace and that synthesised using the explicit co-articulation model.

Overall, the use of the explicit co-articulation model results in a significantly closer approximation to the X-ray articulatory trace, and it is hoped that this will then translate into a reduced error at the outputs of the acoustic models, as described in the following section.

7.3 Acoustic Model: UW Data

In this section an evaluation of the acoustic model as applied to the UW data set is presented. The methods used for preparing the data to be used in training both the artificial neural networks and a set of linear models are discussed, and the implementation of the network training algorithm and the selection of network architectures are described.

Results are presented for the prediction of log spectral vector sequences using both linear regression and MLPs, and a comparison is made between the performances of the networks when trained on both X-ray and synthetic articulatory data.

[13]The model excluding explicit co-articulation uses a linear interpolation between successive articulatory mean positions.

7.3.1 Data Preparation

The inputs to the MLPs used to approximate the acoustic mapping are 16-dimensional articulatory variable positions, and the target vectors produced at their outputs are 24-dimensional log spectral vectors, as described in Section 4.2. The signal supplied to a node j in the hidden layer of a network is a weighted sum of the outputs of the nodes in the input layer, added to a constant "bias" value:

$$\sum_i (w_{ij} y_i) + b_j \tag{7.10}$$

where y_i is the i^{th} network input value, b_j is the bias for hidden node j and w_{ij} is the weight on the connection between nodes i and j.

The parameters w_{ij} and b_j are therefore used to scale these inputs according to both their magnitudes and their relevance to the estimation of the network's outputs. To restrict the values of these parameters to manageable sizes, the data for each articulatory variable were normalised to a mean of zero and a standard deviation of one before training the networks.

Log Spectral Vector Scaling

The use of log spectral vectors for the acoustic signal representation at the network outputs raises an additional scaling problem, since an increase in the volume of a speaker's voice will lead to a corresponding increase in the magnitudes of these spectral vectors. This is an undesirable effect, as the target vectors presented to the networks will then have a random—and hence unpredictable—bias.

Each individual log spectral vector was therefore scaled by computing the mean energy in its 24 coefficients and subtracting this mean value from each of them. The result of this energy scaling is that the outputs of the MLPs no longer represent actual log spectral vectors, but rather spectral "shapes", which are independent of the amplitude of the acoustic signal. The mean and standard deviation of the energy in each log spectral coefficient for each phoneme were therefore also computed over the entire training set, to provide a separate model of absolute spectral energy levels[14].

Figure 7.12 shows the results of this energy scaling process for the vowel `/iy/` and the fricative `/s/`, for the speaker `jw18`. In this figure, the 24-dimensional log spectral vectors from the training data sets for `/iy/` and `/s/` have been overlaid, both before and after subtraction of every vector's mean value from each of its coefficients.

In both cases the removal of the mean bias enables the characteristic spectral shape to be more easily identified. In the data for `/iy/`, this reduction in the spread of log spectral values is fairly even across the frequency range (coefficients 1 to 24). For `/s/` however, the technique is more successful at mid and high-frequencies, with a small *increase* observed in the spread for coefficients 1 to 5 (up to $\approx 500Hz$).

The cause of this increase is the very uneven distribution of energy in the frequency domain for `/s/`. Since frication energy is concentrated at high frequencies, the plot of raw

[14] The use of these two different spectral measures in the scoring of re-synthesised log spectral vectors when performing recognition is detailed in Section 8.3.1.

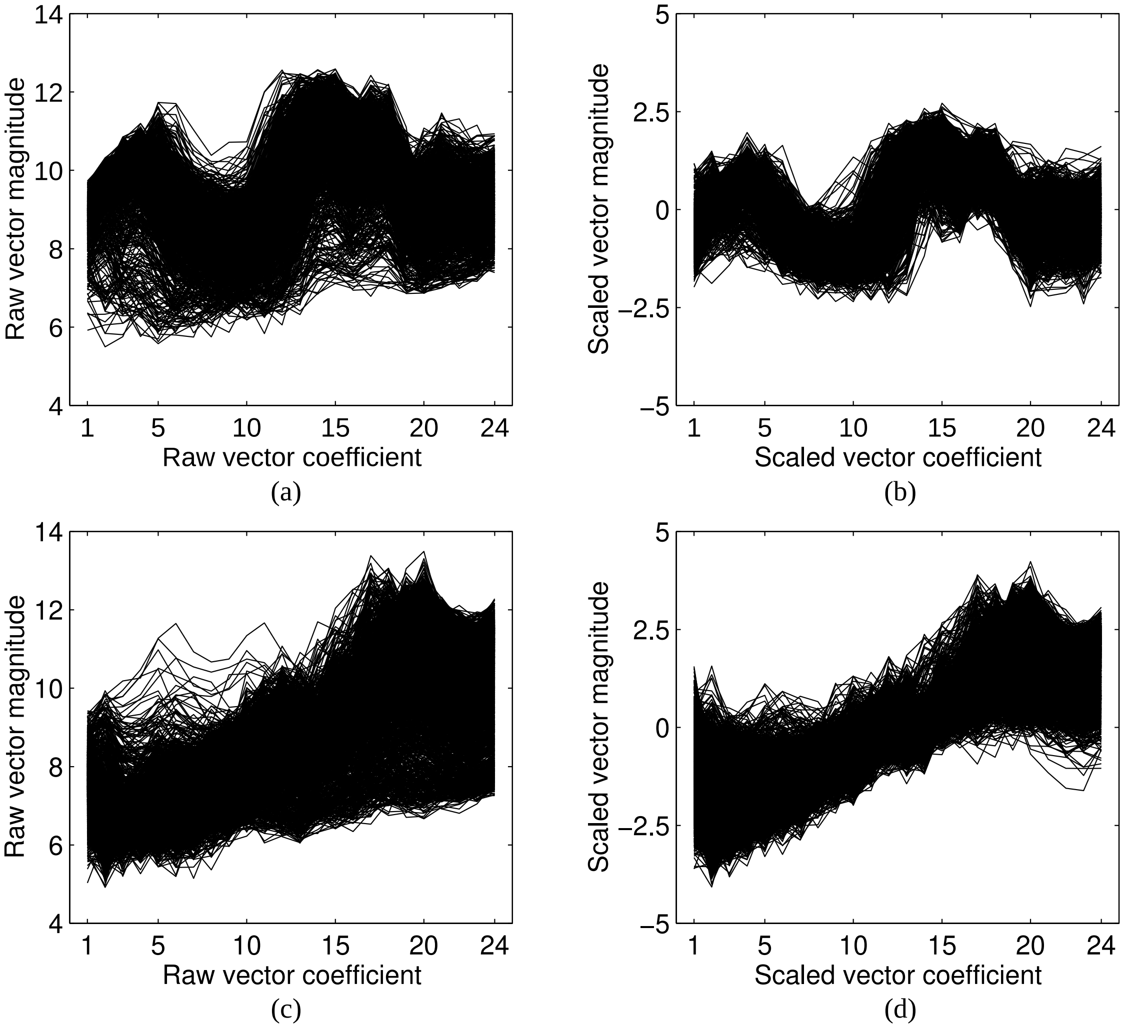

Figure 7.12: Effect of subtracting each log spectral vector's mean value from each of its coefficients over the entire training set for the speaker `jw18`, for `/iy/` and `/s/`: (a) `/iy/` before scaling (b) `/iy/` after scaling (c) `/s/` before scaling (d) `/s/` after scaling.

vector magnitudes shows greater amplitude variation at these frequencies—due to changes in the volume of the speaker's voice—than at low frequencies, where the energy level is close to that of the background noise[15]. The values of the higher log spectral coefficients will therefore dominate the mean log energy calculation, so that the log energy scaling results in a reduction in the spread of values seen in these higher coefficients, with an associated increase in the spread for the first 5 coefficients.

Data Set Merging

Finally, although the cross-validation training technique described in Section 4.4.1 theoretically ensures that the MLPs used for the acoustic mapping do not over-train during

[15]Typical values for the first 5 raw log spectral vector coefficients during silences lie in the range 5.7 to 7.0 for this data set.

optimisation, the effectiveness of this technique in practice is dependent upon the nature of the training, cross-validation and test sets.

In particular, for successful training each of these three data sets should be representative of the input articulatory space for the mapping concerned. If this constraint is satisfied, then by ensuring optimum performance on a separate cross-validation set during training, near-optimum performance will also be obtained on the independent test set[16]. In this case, the prediction of acoustic outputs for unseen articulatory inputs will involve an *interpolation* based on the mapping learned from the training data.

If the training data are not representative of the spaces spanned by the cross-validation and test sets however, a greater proportion of acoustic outputs will be determined by extrapolation. In this case the network's performance on the cross-validation set will be a less reliable indicator of the optimum stopping point during training, and the results obtained on the cross-validation and test sets are more likely to be divergent.

As described in Section 5.2, approximately three quarters of the data for each phoneme from each speaker were used to train the MLPs, and the remaining one quarter were used as a test set. One third of this training data was set aside in turn for cross-validation, leaving half of the original data for optimising the parameters of the networks.

Due to the small size of the X-ray microbeam database, this partitioning results in very small data set sizes—defined here as less than 100 training set vectors—for a few of the least frequent phonemes. To reduce the problem of unrepresentative data sets described above, the data for these infrequent phonemes were therefore merged with those corresponding to more common, and acoustically similar phonemes[17].

`/en/`	+	`/n/`
`/dx.1/`	+	`/dx.2/`
`/ts.1/`	+	`/t.1/`
`/ts.2/`	+	`/s/`

Table 7.4: Data sets corresponding to very infrequent phonemes and phones (left column) which were merged with acoustically similar data sets (right column).

The data sets which were merged are summarised in Table 7.4. As discussed in Section 5.3.2, the flap `/dx/` was originally sub-divided into the two phones `/dx.1/` and `/dx.2/` on the assumption that it is articulated as a brief closure between the tongue and hard palate, with a subsequent acoustically dissimilar release. In practice, such a closure is rarely achieved and `/dx/` is articulated as a *fricative*, with the result that the frames which are aligned to `/dx.1/` and `/dx.2/` are very similar acoustically. After data set merging, the number of distinct phoneme-specific MLP mappings was reduced from 56 to 52 for each UW speaker.

[16] A rough guide to the validity of this assumption in practice can be obtained by comparing the network's mean performance on the cross-validation and test sets, which should be similar.

[17] When merging data sets in this way it is important to avoid combining data with similar articulatory representations but dissimilar acoustic patterns (eg. `/s/` and `/z/`), as vectors which are proximate in the input space should also be proximate in the output space, to ensure a smooth mapping from inputs to outputs

7.3.2 Network Architecture Selection

As described in Section 4.4.2, a separate MLP was used to model the mapping from normalised articulatory positions to normalised log spectral vectors for each of the phonemes. The network parameters were initialised to random values in the range -0.01 to 0.01, and resilient back-propagation was used to optimise their values over a training data set, using cross-validation on a separate data set as a stopping criterion to prevent over-fitting to the training data (Section 4.4.1).

Since each hidden unit in the MLP architecture is able to compute a separate non-linear function of the network's inputs, one approach to network architecture selection is to provide a very large number of hidden nodes—and hence a very powerful network—and to rely on the technique of cross-validation to prevent over-fitting.

As the computation time required to train a large MLP can be very long however, in practical implementations it is desirable to identify the smallest network architecture which can adequately model the data. This is achieved by training a set of neural networks of increasing size on the same data set, until the cross-validation error ceases to decrease significantly.

Figure 7.13 shows a plot of the final network training and cross-validation errors obtained for MLPs trained on the data from the phone `/ow.1/` for the speaker `jw18`, when the size of the network's hidden layer was set to 1, 2, . . . , 9, 10, 12, 15, 20, 25 and 50 nodes respectively.

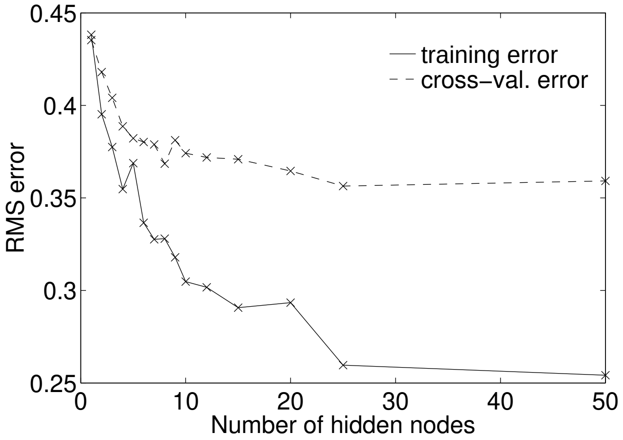

Figure 7.13: Final training set and cross-validation set errors for networks trained on the data from the phone `/ow.1/` for the speaker `jw18`, as a function of the number of hidden nodes used.

In each case, the network was trained from a random initial starting point, and training was terminated once optimum performance on the cross-validation set was obtained. As can be seen from the figure, both the training and cross-validation errors decrease as the number of hidden nodes is increased from 1 to 25, but little or no improvement is seen when this number is increased to 50. Since a network with 25 hidden nodes yields approximately the same cross-validation performance as one with 50 nodes, the smaller architecture is therefore retained.

Figure 7.14 shows a histogram of the network architectures chosen using this method, for networks trained on the data sets incorporating the explicit co-articulation model. The data from all six UW speakers and across all phonemes are plotted[18], and the sizes of the networks are represented by the number of hidden layer nodes, where the number of input and output nodes were fixed at 16 and 24 respectively[19].

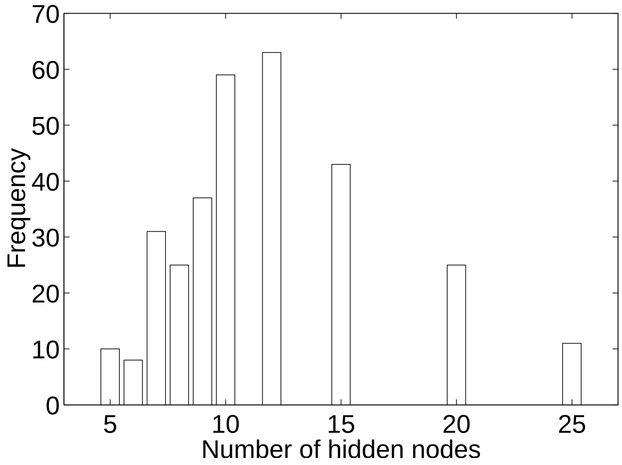

Figure 7.14: Histogram showing the frequencies of networks with various hidden layer sizes, representing all of the networks trained on data incorporating the explicit co-articulation model, for the six UW speakers.

As can be seen from the figure, the median hidden layer size selected was 12 nodes, and a maximum size of 25 nodes was used. The minimum number of hidden nodes was set to 5, since networks of this size trained very rapidly, and no further size reduction was necessitated by speed constraints.

Multiple Random Initialisations

Since the initialisation of the network parameters is a random process, multiple training runs using the same network architecture will typically produce slightly different results. An improvement in performance can therefore often be obtained by training several networks from different random initialisations, and selecting the network which yields the best performance on the cross-validation set.

A further improvement can then be obtained by observing that the network which gives the best performance *on average* usually does not produce the best output for every input vector. As a result, averaging the outputs predicted by a set of different networks will result in a lower error rate than selecting that of any individual network. Once the optimum network architectures had been selected, three networks were therefore trained from different random starting points, and the averages of their outputs were used both as the best estimates of the predicted log spectral vectors, and in the computation of the networks' output error variances (Section 4.4.1).

[18] Hence $6 * 52 = 312$ networks are represented.

[19] Except in the case of jw29, for whom 14 inputs were used.

7.3.3 Acoustic Vector Prediction

The result of training the acoustic models described in the preceding sections on the data from each of the 52 distinct phonemes is a set of:

- Network parameters defining the phoneme-specific mappings from scaled articulatory vectors to energy-normalised acoustic vectors, where these vectors represent quantised scaled spectral energies.
- Variances associated with each network output, which characterise the confidence of the network's predictions for each log spectral vector[20].
- Mean and variance statistics describing the variations observed in the raw log energy in each spectral vector coefficient for each phoneme, computed over all of the spectral vectors in the training set *before* these vectors are scaled for use in training the MLPs (Section 7.3.1).

The use of the last two sets of statistics in the computation of acoustic errors for synthesised spectral vector sequences is described in Section 8.3.1. In this section, the results of training both MLPs and linear regression systems on the scaled articulatory-acoustic data are presented, in terms of the errors observed between the target and predicted acoustic vectors for each phoneme.

Network Error Computation

The MLPs trained on the UW data typically achieved a minimum cross-validation error in less than 1000 optimisation iterations, where an iteration is defined as a single presentation of the entire training set to the network. A measure of the accuracy of the resulting MLP (or linear) mappings can then be obtained by computing the weighted root mean square (RMS) error over the training, cross-validation and test sets. The error at the output of the MLP for the phoneme p is computed over n vectors and at k output nodes as:

$$E_p = \sum_n \sum_k \frac{1}{2}(y_{n,k} - t_{n,k})^2 \tag{7.11}$$

where $t_{n,k}$ is the target output and $y_{n,k}$ is the output value predicted by the network. The combined weighted RMS error over all phonemes is then defined as:

$$E_{RMS} = \sqrt{\frac{\sum_{phonemes} 2E_p}{N_k \sum_{phonemes} N_p}} \tag{7.12}$$

where N_p is the total number of data points for phoneme p and $N_k = 24$ is the total number of network outputs, which is independent of p.

[20] These variances are not local to the spectral vector being predicted, but are globally computed over the training data set, as described in Section 4.4.1.

Linear Regression Versus MLPs

Phoneme-specific acoustic mappings were trained using both MLPs and linear regression[21] for each of the UW speakers. Figure 7.15 shows a bar graph of the resulting combined weighted RMS errors for the speaker jw18, obtained using the linear system, a single MLP network, and the average of three MLP networks trained from different random starting points, respectively.

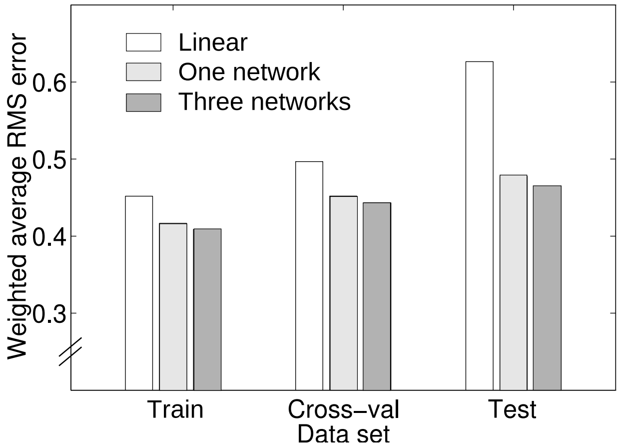

Figure 7.15: Bar graph showing the combined weighted RMS acoustic errors obtained for the speaker jw18 on the training, cross-validation and test data sets. Results are shown for the use of linear regression, one MLP per phoneme, and the average of 3 MLPs trained from different random starting positions on the data for each phoneme, respectively.

Several trends can be observed in this figure. Firstly, the MLP mappings outperform linear regression on each of the data sets, and taking the average of three separate MLP mappings consistently yields slightly less error than using a single mapping alone. Secondly, the differences between the performances of the linear and MLP mappings are greater on the cross-validation data than the training data, and greater still on the test data set. This increasing discrepancy is largely due to poor performance by the linear system on phonemes for which the available data are relatively sparse. As discussed in Section 4.3, the linear system uses a (linear) interpolation to predict acoustic output for articulatory input vectors in the cross-validation and test sets which lie beyond the range of values seen in the training data. By contrast, the thresholding non-linearity used in the MLP mapping (Equation 4.10 in Section 4.4.1), makes it more robust to such outliers since the magnitudes of the network's outputs will be finite even in the extreme case of infinite input values.

Finally, the error for each mapping increases from the training set to the cross-validation set, and from the cross-validation set to the test set. The least error is expected to be obtained on the training set, since this is the data on which the network parameters have been optimised; any significant discrepancies between the errors obtained using the cross-validation and test sets will then be due to the degree to which these sets are truly

[21]The parameters of the linear systems described in this section were estimated using the least mean squared error criterion, as described in Section 4.3.

representative of the input space. If this constraint is not satisfied, then optimising performance on one independent set (the cross-validation data) may not exactly correspond to optimum performance on a second such set (the test data). For the MLP mappings shown the three errors obtained are quite similar to one another, supporting the hypothesis that the networks are adequately trained.

Impact of Co-articulation Model

To assess the impact of using the explicit co-articulation model to generate articulatory trajectories—as opposed to using linear interpolation between mean articulatory positions—three MLP mappings were trained per phoneme on the data for each of the six UW speakers using two separate data sets. In each case the same acoustic target vectors were used, but the articulatory representations differed as to whether the explicit model of co-articulation had been applied. The results obtained are illustrated in Figure 7.16, for the test data sets.

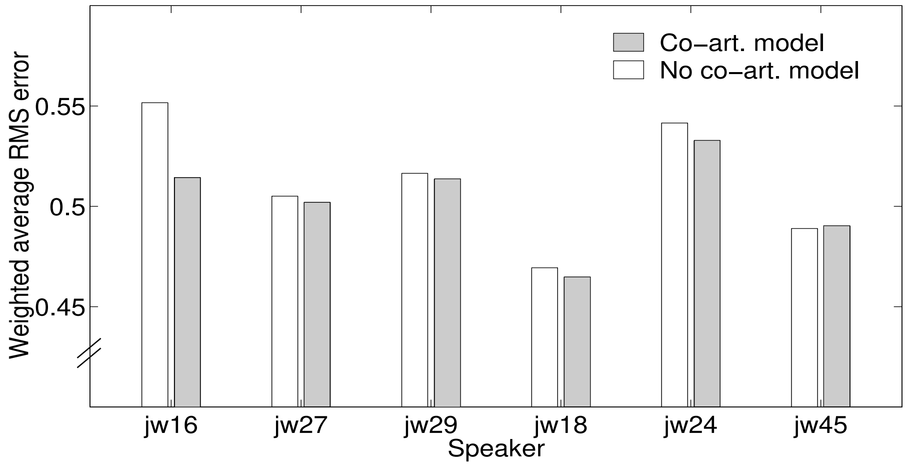

Figure 7.16: Combined weighted RMS acoustic errors obtained on the test sets for the six UW speakers using 3 MLPs per phoneme, for systems trained on articulatory data with and without the explicit co-articulation model.

As shown in the figure, the combined weighted output errors obtained using articulatory data to which the explicit co-articulation model had been applied are less than those obtained using the data sets without the model for every speaker except jw45, and in this last case the two values obtained are very similar. The explicit co-articulation model therefore not only results in more accurate articulatory modelling by comparison with X-ray data (Section 7.2.4), but it appears to yield articulatory trajectories from which the mapping to acoustic vectors is more accurately approximated by the acoustic model.

Performance on X-ray Data

The intended purpose of the production model described in this dissertation is the prediction of acoustic vector sequences from *synthetic* articulatory trajectories, since this allows acoustic outputs to be generated for any arbitrary input phonetic sequence. For comparative purposes however, the acoustic modelling performance obtained by training on actual X-ray data traces was also examined.

For the speaker `jw16` a set of 3 MLPs were trained per phoneme using X-ray articulatory data—excluding vectors which had been labelled as mis-tracks—in an analogous manner to the previously described models trained on synthetic data. The resulting combined weighted RMS errors obtained on the training, cross-validation and test sets using X-ray data, and synthetic data generated both with and without the explicit co-articulation model are illustrated in Figure 7.17.

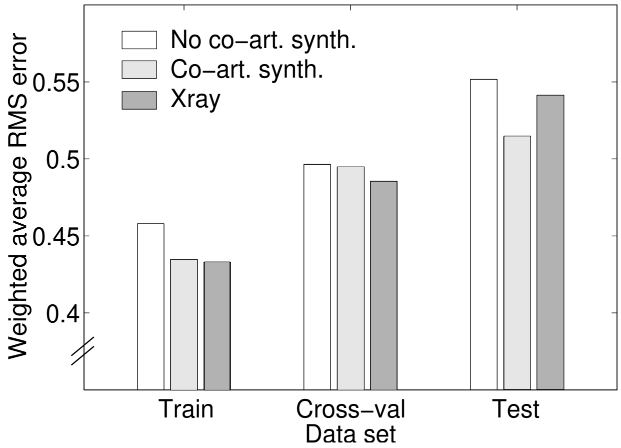

Figure 7.17: Bar graph of combined weighted RMS acoustic errors for the speaker `jw16` on training, cross-validation and test data sets. Results are shown for the average of 3 MLPs trained from different random starting positions on X-ray, synthetic co-articulated and synthetic non-co-articulated data.

The use of X-ray traces at the network inputs results in slightly lower training and cross-validation errors than are obtained using either of the synthetic data sets, but the similarity in the errors obtained using the X-ray and synthetic inputs indicates that the degradation in performance from using synthetic trajectories over X-ray trajectories may not be great. Indeed, for this speaker the use of X-ray data yields an error on the test set which *exceeds* that obtained using co-articulated synthetic data. The high test set error in this case is largely due to the unusually high errors obtained for the phonemes `/jh/`, `/ng/`, `/v/` and `/z/`. The cause of these errors is illustrated in Figure 7.18, which shows a histogram of the horizontal positions of the tongue tip, *T1x*, for each of the test set vectors for the phoneme `/z/`.

The position of the tongue tip is expected to be highly constrained during the production of `/z/`, as evidenced by the majority of articulatory samples which fall in the range $-17mm$ to $-13mm$ in the figure. Due to errors in the automatically-generated alignments and/or mis-tracked pellet samples which were not labelled during preparation of the UW database, a significant number of outlier samples are also included in this data set, and these are responsible for the high error rate.

In the case of the synthetic data sets however, the use of a single Gaussian distribution to model articulatory positions will exclude such outliers from the resulting synthetic trajectories, so that the large error contribution due to these vectors is removed.

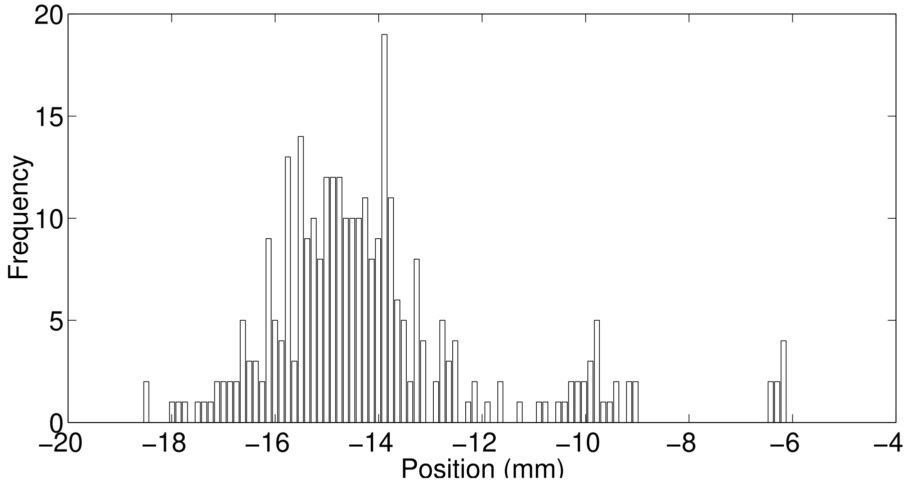

Figure 7.18: Histogram of horizontal tongue tip ($T1x$) positions in the test set data for the production of /z/ by the speaker jw16.

7.4 Articulatory and Acoustic Models: RM Data

Two separate models were implemented for each of the RM speakers das1 and tab0. The first of these used the synthetic articulatory trajectories obtained by inverting an articulatory-acoustic codebook (Section 6.4) to obtain positional and curvature samples, from which 5-dimensional articulatory trajectory vectors corresponding to the acoustic RM data were generated. In the second system, the 16-dimensional articulatory parameter set derived by combining the X-ray data from the six UW speakers (Section 7.2.1) was used to generate these trajectories.

In this section, the results of using the codebook inversion trajectories to derive articulatory parameter statistics are described[22] and the results of using both these statistics and those obtained from the UW speaker-independent data to generate synthetic articulatory trajectories and acoustic vector sequences are presented.

7.4.1 Articulatory Statistics from Codebook Inversion

The result of applying the dynamic programming algorithm described in Section 6.4.2 to the 600 speaker-dependent training data utterances for each of the RM speakers, is a set of 5-dimensional synthetic articulatory trajectories corresponding to each utterance. By using HTK to perform an automatic alignment of the acoustic data to the supplied transcriptions, these trajectories were then sampled in an analogous manner to the X-ray trajectory sampling employed for the UW data set.

Articulatory Positional Variation

Since these synthetic articulatory parameters do not directly represent the positions of physical structures in the vocal tract, but are derived by fitting a mathematical model

[22] The corresponding results for the UW speaker-independent data set were described in Section 7.2.1.

to a set of vocal tract shapes (Section 6.3.1), the observed parameter ranges correspond to those values which yield plausible vocal tract cavity shapes when substituted into Equations 6.1 to 6.6.

The ranges of the parameters a_1 and a_2—which control the tongue body shape through a linear combination of vocal tract area function eigenvectors—are determined by limiting acceptable vocal tract cavity areas to positive values less than $20cm^2$. The ranges of the parameters a_3 and a_4 however—which control the position of the tongue tip—are set to arbitrary values by appropriate scaling in Equations 6.2 to 6.6.

Similarly, the value of the parameter a_5—which determines the area of the opening at the velum—is scaled to yield a range comparable to those of a_3 and a_4. The resulting ranges of the 5 synthetic articulatory parameters are summarised in Table 7.5.

$$
\begin{array}{c}
-0.9 < a_1 < 0.6 \\
-2.0 < a_2 < 0.4 \\
-2.5 < a_3 < 2.5 \\
-2.5 < a_4 < 2.5 \\
-2.5 < a_5 < 2.5
\end{array}
$$

Table 7.5: Approximate ranges of the 5 synthetic articulatory parameters used to generate articulatory trajectories for the RM data.

Discriminatory Usefulness

Due to significant mis-matches in the acoustic signals generated by the synthetic vocal tract model and those produced by the speakers `das1` and `tab0`, the quality of the articulatory trajectories obtained by codebook inversion is poor by comparison with the X-ray data set. The variations observed in the mean positions taken by a given articulator across the set of 56 phonemes were generally not found to be significantly greater than the range of articulatory positions seen during the production of any given phoneme.

This difficulty is illustrated in Figure 7.19, which compares the standard deviations of these mean articulatory positions σ_μ (unshaded), against the average articulatory positional standard deviations σ_{av} (shaded), for each of the five articulators for the speaker `tab0`.

As can be seen from the figure, the value of σ_μ is comparable to that of σ_{av} for a_1, a_2 and a_5, is smaller for a_3, and considerably smaller for a_4. Overall, these initial articulatory statistics provide quite poor discrimination between phonemes, and the application of the explicit model of co-articulation would therefore be of little benefit[23].

Re-estimation of Articulatory Statistics

This situation can be improved by using the mean articulatory positions described above to generate a set of bootstrap articulatory trajectories, from which an improved set of

[23] When Gaussian distributions are fitted to the resulting positional and curvature statistics, the Kolmogorov-Smirnov probabilities computed for these fits are almost uniformly less than 1%, and the corresponding correlation coefficients are on the whole close to zero.

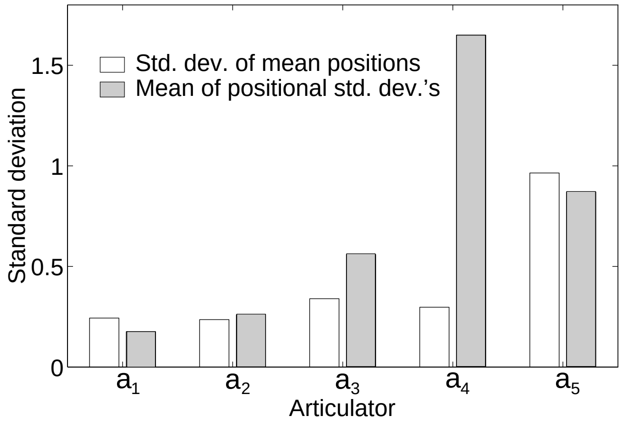

Figure 7.19: Bar graph showing comparisons of the standard deviations of the mean articulatory positions σ_μ (unshaded), against the average articulatory positional standard deviations σ_{av} (shaded), measured at the midpoints of phonemes for the articulatory data produced using codebook inversion for the speaker `tab0`.

statistics can be derived. A set of synthetic articulatory trajectories were initially generated without using the explicit co-articulation model, by simply interpolating between successive mean articulatory positions. The resulting trajectories were then used as input training data for a simple acoustic model comprising a separate 10 hidden-node MLP for the data from each phoneme, which were optimised using the RPROP algorithm and cross-validation.

While the log spectral target vectors used were scaled in the usual way, the corresponding articulatory input vectors used were *unscaled*, since if the positions of scaled articulatory vectors are re-estimated, where this scaling is applied on a phoneme-by-phoneme basis, the re-estimated statistics from the various phonemes will no longer be related in a manner that allows them to be meaningfully re-combined to generate continuous articulatory trajectories. The result of this process is a set of:

- Unscaled synthetic input articulatory trajectories, whose values at the midpoints of phonemes have an uncertainty defined by the initial standard deviation statistics described above.

- Normalised acoustic vector predictions from the MLPs, along with their associated global network error variances.

- Optimised MLP parameters defining the mapping from articulatory to acoustic space.

These three sets of parameters can then be used to re-estimate articulatory positions at the midpoints of phonemes using a technique which is equivalent to applying linearised Kalman filtering on a point-by-point basis. Define:

$\hat{\boldsymbol{x}}_p$	$\equiv$	Initial estimate of the articulatory positional vector $\boldsymbol{x}_p$, at the midpoint of phoneme p
$\hat{P}_p$	$\equiv$	Covariance matrix for $\hat{\boldsymbol{x}}_p$
$h_p()$	$\equiv$	MLP mapping from articulatory to acoustic space for the phoneme p
R_p	$\equiv$	Output error covariance matrix for $h_p()$
H_p	$\equiv$	Jacobian matrix for the mapping $h_p()$
$\boldsymbol{z}_p$	$\equiv$	Target acoustic vector at the midpoint of the phoneme p

where the Jacobian matrix H_p is the matrix of partial derivatives of the MLP's outputs with respect to its inputs. This matrix represents a linearisation of the MLP mapping in the region of an articulatory input vector, and can therefore be used to compute an approximate local inverse of this mapping. Such an inverse mapping H_p can then be used to predict a re-estimated articulatory positional vector $\hat{\boldsymbol{x}}'_p$, using the acoustic error $(\boldsymbol{z}_p - h_p(\hat{\boldsymbol{x}}_p))$ and the covariance matrices R_p and $\hat{P}_p$. The optimum value of $\hat{\boldsymbol{x}}'_p$ in the *linear* sense is found from the linearised Kalman Filtering equation [9, 24, 30]:

$$\hat{\boldsymbol{x}}'_p = \hat{\boldsymbol{x}}_p + \hat{P}_p H_p^T (H_p \hat{P}_p H_p^T + R_p)^{-1} (\boldsymbol{z}_p - h_p(\hat{\boldsymbol{x}}_p)) \tag{7.13}$$

This equation was used to re-estimate the positional samples for each articulator at the midpoints of each of the phonemes. Figure 7.20 shows a bar graph of the standard deviations of the re-estimated mean articulatory positions σ_μ (unshaded), and the average re-estimated articulatory positional standard deviations σ_{av} (shaded), for the resulting set of positional samples.

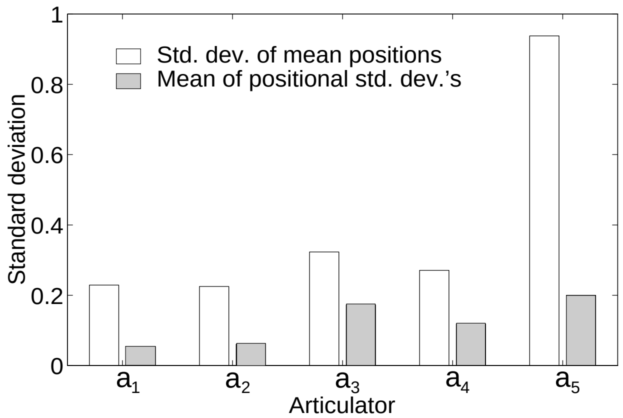

Figure 7.20: Bar graph showing comparisons between the standard deviations of the mean articulatory positions σ_μ (unshaded), against the average articulatory positional standard deviations σ_{av} (shaded), for the re-estimated synthetic articulatory midpoint data for the speaker `tab0`.

By contrast with Figure 7.19, the average articulatory positional standard deviations are now considerably less than the standard deviations of the mean articulatory positions computed across all phonemes, yielding a discriminatory set of articulatory statistics suitable for use as input to the explicit co-articulation model.

Parametric Positional and Curvature Models

These re-estimated positional articulatory statistics were used to generate a corresponding set of trajectory curvature estimates from the RM transcriptions, and the distributions of the resulting positional and curvature values were modelled using single Gaussian functions, as was the case for the UW data.

The probabilities that the statistics being modelled could have been drawn from these Gaussian distributions was once again assessed using the Kolmogorov-Smirnov test, and the results obtained are summarised in Table 7.6.

	Positional models (%) with probabilities greater than:			*Curvature models (%) with probabilities greater than:*		
Speaker	*90%*	*50%*	*10%*	*90%*	*50%*	*10%*
das1	12	41	76	0	7	33
tab0	11	43	77	1	5	30

Table 7.6: Percentage proportions of articulatory positional and curvature distributions with greater than 90%, 50% and 10% probability respectively of being drawn from a single Gaussian distribution, for the RM speakers das1 and tab0.

The probabilities obtained are lower than those corresponding to the UW data in Table 7.2, but are consistent between the two speakers. As observed previously, the curvature statistics have a poorer match to Gaussian distributions than do the corresponding positional statistics.

Correlation Coefficients

The correlation coefficients between the positional and curvature statistics were computed in an analogous manner to those computed for the UW data. A total of 12.1% and 12.9% of the correlation coefficients for das1 and tab0 respectively were found to be insignificant at the 5% significance level using Student's T-test, excluding coefficients with absolute values less than 0.1.

Since the deviations from mean articulatory positions observed in this data set were generated by a re-estimation technique based on linearised Kalman filtering, they no longer represent physiologically interpretable movements. Unlike the results obtained on the UW data therefore, both positive and negative correlations are observed, where the direction of the positional adjustment chosen is that which reduces the error at the output of the MLP mapping concerned, rather than that which minimises the articulatory effort.

Figure 7.21 shows the mean significant correlation *magnitudes* for each of the articulators, averaged over all of the phonemes for the speaker tab0. In each case, a mean correlation coefficient magnitude close to 0.4 is observed, by contrast with the variable correlations seen for the UW articulators.

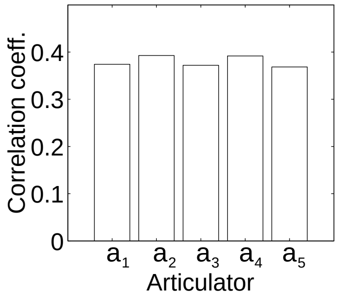

Figure 7.21: Bar graph of mean significant correlation coefficient magnitudes for each of the articulators over all phonemes, for the speaker `tab0`.

7.4.2 Acoustic Vector Prediction

Both 5-dimensional and 16-dimensional articulatory trajectories corresponding to each of the 600 training sentences for `das1` and `tab0` were generated using the explicit model of co-articulation, from the re-estimated synthetic articulatory statistics and the speaker-independent X-ray statistics respectively. Since X-ray trajectories corresponding to the RM data are unavailable, a direct evaluation of the accuracy of these articulatory representations is impossible. The accuracy of the predicted *acoustic* vector sequences was therefore evaluated, and used as an indirect measure of the utility of the two representations.

As was the case for the UW data set, a separate MLP was used to approximate the mapping from articulatory vectors to acoustic vectors for each phoneme, and both the articulatory and acoustic data were scaled before training the networks. Due to the increased size of the RM training data set, it was not necessary in this case to combine phoneme-specific data sets to ensure the networks were adequately trained.

Network Architecture Selection

Networks with 5, 10, 15, 20, 25, 30, 40 and 50 hidden nodes were trained using RPROP to determine the optimum network size for each phoneme, with the results illustrated in Figure 7.22.

This figure shows histograms of the number of hidden nodes used in the MLPs trained on the data from the two RM speakers, using both articulatory trajectories derived from speaker-independent X-ray data, and those derived from the re-estimated synthetic data. The network sizes chosen were greater on average than those used in the acoustic model for the UW data, indicative of an increase in the complexity of the mapping—one cause of which is the relative inaccuracy of the articulatory data supplied at the inputs, compared with the UW speaker-specific X-ray articulatory data.

While the performance of some of the networks indicated that a further reduction in output error could be obtained by using more than 50 hidden nodes, this value was imposed as an upper limit in order to restrict training times to manageable durations, hence the high frequency of network architectures of this size.

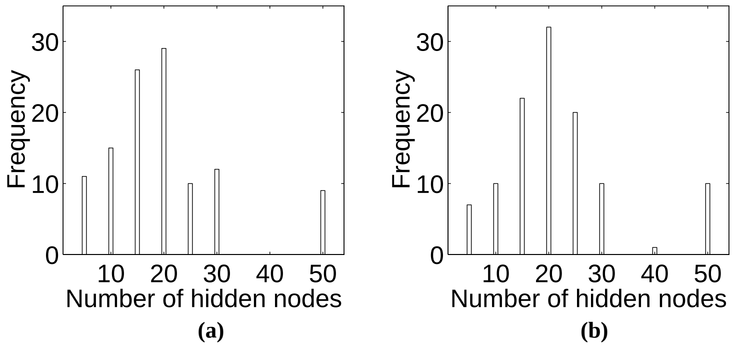

Figure 7.22: Histograms showing the frequencies of the networks chosen in terms of hidden layer size, representing all networks used on the data for both RM speakers: (a) networks trained on speaker-independent X-ray data (b) networks trained on re-estimated synthetic data.

Network Performance

Once the appropriate network architectures had been identified, three MLPs were trained on the data for each phoneme, and their outputs were averaged. Due to the larger sizes of the network architectures, an average of approximately 3000 RPROP iterations were required to obtain optimum performance on the cross-validation set for each phoneme.

The resulting combined weighted RMS errors were computed on the training, cross-validation and test data from each speaker using the two different articulatory input sets. The results obtained for `das1` are shown in Figure 7.23, and those obtained for `tab0` in Figure 7.24.

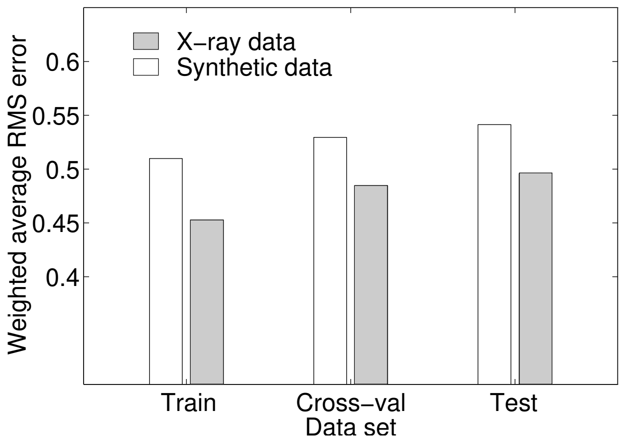

Figure 7.23: Bar graph of combined weighted RMS acoustic errors computed on the training, cross-validation and test data for the speaker `das1`, using both speaker-independent X-ray articulatory trajectories and re-estimated synthetic articulatory trajectories.

As can be seen from the figures, the networks trained on articulatory data generated from the speaker-independent X-ray statistics yield a lower error than do those trained on the speaker-dependent re-estimated synthetic statistics. In addition, the errors obtained are similar in magnitude to those observed for the networks trained on the UW data sets

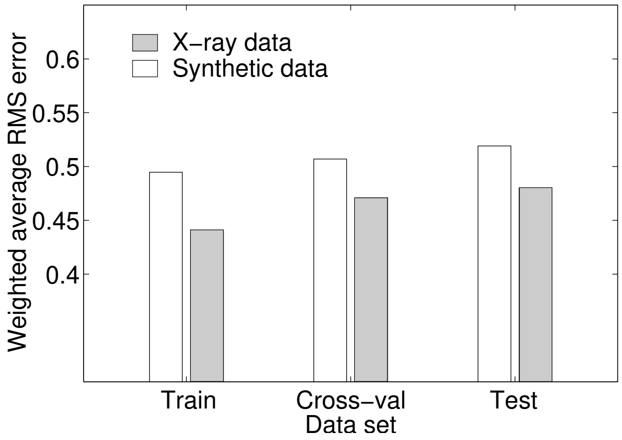

Figure 7.24: Bar graph of combined weighted RMS acoustic errors computed on the training, cross-validation and test data for the speaker `tab0`, using both speaker-independent X-ray articulatory trajectories and re-estimated synthetic articulatory trajectories.

illustrated in Figure 7.16.

It is therefore anticipated that the acoustic models trained using speaker-independent X-ray articulatory data are likely to yield better recognition accuracy when used in a recognition-by-synthesis framework than the corresponding synthetic models.

7.5 Non-linearity of Acoustic Models

As can be seen from Figures 7.14 and 7.22, in many cases the number of hidden nodes used in the MLP mappings is less than the dimensionality of the networks' output spaces. Since the networks are being used as function approximators, the activation function at the output nodes is the identity function, while that at the hidden nodes is a sigmoid function whose output is limited to the range $(0, 1)$ as described in Section 4.4.1.

This implies that the ability of these networks to model non-linearities in the mapping from articulatory positions to parameterised acoustic vectors will be severely limited, since the networks' outputs are computed as linear sums of lower-dimensional representations in hidden-unit space, which in turn are restricted to the range $(0, 1)$ in each dimension. As a result, it is expected that in these cases a linear system should approximate the acoustic mapping at least as accurately as the MLPs, so that the acoustic model employed here is in fact only partially non-linear[24].

The application of both the UW and RM models to the task of augmenting the performance of a standard continuous speech recognition system is the topic addressed in Chapter 8.

[24]The fact that the use of larger MLP architectures did not yield significant improvements in acoustic modelling accuracy in these cases supports the hypothesis that for the phonemes concerned the MLPs did not discover useful non-linear properties in the mapping.

Chapter 8

Recognition System

8.1 Introduction

This chapter describes the application of the articulatory speech production model to the task of speaker-dependent continuous speech recognition. The recognition framework in which the SPM is used was illustrated in Figure 1.1, and is repeated in Figure 8.1 below for ease of reference.

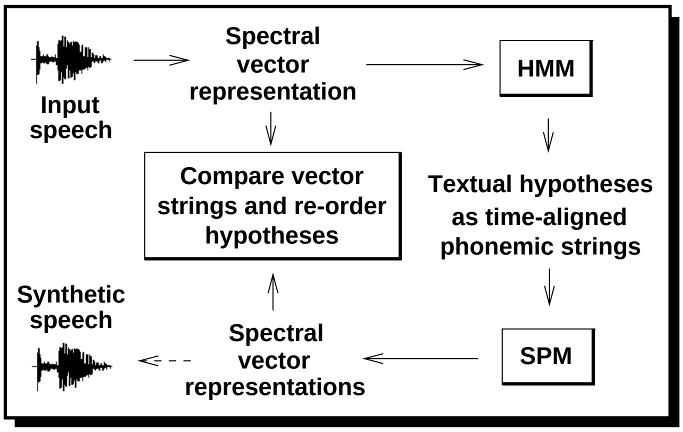

Figure 8.1: Recognition system overview.

The HMM-based recogniser provides an ordered list of N hypothesised transcriptions for each of a set of input utterances, along with their phonetic time-alignments. From these, parameterised speech vectors are synthesised by the SPM. By computing the errors between the synthetic and original speech vectors in the spectral domain, the N-best lists are re-ordered, such that those entries yielding the least spectral error are selected as the most probable transcriptions. Recognition accuracy will improve if:

1. The most probable transcriptions hypothesised by the HMM (the $n = 1$ entries) contain errors for some subset of the utterances being recognised.

2. For a subset of the instances where these $n=1$ transcriptions are erroneous, better transcriptions appear in the hypothesised lists for $n=k$, where $1<k\leq N$.

3. The SPM or a combined SPM-HMM system is able to provide a sufficiently accurate acoustic model to identify these improved transcriptions.

In Section 8.2, the techniques used to generate N-best transcription hypothesis lists using the HTK HMM-based recognition system are described. Section 8.3 subsequently presents the methods employed for re-synthesis of parameterised vectors, and the spectral error measure used to compare these vector sequences with those of the original input speech. Finally, Section 8.5 describes the results obtained when using the combined system to recognise test utterances taken from the UW and RM corpora described in Chapters 5 and 6 respectively.

8.2 Initial Recognition Phase: HMMs

During the initial recognition phase, an HMM-based system is used to transcribe a set of utterances represented by MFCC-parameterised acoustic vectors. Since the aim is to compare the *acoustic* modelling ability of the HMMs with that of the SPM, a "null" grammar[1] is used when performing recognition[2]. The HMMs used for recognition were the same models which were used to align the training data as described in Sections 5.4 and 6.2. This section describes the techniques used to generate N-best hypothesis lists for the test data using these model sets.

8.2.1 Depth of N-best Lists

During the decoding of an acoustic signal within an HMM-based system, a partial search of the space of all possible word strings is performed. In the conventional approach, only the most probable hypothesised transcription identified by the decoder is retained at the end of this search. The recognition algorithm can easily be modified however, to retain the best N transcription hypotheses in order of decreasing estimated probability[3] [176].

The amount of linguistic diversity observed in the resulting N transcriptions will be dependent upon—amongst other factors—the length of the acoustic vector sequence at the input to the recogniser. For example, in a recognition task with a vocabulary of N_v words, the total number of possible transcriptions of length l is N_v^l, which can be a very large number. If the utterance is segmented into two sub-sections of length m and p respectively where $l = m + p$, then recognition hypotheses can be independently generated and re-scored for each sub-section. In this latter case only $N_v^m + N_v^p$ different possibilities

[1] A null grammar is one in which each word in an utterance is followed by any word in the vocabulary with equal likelihood. While the grammar used for the RM corpus was a true null grammar, that used on the UW data was slightly modified to require a tone at the start of each utterance, followed by an arbitrary word string with optional trailing background comments (Section 5.2.2). The term "null" will be retained when describing the grammars used for both systems however, for ease of reference.

[2] A side effect of this constraint is to increase the number of errors made by the HMM-based system, which otherwise would be very low due to the relatively small vocabularies used in the UW and RM corpora.

[3] A review of some N-best search algorithms and their relative performances can be found in the articles by Schwartz et al. [152, 153].

need be considered, which typically yields a considerable reduction in the size of the search space for large values of N_v and l.

While each utterance in the RM test set is limited in length to a single sentence, many of the utterances in the UW corpus comprise multiple sentences. Before generating N-best lists, the longer UW test set utterances were therefore hand-segmented into individual sentences[4]. In each of the experiments described in this chapter, the value of N was set to 100, and a maximum of 10 partial paths terminating in any given node (representing an HMM model or a word end) were retained during the decoding of each utterance[5].

8.2.2 Grammar Scale Factor

The N textual hypotheses for each test set utterance were generated using the HTK Version 2.0 Viterbi decoder [176]. This decoder determines the most probable transcriptions of an utterance by combining the probability scores predicted by an acoustic model with those produced by a syntactically-based language model (Section 1.2.2). The acoustic model predicts the probability[6] P_a that the observed acoustic vectors could have been produced by a given textual hypothesis, while the language model predicts the *a priori* probability of the transcription, P_l.

The use of a null grammar to generate these hypotheses implies that P_l will be a function only of the number of words in the utterance, w, and will be independent of the identities of these words[7]. In a vocabulary of N_v words, the probability of seeing any given word at a particular location in the utterance is therefore $1/N_v$, and the overall *a priori* probability for an utterance comprising w words is:

$$P_l = \left(\frac{1}{N_v}\right)^w \tag{8.1}$$

This value is combined with the corresponding acoustic probability P_a to yield an overall score for the utterance, P_{hmm}. Since the values taken by both P_a and P_l can be extremely low, these probabilities are typically computed in the logarithmic domain to render their numerical values tractable.

By scaling P_l to increase or decrease its value relative to P_a in this computation, it is possible to bias the recogniser in favour of shorter or longer transcriptions as follows:

---

[4] A further increase in efficiency could be made by computing a *lattice* of possible transcriptions for an utterance, and then re-scoring parts of the lattice, as described in Section 9.5.3.

[5] Ostendorf et al. have described an N-best re-scoring system in which only $N = 20$ hypotheses were generated for each utterance, as these were found to be sufficient to include the correct transcription 98% of the time [119]. This system used a statistical class grammar however, and the use of a null grammar in the current system necessitates a deeper list of hypotheses.

[6] In fact the acoustic model does not compute a true probability, but rather a value which is proportional to this probability, estimated from the appropriate acoustic pdfs. This does not affect the recognition performance however, which is a function of the *relative* probabilities of the different hypothesised transcriptions; the term "probability" will be retained in this discussion for ease of description.

[7] This is only approximately true for the UW data, since the grammar in this case also specifies an obligatory tone at the start of each utterance and optional background comments following the utterances, as described in Section 5.4. The effects of these restrictions on the corresponding language model probabilities are slight however, and will be neglected in this discussion.

$$\log(P_{hmm}) = \log(P_a) + G_s \log(P_l) \tag{8.2}$$

where G_s is a weighting constant known as a *grammar scale factor*. Larger values of G_s will tend to result in shorter[8] hypothesised transcriptions at the output of the recogniser, as the penalties resulting from the use of additional words will then be greater.

The use of a grammar scale factor $G_s > 1$ frequently yields a significant increase in recognition accuracy. This effect is due to the high incidence of spurious short words (such as "a") inserted into the hypothesised transcriptions when recognition is based predominantly on the acoustic error. These words are typically inserted at locations in the utterance where they provide a good short-term match to the acoustic signal. For example, the lack of a glottal stop in the RM phoneme set (Table 5.4) means that a word such as "utterance" may be transcribed as "a utterance", where the additional "a" is used as an approximation to the otherwise unmodelled stop release. By contrast, if the value of G_s used is too large, valid words will be deleted from the transcriptions produced (eg. by replacing a correct word string such as "in the mission" with the acoustically less accurate but shorter "intermission"). The value of G_s required for optimum recognition performance is task-specific and in general must be determined empirically.

8.3 Secondary Recognition Phase: SPM

This section describes the methods used to perform spectral comparisons between the synthetic parameterised acoustic sequences generated by the SPM and the vector sequences corresponding to the original speech. Both the absolute spectral energy and the normalised spectral "shape" of the acoustic vectors are compared, and a partial compensation for transcription alignment errors produced by HTK is provided.

8.3.1 Re-scoring *N*-best Transcriptions

The acoustic model described in Chapter 4 uses a separate set of MLPs to predict energy-normalised log spectral vector shapes for each phoneme on a vector-by-vector basis. A global error variance is computed over the training data set for each MLP output, which is subsequently used as a confidence measure for these acoustic vector predictions. A scaled log spectral error measure $E_{s_{i,p}}$ can therefore be computed as:

$$E_{s_{i,p}} = \left(\frac{y_{i,p} - \left(t_i - \frac{1}{24} \sum_{i=1}^{24} t_i \right)}{\sigma_{y_{i,p}}} \right)^2 \tag{8.3}$$

where $\boldsymbol{t}$ is the target (unscaled) log spectral vector corresponding to the original speech, $\boldsymbol{y}_p$ is the energy-normalised log spectral vector predicted at the output of the MLP mapping for the phoneme p, and $\sigma_{y_{i,p}}$ is the error variance for the i^{th} coefficient of $\boldsymbol{y}_p$. The vector $\boldsymbol{t}$ is scaled by subtracting its mean value from each of its 24 coefficients, to match the acoustic representation used in the MLP mappings.

[8] In terms of the total number of *words* used.

This error measure is well suited to comparing the relative concentrations of energy with frequency in an acoustic vector, but does not characterise the raw log spectral energy in each of the vector's coefficients. As described in Section 7.3.1, the energy normalisation applied to the data used in the MLP mappings is more effective in identifying characteristic spectral shapes for some phonemes (eg. vowels) than others (eg. fricatives). As a result, a separate error measure based on the raw spectral energy, $E_{e_{i,p}}$ is also computed as follows:

$$E_{e_{i,p}} = \left(\frac{\mu_{e_{i,p}} - t_i}{\sigma_{e_{i,p}}} \right)^2 \tag{8.4}$$

where t_i is the i^{th} coefficient of the target log spectral vector $\boldsymbol{t}$, $\mu_{e_{i,p}}$ is the mean energy computed from the i^{th} spectral coefficient of each of the *training set vectors* for the phoneme p, and $\sigma_{e_{i,p}}$ is the corresponding standard deviation. The statistics $\mu_{e_{i,p}}$ and $\sigma_{e_{i,p}}$ are therefore broadly indicative of the raw log spectral energy which is expected to be observed at any given frequency during the production of a particular phoneme. Since there is considerable variation in this energy level due to both the phonetic context and to fluctuations in the intensity of the speaker's voice, the values of $\sigma_{e_{i,p}}$ are typically large, and this raw energy-based error measure is less useful for discriminating between different acoustic representations than the spectral shape measure defined in Equation 8.3.

These two scaled errors are then combined in a weighted sum of squares to yield an overall error measure, which is computed over all of the vectors in the two sequences being compared:

$$E = \sum_{vectors} \left[K_s \sum_{i=1}^{24} E_{s_{i,p}} + K_e \sum_{i=1}^{24} E_{e_{i,p}} \right] \tag{8.5}$$

The constants K_s and K_e are weighting factors on the error terms, where $K_s > K_e$ since $E_{s_{i,p}}$ is more useful in discriminating between hypotheses than $E_{e_{i,p}}$, due to the larger number of parameters used in its estimation and the fact that its computation is local to the vector concerned[9]. If the distributions describing both the spectral energy variations and the errors at the outputs of the MLPs are themselves modelled as Gaussians, a simple modification to this measure leads to a probabilistic interpretation of the spectral errors. The logarithm of a Gaussian pdf with mean μ and standard deviation σ for a random variable x is given by:

$$-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \tag{8.6}$$

which has a form similar to the expressions in Equations 8.3 and 8.4, apart from an additional constant term related to the variance of the distribution. The effect of adding this constant to the basic scaled squared difference term is to attach greater significance to those errors computed from distributions with lower variances, since these distributions are assumed to be more accurately determined. The logarithm of the probability that

[9]In fact only the normalised coefficients predicted by the MLP are local to the vector in question, while the standard error deviations are global to the phoneme. This contrasts with $E_{e_{i,p}}$, which is computed from means and standard deviations which are both global to the phoneme.

a hypothesised transcription corresponds to an observed acoustic vector sequence can therefore be estimated by the SPM as:

$$-\log(P_a) \propto \sum_{vectors} \sum_{i=1}^{24} \left[K_s \left(\log(2\pi\sigma^2_{s_{i,p}}) + E_{s_{i,p}} \right) + K_e \left(\log(2\pi\sigma^2_{e_{i,p}}) + E_{e_{i,p}} \right) \right] \tag{8.7}$$

where P_a is proportional to the actual acoustic probability as before. As in the case of the HMM-based system, a language model probability P_l weighted by a grammar scale factor G_s can be combined with this expression to obtain an overall logarithmic probability:

$$\log(P_{spm}) = \log(P_a) + G_s \log(P_l) \tag{8.8}$$

Finally, rather than discarding the HMM-produced logarithmic probability scores associated with each transcription in an N-best list, and re-ordering the list based on the SPM-produced score alone, these two scores can be combined in a weighted sum:

$$\log(P_{comb}) = (1-\beta)\log(P_{htk}) + \beta \log(P_{spm}) \tag{8.9}$$

so that when $\beta = 0$ the HTK-computed score alone is used and when $\beta = 1$ only the SPM-generated score is employed, and β is a constant whose optimum value must be determined empirically[10].

8.3.2 Boundary Alignment Errors

Since the phonetic alignments automatically generated by HTK are imperfect, the phonetic labels associated with the observed acoustic vectors are often incorrect in the region of phonetic boundaries. The mis-placement of these boundary positions by one or more frames frequently leads to large acoustic error contributions from the frames concerned. These errors arise from presenting the articulatory input vectors for incorrectly labelled frames to the inputs of the wrong MLPs, and from computing $E_{e_{i,p}}$ using the incorrect statistics. As a result, both $E_{s_{i,p}}$ and $E_{e_{i,p}}$ are initially dominated by errors within one frame of phonetic boundaries.

This problem is particularly pronounced in the case of stops and nasals, where the acoustic signal either has low energy, is rapidly changing, or both. For example, even a one-frame error in the position of a boundary demarcating the release of a stop closure (and hence the onset of the associated burst waveform) will lead to a large spectral error in one of the two frames adjacent to the boundary.

To alleviate this problem, boundaries delimiting stops and nasals are permitted to shift by one frame during the error computation if this results in a local error reduction. To move a phonetic boundary in this way would nominally require the generation of a new set of articulatory trajectories and the re-synthesis of the corresponding acoustic vectors for each possible boundary configuration. Such an approach would involve an impractical

[10] Since neither the HMM-produced or SPM-produced "probability" estimates represent actual probabilities, the magnitudes of P_{htk} and P_{spm} will in general be highly dissimilar.

amount of computation however, and hence an approximate technique is used which avoids this computational overhead.

A boundary shift of one frame to the right of the HTK-designated position between two neighbouring phonemes p_1 and p_2 involves the deletion of the first frame of phoneme p_2 and its replacement with an additional p_1 frame added to the end of the existing p_1 sequence. As a first approximation to a suitable set of coefficients to use for this new frame, the coefficients of the vector at the end of the previous p_1 sequence can be copied into the new vector position[11]. Using this technique, a one-frame shift of each stop and nasal boundary to the left or right is hypothesised. If one of these shifts would result in an error reduction and the other in an error increase, the boundary is considered to be mis-aligned, and is moved one frame in the appropriate direction. In all other cases the boundary's position is left unchanged. The result of this approximate re-alignment process is a significant reduction in the contribution to the total error from frames in the vicinity of phonetic boundaries.

8.4 Assessing Recognition Performance

This section defines the measures used to evaluate the performances of both the HMM-based speech recogniser and the combined HMM-SPM recognition system.

8.4.1 Performance Measures

A simple measure of the performance of a recognition system is the percentage of labels correctly identified, computed as:

$$\frac{100\,(L - D - S)}{L} \tag{8.10}$$

where L is the total number of labels, D is the number of labels in the correct transcription which did not appear in the recognition sequence ("deletions") and S is the number of substituted labels. Since spurious labels can also be introduced into the recognition sequence however, it is more usual to report the percentage "accuracy", defined as:

$$\frac{100\,(L - D - S - I)}{L} \tag{8.11}$$

where I is the number of these insertions. This latter performance measure was employed in all cases, where word-level labels were selected as the recognition units used.

8.4.2 Homonyms

The use of a null grammar in both the SPM and HTK recognition algorithms implies that random errors may be introduced as a result of homonyms in the task vocabularies. For example, the three words "to", "two" and "too" are acoustically indistinguishable, and in the absence of a grammar to preferentially select between them according to the

[11] Numerous alternative approximation methods could be employed here: for example, a linear extrapolation from the last two p_1 vectors.

context, a label selection will be made arbitrarily during recognition. This effect gives rise to a random component in the error rate, which is undesirable when attempting to compare the acoustic modelling accuracies of the two systems based on their recognition performances alone.

Accordingly, homonyms such as these were grouped into equivalence classes during recognition, in an attempt to remove these random effects. Although this approach removes the majority of the random component in the error rate, a similar effect caused by *phrases* with identical phonetic transcriptions remains. For example, the much-quoted phrases "I scream", "eyes cream" and "ice-cream" each have identical (or nearly identical) acoustic realisations, but are not rendered equivalent by the simple homonym equivalences described above. Fortunately however, in the relatively small task vocabularies used in the UW and RM corpora, such acoustically equivalent phrases are very rare, and the use of homonym equivalences suffices[12].

8.5 Recognition Results

In this section the results of using the SPM to re-order N-best hypothesised transcription lists generated by HTK are presented, for data drawn from both the UW and RM corpora. In each case, the error rates achieved when re-scoring the transcriptions using both the SPM log probability measure and the combined measure given in Equation 8.9 are presented. In addition, typical recognition results obtained using the simple probabilistic model described in Appendix B are illustrated.

8.5.1 University of Wisconsin Data

The baseline ($N=1$) word recognition accuracies achieved by the speaker-dependent HTK systems for each of the UW speakers are listed in Table 8.1. The HMMs used were 5-mixture monophone models—ie 5 distinct Gaussian distributions[13] were used to model the 39 acoustic parameters in each of the 3 "emitting" states of each HMM[14]. Separate HMMs were used to represent the 47 *phonemes* in Tables 5.4 and A.1, and additional models were provided for silences, inter-word spaces[15], and the "tone" and background comment models described in Section 5.4.1[16]. This yields a total of ≈ 60000 parameters used in the UW HMM-based acoustic models.

In each case the grammar scale factor G_s was set to a value of 11, which was empirically

[12] One technique for removing all such random effects from the computed recognition accuracy would be to report *phoneme* recognition errors rather than word recognition errors. Since phrase-level effects are not significant in the UW and RM databases however, a word-level system coupled with homonym equivalences was retained here.

[13] Each characterised by two parameters specifying the mean and variance of the distribution respectively.

[14] Or two emitting states in the case of stop and diphthong phonemes.

[15] The inter-word space, or "sp" model contains only one emitting state, by contrast with the 3 emitting states used in the standard models.

[16] A small number of additional parameters were also used to model the transition probabilities between the component states of each of these HMMs.

Speaker	*Baseline accuracy*
jw16	78.9%
jw27	69.2%
jw29	74.4%
jw18	80.1%
jw24	65.5%
jw45	70.3%

Table 8.1: Baseline word recognition accuracies for UW speakers, using speaker-dependent HTK models.

found to yield the best recognition results[17]. As can be seen from the table there is considerable variation in the recognition performance between speakers—in this case, up to a 15% difference in the absolute recognition accuracies. Variations such as these are frequently encountered in speech recognition tasks, and arise from differences in factors such as the speaking rate, intelligibility and recording environment[18].

SPM-based Re-scoring

Figure 8.2 shows a plot of the recognition results obtained by using the SPM-computed log probabilities to re-order transcriptions up to a depth of $N = 100$ for the six UW speakers. A grammar scale factor of 30 and a $K_s : K_e$ ratio of 5 were used in each case[19], and the articulatory trajectories were synthesised using the explicit co-articulation model described in Chapter 3.

In this figure, recognition accuracy is plotted against the depth, M, of the hypothesis list for four separate recognition systems, where $1 \leq M \leq N$. In the plot for each speaker:

1. The uppermost curve represents the best possible recognition accuracy which can be achieved at a given depth M by hand-selecting the optimum transcription n, where $1 \leq n \leq M$. This curve is monotonically increasing, since as M increases there is an increasing chance of observing the correct transcription at a depth less than or equal to M. This provides an upper limit for the performance that could be achieved by any re-scoring algorithm.

2. The lowest curve represents the *mean* performance obtained over 100 repetitions of selecting a transcription n at random for $1 \leq n \leq M$ at a given depth M. This curve

[17] Ideally the optimum value of G_s should be determined based on the recognition performance achieved on a separate cross-validation set. Due to insufficient data in the corpora used in this initial evaluation however, G_s values were determined based on test set performance. While this is expected to yield slightly higher recognition rates than would otherwise be achieved, only the relative recognition rates achieved by HTK and the SPM are of interest to this study.

[18] Although the recording environment was nominally identical for each of these speakers, slight differences in variables such as the microphone placement relative to the lips and nostrils are inevitable.

[19] Once again, these parameters should ideally be determined based on the performance on a separate cross-validation set, but in this case they were obtained based on test set performance. Further performance increases could be obtained by optimising separate phoneme-specific values for the ratio $K_s : K_e$ (Section 9.5.3).

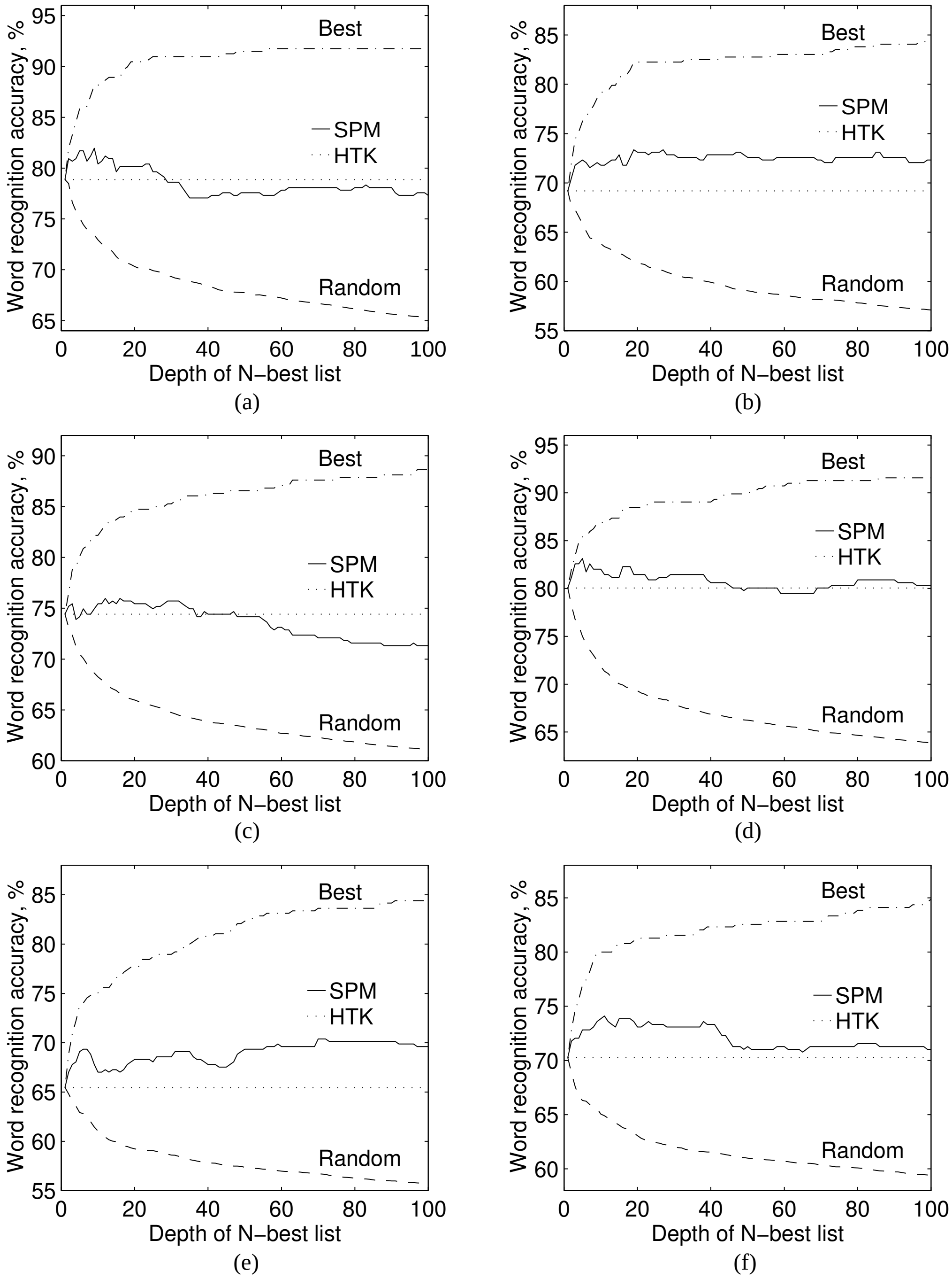

Figure 8.2: Word recognition accuracy as a function of the depth of the N-best list used as input to the SPM for the UW speakers: (a) jw16 (b) jw27 (c) jw29 (d) jw18 (e) jw24 (f) jw45. In addition to the performances achieved by HTK and the SPM, the best achievable performances and the performances achieved by randomly selecting transcriptions are also shown.

will in general decrease monotonically, since as M is increased, the number of errors *on average* in the M^{th} transcription is expected to increase.

3. The HTK curve simply represents the baseline performance as a reference for assessing the performance of the other models. It is independent of M since the HTK system always selects the $M = 1$ transcription.

4. The fourth curve represents the recognition accuracy achieved by re-ordering the transcription lists according to the log probability scores computed by the SPM.

A significant improvement in recognition accuracy is observed using SPM-based re-scoring as M increases from 1 to 5 for each speaker except jw29—for whom an improvement is not obtained until $7 \leq M \leq 12$. The performance for the speakers jw16, jw29, jw18 and jw45 tails off as M increases to 100, and in the case of jw16 and jw29 the recognition accuracy for $M = 100$ is actually worse than that of HTK.

In each case the curves appear to be converging to the "true" recognition performance that would be achieved by the SPM if all possible hypotheses were re-scored (ie as the total number of transcriptions available, $N \to \infty$). While the performance of the SPM is significantly better than that achieved by randomly selecting transcriptions from the N-best list, it reliably matches or exceeds the performance of HTK only for values of M less than 30—and in the case of jw29 the performance is actually worse than that of HTK for some values of M in this range.

Finally, the re-scoring performance of the SPM appears to be *negatively* correlated with the absolute recognition performance of HTK. This is demonstrated by the larger and more sustained recognition accuracy improvements obtained for jw27, jw24 and jw45, for whom the baseline recognition accuracy achieved by HTK is relatively poor. This result is not surprising, since the poor recognition performance achieved by HTK for these speakers is assumed to be attributable to relatively poor acoustic modelling; hence it is more likely that the SPM will be able to provide improved acoustic discrimination. The re-scoring algorithm therefore appears to have the desirable property of being more effective for those speakers for whom it is most needed.

Simplified Probabilistic Model

The deterioration in recognition performance observed as $M \to N$ for the majority of the speakers in Figure 8.2 can be explained by considering the expected performance of the simplified probabilistic re-scoring algorithm described in Appendix B.

Following the development in this appendix, let M denote the depth of the list re-ordered so far, and k be the index of the best transcription hypothesised by the probabilistic model for $1 \leq k \leq M$, with associated error E_k. Then the probability of selecting the $(M+1)^{th}$ transcription as the most likely hypothesis when the depth of the list to be re-scored is increased by one, is determined by the choice of the constants used in Equations B.3 and B.4. As an example of the typical recognition performance obtained using such a model, these values were set to:

$$P\left(\text{choose}(M+1) \mid E_{M+1} < E_k\right) = 0.5 \tag{8.12}$$

$$P\left(\text{choose}(M+1) \mid E_{M+1} > E_k\right) = 0.02 \tag{8.13}$$

These conditional probabilities were then used to compute the values of P_{better} and P_{worse}, which are the probabilities that the recognition accuracy will increase or decrease as a result of the addition of the $(M+1)^{th}$ transcription, respectively. Figure 8.3 shows a plot of the average recognition performance obtained from 100 simulation runs, in which this simplified probabilistic model was used to re-score the transcriptions for the speaker jw18 as M was increased from 1 to 100. As was the case in Figure 8.2, the best possible recognition accuracy, the baseline HTK performance and that obtained over 100 iterations of random transcription selection are graphed for reference purposes.

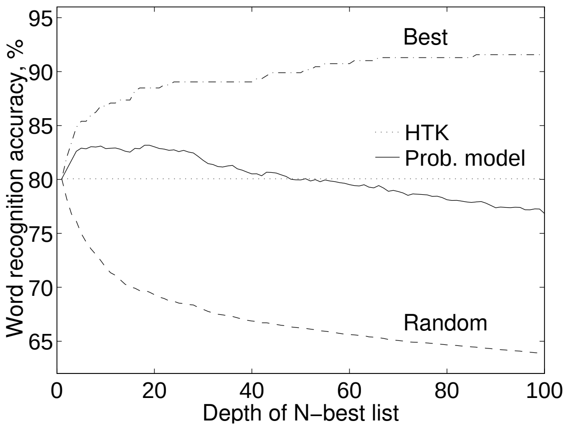

Figure 8.3: Word recognition accuracy as a function of the depth of the N-best list used as input to a simplified probabilistic model of transcription re-scoring for jw18. In addition to the performances achieved by HTK and the probabilistic model, the best achievable performance and the performance achieved by randomly selecting transcriptions are also shown.

As can be seen from the graph, the performance of the probabilistic model exceeds that of HTK up to $M \approx 50$, but is inferior for larger values. For low values of M (eg. $2 \le M \le 5$), the value of $P\left(E_{M+1} < E_k\right)$ is relatively high, so that $P_{better} > P_{worse}$ and recognition accuracy increases. For high values of M however, $P\left(E_{M+1} < E_k\right)$ becomes very low, and hence $P_{better} < P_{worse}$ and recognition performance deteriorates. In this example, performance improves rapidly for $2 \le M \le 5$, is approximately steady for $5 \le M \le 20$, and degrades thereafter.

In a practical re-scoring algorithm, the value of $P(\text{choose}(M+1) \mid E_{M+1} > E_k)$ should decrease as M becomes very large, so that performance ceases to degrade as $M \to N$, where N is the total number of transcription hypotheses available. This is in agreement with the results observed for jw18 in Figure 8.2, which exhibit a gain in accuracy compared with HTK for $2 \le M \le 5$, steady or deteriorating performance for $5 \le M \le 40$ and approximately constant recognition accuracy for larger values of M.

Combined HTK and SPM-based Re-scoring

A significant performance improvement can be obtained by combining the separate logarithmic probability scores computed by HTK and the SPM according to Equation 8.9. Since these scores are typically of very dissimilar magnitudes, this necessitates the empirical determination of an additional parameter β, which was also obtained in this case by assessing the algorithm's performance on the test data[20], due to the absence of a cross-validation set[21].

The results obtained for each of the UW speakers, including the β values used, are illustrated in Figure 8.4. In this figure, the baseline HTK performance curves and the best achievable recognition curves are shown as before, but the accuracy achieved by randomly selecting transcriptions has been omitted for clarity. In each case two combined HTK-SPM system curves are plotted, where these correspond to synthesising articulatory trajectories in the SPM both with and without the explicit co-articulation model, respectively.

When the HTK-computed scores are combined with those predicted by the SPM, two effects are clearly apparent:

1. There is no longer a significant degradation in recognition performance as M increases.

2. The recognition improvements obtained for low values of M are substantially higher than those achieved using the SPM-generated scores alone[22].

As was the case in Figure 8.2, the majority of the improvement in recognition accuracy is observed for $2 \leq M \leq 5$. In addition, with the exception of the speaker jw29, the rate of improvement in this range is close to the best that could be achieved by any re-scoring algorithm. For values of M greater than 10 however, little or no further improvement is seen, since HTK and the SPM appear to be making similar acoustic scoring errors. The absence of a continued degradation in recognition performance at large M is due to the very low probability scores assigned to these transcriptions by HTK, which prevent them from erroneously being selected as the best transcription hypotheses during re-scoring.

With the exception of the speakers jw27 and jw45, the recognition accuracies achieved when re-scoring from articulatory trajectories synthesised with the explicit co-articulation model are higher than the corresponding results obtained without this model. In the case of the two former speakers, the differences between the recognition performances observed with and without the co-articulation model do not differ significantly. The results obtained therefore indicate that the more accurate articulatory modelling achieved using the explicit

[20]Kannan has proposed techniques for automatically optimising the parameters used to combine multiple probability scores for N-best hypothesis re-ordering, based on the minimisation of the average word error [77]. In this case however, suitable values were determined by a random sampling of parameter space.

[21]Schwartz et al. have suggested that a cross-validation set containing as many as 300 utterances may be required to determine reliable values for the parameters used to combine the various acoustic and language model probability scores in such a re-scoring algorithm [153].

[22]This is consistent with the N-best re-scoring performance reported by other authors, who found that a re-scoring algorithm whose performance was worse than that of a standard HMM-based system when used alone, yielded an increase in word accuracy when combined with the HMM-generated scores [5, 119].

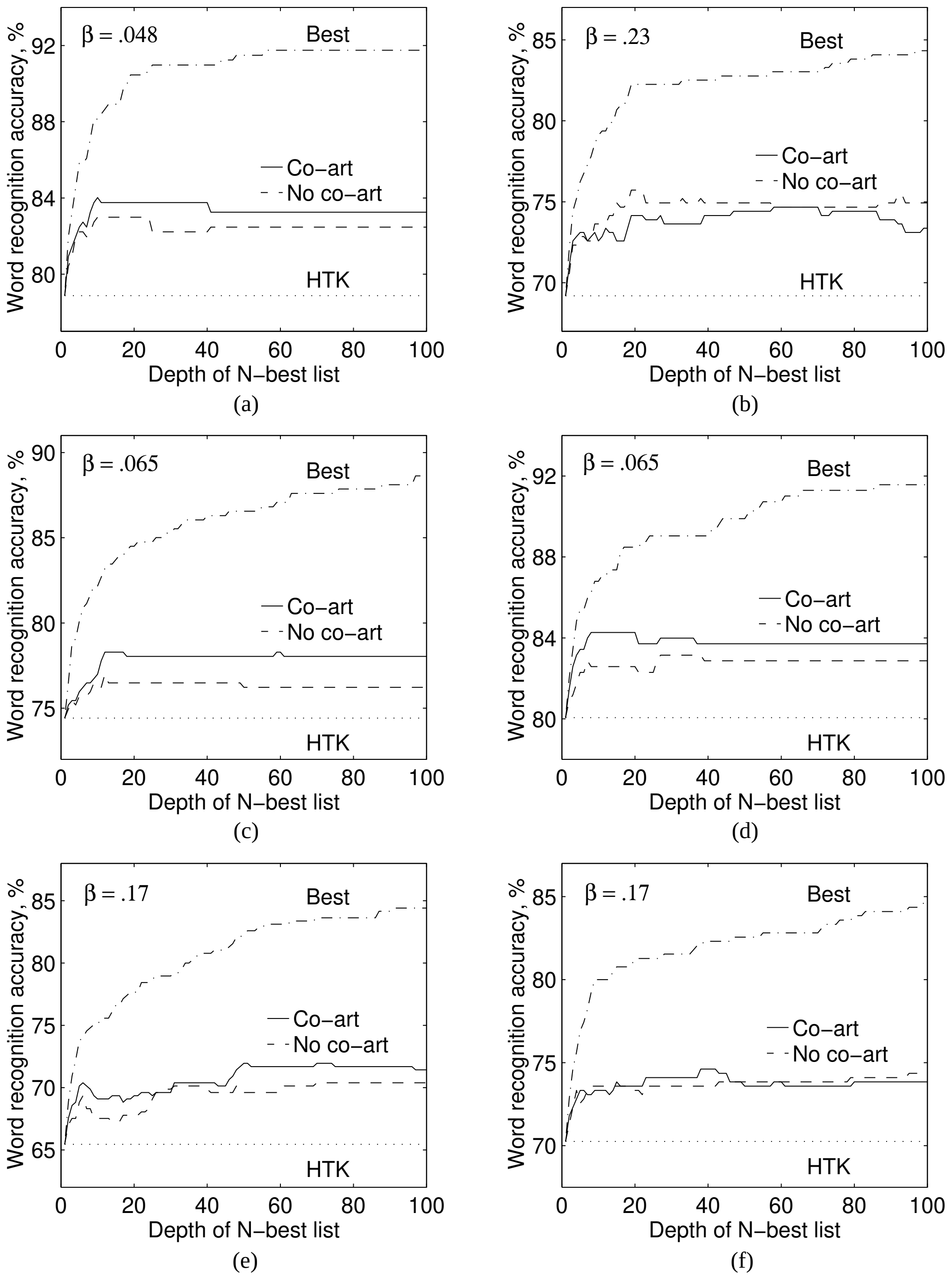

Figure 8.4: Word recognition accuracy as a function of the depth of the N-best list used as input to the combined HTK-SPM re-scoring algorithm, for the UW speakers: (a) jw16 (b) jw27 (c) jw29 (d) jw18 (e) jw24 (f) jw45. The performances of the model using articulatory trajectories synthesised both with and without the co-articulation model are illustrated, along with the baseline HTK performances and the best achievable performance curves.

co-articulation model results in improved recognition accuracy when used as a basis for the re-scoring of N-best transcriptions.

Combined HTK-SPM weighting parameter

The optimum value of β for each speaker is negatively correlated with the baseline recognition performance achieved by HTK. Thus the β values for jw27, jw24 and jw45—for whom the baseline recognition rates are relatively low—are significantly larger than the corresponding values for the other speakers. This is to be expected from the nature of the combined score $\log(P_{comb})$ defined in Equation 8.9, in which the value of β determines the relative weight that is given to the SPM-computed scores. When the recognition performance achieved by HTK is relatively poor, there is greater scope for improving this performance through the use of the SPM-generated log probabilities, as illustrated in Figure 8.2. Thus the optimum values of β will be higher in these cases, reflecting an increased confidence in the SPM-computed scores relative to those generated by HTK.

The value of β used will also affect the number of utterances whose transcription lists are re-ordered by the re-scoring algorithm. For very low values of β the combined score will be very similar to the HTK-produced log probability, and relatively few transcriptions are likely to be re-ordered. As β increases, it is expected that a corresponding increase in the number of transcriptions which are re-ordered will be observed. The number of transcriptions re-ordered in this way for the data from each of the UW speakers are listed in Table 8.2, for both the optimum value of β and the use of the SPM score alone ($\beta = 1$).

Speaker	*Fraction re-ordered, optimum* β	*Fraction re-ordered,* $\beta = 1$
jw16	37.3%	86.3%
jw27	72.5%	82.4%
jw29	51.0%	86.3%
jw18	43.1%	74.5%
jw24	76.5%	88.2%
jw45	66.7%	84.3%

Table 8.2: Percentages of test set utterances for each UW speaker whose N-best transcription lists were re-ordered using both the combined HTK-SPM probability scores (optimum β) and the SPM scores alone ($\beta = 1$).

As expected, the number of utterances whose hypothesis lists are re-ordered is higher when using the SPM score alone than when using the optimum value of β to combine the HTK and SPM scores. In addition, the higher optimum values of β for jw27, jw24 and jw45 lead to a greater number of re-ordered transcriptions than is the case for the other speakers.

By varying the value of β from 0 to 1, the sensitivity of the recognition performance to the relative weights attached to the HTK and SPM scores can be determined. Figure 8.5 shows a plot of the recognition accuracy achieved as a function of β for the re-scoring of an N-best list of depth 10 for the speaker jw18, using articulatory data synthesised with the explicit co-articulation model.

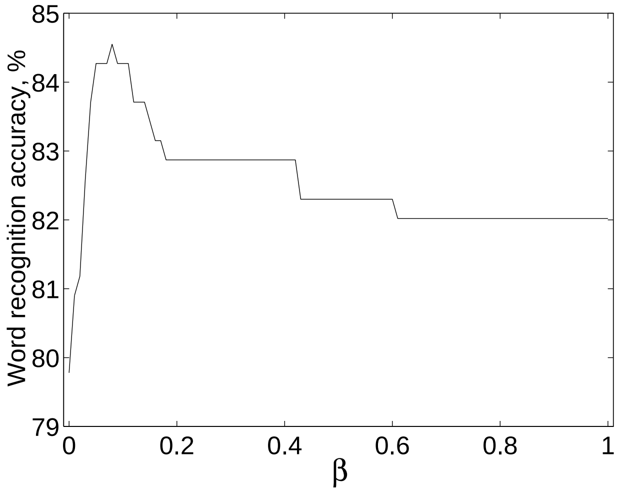

Figure 8.5: Word recognition accuracy as a function of β, the weighting constant for the HTK and SPM scores, where $\beta=0$ corresponds to using the HTK score alone, and $\beta=1$ to using only the SPM score. The curve shown corresponds to re-scoring an N-best list of depth 10 for the speaker `jw18`, as beta is stepped from 0 to 1 in intervals of 0.01.

When $\beta=0$ the baseline HTK performance is obtained. As β increases, an increasing weight is placed on the SPM-generated score, and recognition accuracy increases sharply until $\beta\approx0.07$. Accuracy subsequently declines more steadily, to a value corresponding to the use of the SPM score alone. While there is a relatively narrow band of β values which yield recognition accuracies above 83%, *any* value of $\beta>0$ yields improved performance compared with HTK for an N-best list of this depth.

The final relative error rate reductions achieved by re-scoring $N=100$ transcriptions for each of the 51 test set utterances[23] for each of the UW speakers are listed in Table 8.3. The articulatory data were synthesised using the explicit co-articulation model, and combined HTK-SPM probability scores computed from the optimum β values were employed in the re-scoring algorithm. In each case a significant reduction in the recognition error rate was obtained.

8.5.2 Resource Management Data

The baseline word recognition accuracies achieved by the speaker-dependent HTK systems for the RM speakers `das1` and `tab0` are listed in Table 8.4. The grammar scale factor G_s was set to 13 for `das1` and 17 for `tab0`, and the resulting word recognition errors are considerably less than those listed in Table 8.1 for the UW data.

This significant performance improvement is largely due to the more accurate acoustic modelling provided by the RM HTK models. A total of 5654 distinct cross-word tree-clustered triphone models were used in the RM system, each with 3 emitting states[24] and

[23] Comprising a total of 390 words.

[24] Two emitting states for stops and diphthongs.

Speaker	*Baseline HTK error*	*Combined HTK-SPM error*	*Relative Improvement*
jw16	21.1%	16.8%	20.4%
jw27	30.8%	26.6%	13.6%
jw29	25.6%	22.0%	14.1%
jw18	19.9%	16.3%	18.1%
jw24	34.6%	28.6%	17.3%
jw45	29.7%	26.2%	11.8%

Table 8.3: Relative reductions in word recognition error rates for the UW speakers, achieved over the baseline HTK performance by using combined HTK-SPM scores.

Speaker	*Baseline accuracy*
das1	91.51%
tab0	91.43%

Table 8.4: Baseline word recognition accuracies for RM speakers, using speaker-dependent HTK models.

a single Gaussian distribution to model each of the 39 acoustic parameters. Including the parameters used to model transition probabilities between states, this yields a total of ≈ 1.2 *million* parameters in the RM acoustic models. This represents a 20-fold increase in the number of parameters used compared with the UW system, and hence the baseline recognition performance achieved is greatly improved.

SPM-based Re-scoring

The results obtained when using the SPM-computed log probabilities to re-score utterance hypothesis lists of length $N=100$ are shown in Figure 8.6. A $K_s:K_e$ ratio of 10 was used for each speaker, and grammar scale factors of 25 and 15 were used for das1 and tab0 respectively. A higher $K_s : K_e$ ratio is used for the RM speakers compared with the UW speakers since more data are available to train the MLPs in the RM corpus, and the corresponding acoustic error variances used to compute $E_{s_{i,p}}$ are lower; hence greater significance is attached to the log probabilities computed based on $E_{s_{i,p}}$ than $E_{e_{i,p}}$. As in the case of the UW data, the values of these parameters were empirically determined based on test set recognition performance. The articulatory representation used for each speaker was that generated from the speaker-independent UW statistics using the explicit co-articulation model.

While the error rates achieved by the SPM are consistently above those achieved by randomly selecting transcriptions, there is only a very slight improvement in recognition accuracy in both cases for $2\leq M\leq 5$. In general, the performance achieved by the SPM is almost uniformly inferior to that of HTK for both speakers.

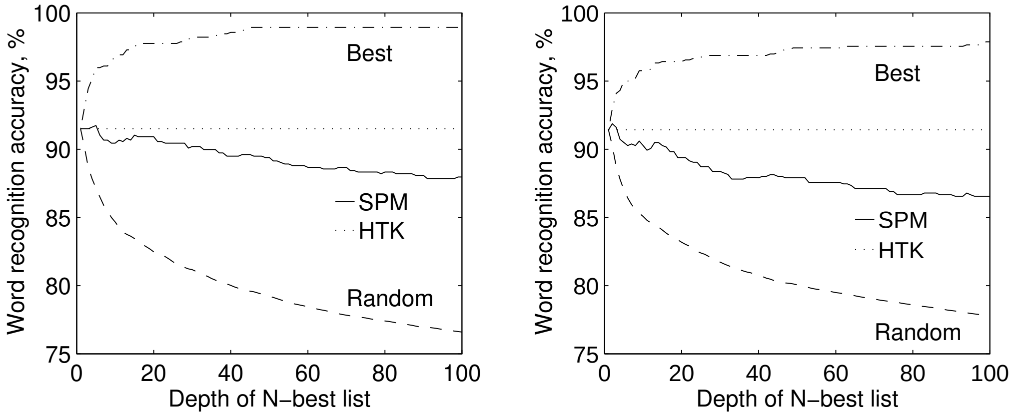

Figure 8.6: Word recognition accuracy as a function of the depth of the N-best list used as input to the SPM for (a) das1 (b) tab0. In addition to the performances achieved by HTK and the SPM, the best achievable performances and the performances achieved by randomly selecting transcriptions are also shown.

Simplified Probabilistic Model

Figure 8.7 shows a plot of the recognition performance obtained by using the simplified probabilistic model described in Appendix B to re-score transcriptions up to a depth of $N = 100$ for the speaker das1, where the conditional probabilities in Equations 8.12 and 8.13 have been retained in this example.

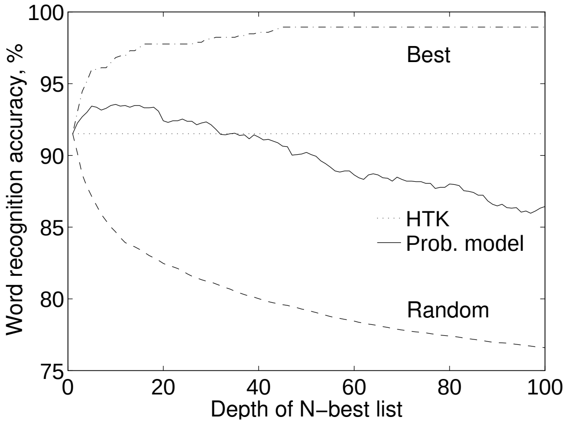

Figure 8.7: Word recognition accuracy as a function of the depth of the N-best list used as input to a simplified probabilistic model of transcription re-scoring for das1. In addition to the performances achieved by HTK and the probabilistic model, the best achievable performance and the performance achieved by randomly selecting transcriptions are also shown.

As was observed in Figure 8.3, the recognition accuracy improves for $2 \leq M \leq 5$. In this case however, the performance starts to degrade at a lower depth ($M \approx 20$), and the rate of degradation is also considerably increased. This behaviour is due to the decreased probability of encountering the best transcription of an utterance at any given depth

$M > 1$, since many more utterances are correctly transcribed by the $M = 1$ entries in the case of the RM data, as compared with the UW results.

Combined HTK and SPM-based Re-scoring

The results obtained for each of the RM speakers using combined HTK-SPM log probability scores, including the β values used, are illustrated in Figure 8.8. Although $N = 100$ transcriptions were re-scored for each test set utterance[25], the plots show only the results obtained up to $M = 20$, in order to provide additional detail at low values of M; both the combined system curves remain unchanged for $M > 20$.

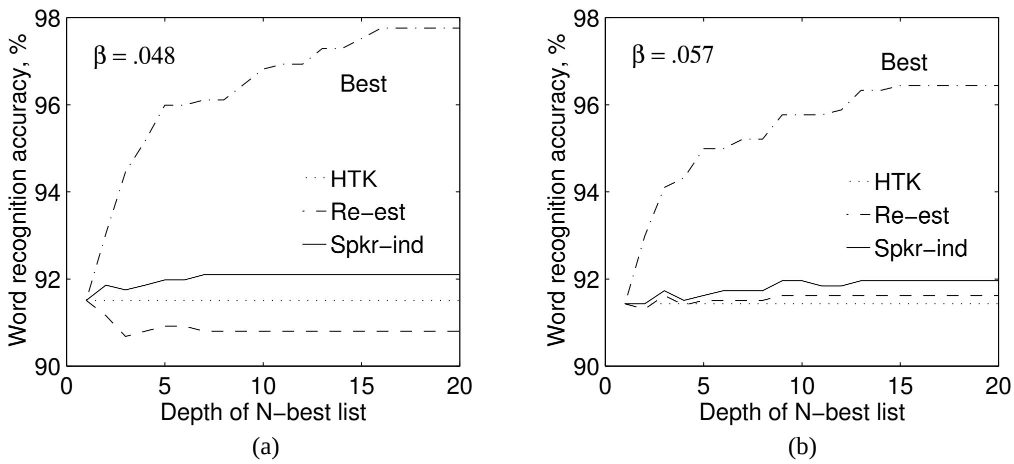

Figure 8.8: Word recognition accuracy as a function of the depth of the N-best list used as input to the combined HTK-SPM re-scoring algorithm, for (a) `das1` (b) `tab0`. The performances of the model using articulatory trajectories generated from both re-estimated synthetic statistics and the speaker-independent UW statistics are illustrated, along with the baseline HTK performances and the best achievable performance curves.

In this figure, the baseline HTK word recognition accuracies and the best achievable recognition curves are indicated, but the accuracies achieved by randomly selecting transcriptions have once again been omitted for clarity. In each case two combined system curves are plotted. These correspond to synthesising articulatory trajectories in the SPM from re-estimated synthetic articulatory statistics and speaker-independent X-ray statistics drawn from the UW data respectively. In each case the explicit model of co-articulation was used during trajectory generation.

As was the case for the UW data, the recognition performance of the combined HTK-SPM system exceeds that of the SPM-based system alone. For both of the speakers shown the error rate achieved using articulatory trajectories synthesised from speaker-independent X-ray statistics is less than the corresponding error using re-estimated synthetic statistics. This difference is consistent with the significantly lower acoustic errors obtained at the outputs of the MLPs used to approximate the acoustic vectors from speaker-independent X-ray articulatory data, as illustrated in Figures 7.23 and 7.24.

[25] This was true of all the test set utterances with the exception of two utterances for `tab0`, for which only 2 and 39 hypotheses were generated respectively, due the small number of words in their transcriptions.

For both speakers the combined system trained on speaker-independent X-ray data achieves an improvement in recognition accuracy compared with HTK. The magnitudes of these improvements are less than those achieved for the UW data, supporting the hypothesis that greater performance improvements are achieved when the absolute recognition accuracy achieved by HTK is low. Relatively small values of β are used, and once again the reduction in recognition error rates is achieved when $M < 10$, while recognition accuracy is steady for larger values of M.

Combined HTK-SPM weighting parameter

The number of transcriptions re-ordered for the data from each speaker are listed in Table 8.5, for both the optimum value of β and the use of the SPM score alone ($\beta = 1$). In each case the values correspond to the use of speaker-independent articulatory statistics for generating articulatory trajectories. As observed with the UW data, the number of utterances whose hypothesis lists are re-ordered is higher when using the SPM score alone than when using the optimum value of β to combine the HTK and SPM scores.

Speaker	*Fraction re-ordered, optimum* β	*Fraction re-ordered,* $\beta = 1$
das1	10%	68%
tab0	16%	75%

Table 8.5: Percentages of test set utterances whose N-best transcription lists were re-ordered by the combined HTK-SPM re-scoring algorithm for the RM speakers. Values are given for the optimum β value and for $\beta = 1$.

Figure 8.9 shows a plot of the recognition accuracy obtained as a function of the parameter β for the re-scoring of an N-best list of depth 10 for the speaker das1, using articulatory data synthesised using the explicit co-articulation model from speaker-independent X-ray statistics.

When $\beta = 0$ the baseline HTK performance is obtained. As β increases, an increasing weight is placed on the SPM-generated score, and improved recognition accuracy is obtained for $0 < \beta < 0.4$. Recognition performance subsequently declines for larger values of β, to a level significantly below that achieved by HTK when $\beta = 1$.

The final relative error rate reductions achieved by re-scoring $N = 100$ transcriptions for each of the 100 test set utterances[26] comprising 848 and 898 words for das1 and tab0 respectively are listed in Table 8.6. The articulatory data were generated using the explicit model of co-articulation from speaker-independent X-ray statistics, and the combined HTK-SPM probability score computed using the optimum β value was used to re-order the transcriptions. In each case a reduction in the relative error rate is achieved, although the magnitudes of these reductions are considerably less than those achieved for the UW corpus.

[26] Except in the case of the shortened lists for the two tab0 utterances as described previously.

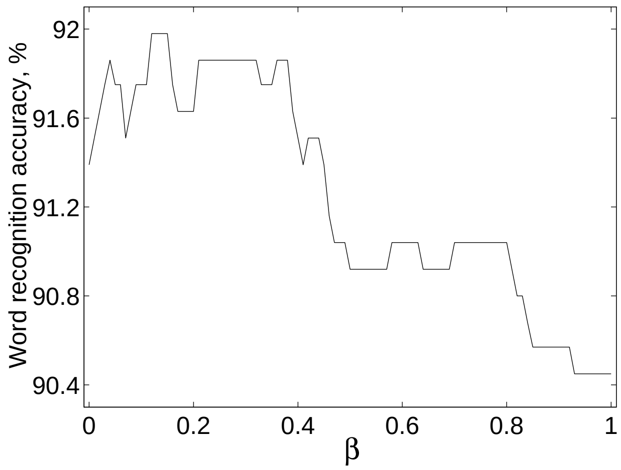

Figure 8.9: Word recognition accuracy as a function of β, the weighting constant for the HTK and SPM scores, where $\beta = 0$ corresponds to using the HTK score alone, and $\beta = 1$ to using only the SPM score. The curve shown corresponds to re-scoring an N-best list of depth 10 for the speaker das1, as beta is stepped from 0 to 1 in intervals of 0.01.

Speaker	*Baseline HTK error*	*Combined HTK-SPM error*	*Relative Improvement*
das1	8.49%	7.90%	6.9%
tab0	8.57%	8.06%	6.0%

Table 8.6: Relative reductions in word recognition error rates for the RM speakers, achieved over the baseline HTK performance by using combined HTK-SPM scores.

Chapter 9

Summary and Conclusions

9.1 Introduction

This dissertation has described the design and implementation of a novel self-organising articulatory speech production model suitable for use in a continuous speech recognition framework. In this chapter, the design and implementation of the component models are briefly reviewed, and the results obtained from evaluations performed on two separate speech databases are summarised. Finally, some recommended future areas of research are identified, and a concluding overview of the dissertation is presented.

9.2 Articulatory Model

The approach taken to articulatory modelling in the SPM can be summarised as follows:

- The model focuses on articulatory positions at the midpoints of phonemes, which are characterised by smoothly-varying parametric pdfs during system training, and directly predicted from time-aligned phonetic strings during articulatory synthesis.
- Systematic co-articulatory effects are assumed to be motivated principally by a desire to achieve an economy of articulatory movement during periods of maximum articulatory effort, within allowable perceptual constraints.
- The amount of articulatory effort required in a particular context is proportional to the acceleration of the articulator concerned. This acceleration is estimated by a simple approximation to the curvature of an articulatory trajectory.
- Correlations between this curvature measure and the distribution of positional values are used to predict co-articulated articulatory positions at the midpoints of phonemes, from which articulatory trajectories are constructed by linear interpolation.

The model uses a set of exemplar articulatory trajectories to automatically extract suitable values for each of its parameters, and the resulting system is used to predict articulatory movements corresponding to time-aligned phonetic strings.

9.2.1 Initial Articulatory Trajectories

In the UW corpus, explicit X-ray articulatory trajectories are provided which can be used to train the articulatory models. By contrast, the RM corpus provides only acoustic data, and hence initial articulatory trajectories were synthesised for this data set both from speaker-independent statistics extracted from the UW corpus, and by using dynamic programming to invert an articulatory-acoustic codebook.

Synthetic Trajectories from Codebook Inversion

A codebook containing approximately 100,000 [articulatory vector, acoustic vector] pairs was generated by using a set of synthetic articulatory parameters to define piece-wise constant vocal tract shapes. These shapes were used to control an explicit Kelly-Lochbaum vocal tract model, through which the appropriate excitation signals were propagated to synthesise the corresponding acoustic waveforms.

Dynamic programming was then used to determine the most probable articulatory vector sequence given an acoustic vector string, based on both the acoustic and articulatory errors. The resulting synthetic articulatory trajectories yielded poor discrimination between the phonemes, and an improved articulatory set was therefore computed using linearised Kalman filtering.

9.2.2 Positional Characterisation

The articulatory parameters were characterised by sampling their positions at the midpoints of phonemes, the temporal locations of which were determined from automatically-generated phonetic alignments. As a first approximation to articulatory behaviour, the positional variations observed in each articulatory parameter were modelled independently, and single Gaussian distributions were used to model positional variability at the midpoints of phonemes for each [articulator, phoneme] pair. The resulting probabilistic distributions provide an appealing model of articulatory movement compared with fixed point targets and window models, since the need for context-dependent targets is greatly reduced or eliminated, and a statistically-based model of co-articulation based on most probable articulatory positions is possible.

The positional sample distributions for the UW X-ray data and the re-estimated synthetic RM data were generally found to be unimodal. When the Kolmogorov-Smirnov statistic was used to assess the probability that the observed data were drawn from the hypothesised parametric distributions however, the use of a single Gaussian to model each data set was found to be an accurate representation in only a minority of cases, due to skews on the sample data sets.

Sources of Variability

Both random and systematic effects give rise to the positional variations observed in the resulting samples, and the articulatory model seeks to characterise those systematic components which are correlated with the phonetic sequence. These in turn arise from both prosodic and co-articulatory effects, and only the latter are explicitly modelled in the system, while the former are treated as additional random effects since they are more strongly correlated with the phrase structure than the identities of the individual words uttered.

9.2.3 Explicit Co-articulation Model

Co-articulation is the allophonic variation in a phoneme according to its phonetic context. Both carryover and anticipatory co-articulatory effects are observed in speech, whereby the articulatory constraints of preceding and following sounds influence the articulation of a given sound, respectively. These context-driven positional variations are assumed to be motivated principally by a desire to achieve an economy of articulatory movement during periods of maximum articulatory effort. They are constrained in magnitude by the degree of specification of the articulator concerned, since the perceptual constraints for the phoneme being produced must also be satisfied.

The articulatory effort required in a particular context is therefore assumed to be correlated with the achieved articulatory position. The required effort itself is proportional to the acceleration of the articulator in question, which in the SPM is modelled independently of those of the other articulators as a first approximation. Relative accelerations are estimated from the relative *curvatures* of the articulatory trajectories, which are approximated by computing the differences between the linear interpolation gradients leading into and out of the midpoints of the phonemes.

Correlations between Positions and Curvatures

Correlations between articulatory positions and accelerations (and hence trajectory curvatures) are then used to predict co-articulatory positional variations at the midpoints of phonemes. Strong correlations were observed between these two variables at the midpoints of phonemes for most [articulator, phoneme] pairs. Both positive and negative correlations were observed for the RM data, since the articulatory trajectories in this case were estimated using linearised Kalman filtering, and were not based on physiological constraints. For the UW corpus however, the correlations were almost uniformly positive, and were interpretable in terms of the expected degree of articulatory specification. For both the UW and RM data sets, the significance of each correlation coefficient was assessed using Student's T-test, and insignificant coefficients were set to zero before using the correlations to predict positional variations.

9.2.4 Synthesis of Trajectories

Complete articulatory trajectories were synthesised from time-aligned phonetic strings by specifying co-articulated articulatory positions at the midpoints of the phonemes, and interpolating linearly between them. This technique provides an implicit model of co-articulatory effects, since the phonetic context will strongly influence the rate of change in an articulator's position leading into and out of a phoneme. Even in the absence of the explicit model of co-articulation at the midpoints of phonemes, strong co-articulatory effects will therefore be observed.

A direct assessment of the articulatory modelling accuracy achieved was only possible for the UW data set, as reference articulatory trajectories were unavailable for the RM corpus. Articulatory trajectories were therefore synthesised both with and without using the explicit model of co-articulation, for each of the UW test set utterances. For all six UW speakers, the error contribution from each articulator decreased when the explicit co-articulation model was applied, and the overall articulatory error computed on an utterance-by-utterance basis was also reduced in every case.

9.3 Acoustic Model

The rôle of the acoustic model is to predict parameterised acoustic vectors from synthetic articulatory trajectories. The key features of the acoustic model described in this dissertation are the provision of:

- A robust partially non-linear mapping from articulatory space to acoustic space which is used to predict the most likely parameterised acoustic vector sequence corresponding to a set of articulatory trajectories.
- A characterisation of the expected acoustic variability at each frequency and at each point in time, in terms of both the raw log spectral energy and the distribution of energy as a function of frequency within each acoustic vector.
- A separate mapping from articulatory vectors to acoustic vectors for each phoneme in the input set.
- A modelling paradigm in which the values of the parameters of each acoustic model were extracted automatically during optimisation on a training data set.

Separate multi-layer perceptron models were used to approximate the mapping from articulatory vectors to acoustic vectors for each phoneme.

9.3.1 Data Preparation

The articulatory input data were normalised to a mean of zero and a standard deviation of one before presentation to the networks. This reduced the need for scaling of the input values by the first layer of the MLP, and restricted the magnitudes of the MLP's parameters to relatively small values.

The acoustic signal was parameterised using 24 Mel-scaled log spectral coefficients for each speech frame. The relatively high dimensionality of these acoustic vectors permitted accurate acoustic matching to be performed, and the use of a parameterisation in the log frequency domain was chosen to approximate the signal analysis which takes place in the human cochlea.

The log spectral target vectors were also scaled, by subtracting the mean of each individual vector from its coefficients. This energy normalisation was used to remove biases on the raw acoustic vectors, which result from changes in the intensity of a speaker's voice. The method was most successful in removing biases where the energy was distributed across the frequency range (such as in vowels) than when it was concentrated in a frequency sub-band (such as in the unvoiced fricatives). The expected raw log spectral energy at each frequency for a given phoneme was modelled separately, by computing the mean and variance of each log spectral vector coefficient over all of the acoustic training vector examples for each phoneme.

9.3.2 Network Training

A separate set of MLPs was used to approximate the acoustic mapping to normalised log spectral vectors for each phoneme[1]. This greatly reduced the complexity of the modelling task for each acoustic model, and obviated the need to explicitly specify the excitation sources during synthesis, since these are implicitly defined by the phoneme's identity.

The parameters of the MLPs were optimised using the RPROP algorithm, and cross-validation was used to prevent over-fitting to the training data. The number of hidden nodes used in each network's hidden layer was selected based on the network's performance on the cross-validation set, and three networks were trained on the data for each phoneme from different random initialisations. The outputs of each set of three networks were averaged when predicting normalised log spectral vector values, and the error variances at the network outputs were computed over the training sets and used to characterise the uncertainty in these predicted values[2].

9.3.3 Acoustic Vector Prediction

The performance of the acoustic models in predicting normalised acoustic vectors was assessed by computing a weighted sum of the root mean squared errors obtained on each phoneme-specific data set. For each of the UW speakers, the MLP mappings were found to outperform regression systems trained using a least mean squared error criterion. In every case, averaging the outputs of three separate networks led to improved accuracy compared with using a single network alone.

MLP mappings were trained on articulatory trajectories synthesised for each of the utterances in the UW corpus both with and without the explicit model of co-articulation. The data incorporating the co-articulation model yielded a reduced acoustic error for every speaker except jw45, for whom the two errors were very similar. The acoustic errors obtained when training the networks on both synthetic and X-ray articulatory data were also compared for one UW speaker. Similar errors were obtained in each case, with the models trained on synthetic data generated with the explicit co-articulation model actually achieving *less* error than the equivalent X-ray system on the test data. This last effect was found to be due to outliers in the X-ray articulatory data set, which were excluded by the parametric models used in the synthetic case.

Acoustic models were then optimised for the RM corpus, using both the synthetic re-estimated articulatory data and articulatory trajectories generated from speaker-indepen-

[1]Except in the case of the UW corpus, in which some of the individual phonetic data sets were merged with acoustically similar sets before training the models.

[2]These error variances were also used in the re-estimation of articulatory positions using linearised Kalman filtering for the synthetic RM data set.

dent UW statistics. Significantly greater acoustic accuracy was obtained using the speaker-independent UW statistics, indicating that this technique was more successful than the inversion of the articulatory-acoustic codebook for generating articulatory trajectories for speakers for whom X-ray data are unavailable.

9.4 Recognition System

The SPM was used to augment the performance of the HTK speaker-dependent continuous speech recognition system according to the following algorithm:

- HTK was used to generate N hypotheses as to the probable transcription of each test set utterance, along with their phonetic time-alignments.
- From each of these transcriptions a parameterised acoustic representation was synthesised by the SPM, comprising both the most probable acoustic vector sequence and the associated variance predictions.
- The parameterised acoustic vectors corresponding to each of the N hypothesised transcriptions were compared with those of the input speech, and the transcription which yielded the best acoustic match was chosen as the most likely hypothesis.

The goal of using the SPM in this way was to provide a more accurate description of contextual effects in the acoustic signal by explicitly modelling co-articulation in the articulatory domain. The conventional system was retained in this framework as it provides both an efficient method for searching for likely hypotheses, and a technique for generating initial probability estimates for each of the resulting transcriptions.

9.4.1 Generation of N-best Lists

The HMM-based system used both an acoustic model and a language model to generate transcription hypotheses. Since the goal of implementing the combined recognition system was to compare the acoustic modelling accuracy of HTK with that of the SPM, a null grammar was used when generating these N-best hypothesis lists. The language model scores associated with the grammar were scaled by an empirically-determined constant before being combined with the corresponding acoustic model scores, to reduce the number of spurious words which were otherwise inserted into the transcriptions produced.

The N-best hypotheses were generated from word-level lattices, and the test set utterances in the UW corpus were hand-segmented into sentence-length acoustic files before performing recognition. A total of 100 transcription hypotheses were generated for the 51 test utterances per speaker in the UW corpus, and the 100 test utterances for each of the two RM speakers `das1` and `tab0`.

Aligning the Transcriptions

An alignment of the transcriptions at the sub-phonemic level was required to identify the component phones of stop and diphthong phonemes. The HMMs used to model these phonemes were therefore modified to contain just two emitting states, which align to the closure and burst sections of stops and the initial and final voiced sections of diphthongs, respectively. The most probable alignment of each transcription to the individual *states* of the corresponding HMMs was then automatically generated, and used to identify these phone boundaries.

9.4.2 Re-scoring Transcriptions

The SPM was used to re-synthesise parameterised acoustic vectors corresponding to each transcription, which were then compared with a parameterised version of the original speech waveform. Two separate acoustic probability measures were computed, which were combined in a weighted sum to yield an overall acoustic score. The first of these was computed by comparing the normalised log spectral vectors predicted by the MLPs with the normalised vectors corresponding to the original speech waveform. The squared differences between these vectors were scaled by the error variances at the MLP's outputs, and converted to log probability scores.

The second measure was introduced to model the expected unscaled, or "raw" log energy in each spectral vector, which was not modelled by the MLPs. The mean and variance of each log spectral vector were computed over all of the vectors in the training data set on a phoneme-by-phoneme basis. During acoustic scoring, the spectral vectors corresponding to the original speech were then compared with the mean log spectral vectors for the phonemes concerned, and probability scores were computed based on the differences between the observed vectors and these means.

As was the case for the HMM-based system, the resulting acoustic scores were combined with those predicted by a language model. The N-best transcription hypotheses were then re-ordered, either based on the SPM-produced log probability scores alone, or using a weighted sum of these scores and those originally estimated by HTK.

Compensation for Alignment Errors

Due to errors in the HTK-generated phonetic alignments, frames were occasionally mislabelled in the regions of phonetic boundaries. The result of this effect was a disproportionate contribution to the acoustic error from predicted acoustic vectors within one frame of these boundary locations. This effect was particularly pronounced in the case of stops and nasals, hence an approximate technique was developed for re-aligning the boundaries of these phonemes by up to one frame. As a result of this boundary re-alignment algorithm, the contributions to the total error from these vectors were reduced to levels similar to those from other frames.

9.4.3 Recognition Results

Recognition performance was assessed by measuring the word recognition accuracies of the various systems. Homonyms in the task vocabularies were grouped into equivalence classes during recognition, since the null grammar was unable to distinguish between them, and they otherwise gave rise to a random component in the computed error rates.

University of Wisconsin Corpus

The baseline word recognition accuracies achieved by HTK on the test set utterances for the six UW speakers varied from a low of 65.5% to a high of 80.1%. The N-best lists corresponding to each of these utterances were initially re-ordered based on the SPM-computed log probability scores alone. Significant increases in recognition accuracy were observed for $2 \leq N \leq 5$ for all speakers except jw29, but recognition performance was either steady or declined for larger values of N in each case. The relatively poor performance at high values of N can be explained in terms of the decreasing probability of observing the optimum transcription as N increases, as demonstrated by the use of a simplified probabilistic re-scoring model.

The recognition accuracies achieved by the SPM at $N = 100$ were significantly higher than those achieved by HTK for two of the speakers, were significantly less for one speaker, and were comparable for the remaining three. In addition, the performance of the SPM was negatively correlated with that of HTK, so that performance increases were generally greater for speakers for whom the baseline HTK recognition accuracy was relatively low.

The transcription hypothesis lists were then re-ordered based on a weighted combination of the SPM and HTK probability scores, where the weighting coefficient used was empirically determined[3]. Significantly improved recognition accuracies were obtained, and the performance gains achieved did not decline as N increased. In the majority of cases higher recognition accuracy was obtained when using articulatory data generated using the explicit co-articulation model than without it, and maximum relative reductions in the word error rates of between 10% and 20% were obtained across the six speakers.

Resource Management Corpus

The performance increases achieved on the RM data set were significantly smaller than those obtained for the UW corpus, reflecting the much greater acoustic modelling accuracy achieved by the RM HMMs. When the SPM-generated scores alone were used to re-order the N-best lists, only very slight performance improvements were observed at low values of N, and the SPM's recognition accuracy at $N = 100$ was significantly less than that of HTK for both speakers.

The performance achieved using combined HTK-SPM probability scores was significantly better than the performance achieved using SPM-generated scores alone. SPM systems were trained on articulatory data generated using both re-estimated synthetic articulatory statistics and speaker-independent X-ray statistics, and improved recognition accuracy was obtained in the latter case for both speakers[4]. Using the combined HTK-SPM scores and an SPM system trained from articulatory trajectories generated from speaker-independent X-ray articulatory statistics, maximum relative error rate reductions of 6.9% and 6.0% were obtained for the two RM speakers studied.

[3]The recognition results were found to be sensitive to the value of this weighting coefficient when the proportion of the SPM-generated score used was small, but were relatively insensitive to it when the HTK-generated score was assigned the smaller weight.

[4]This result is as expected from the performance of the MLPs used in the RM acoustic models.

Number of Parameters

The articulatory models used in the SPM comprise 224 parameters to model the means and variances of the positional and curvature distributions over 56 phonemes for each articulator. A further 56 correlation coefficients are provided per articulator to yield a total of 280 parameters over 16 articulators, or 4480 parameters in the articulatory models overall. The MLPs used in the acoustic model for the UW corpus contained a total of $\approx 21,000$ parameters for each random initialisation across the 52 phonemes[5], to give a total of $\approx 63,00$ parameters in the acoustic models used to approximate log spectral shapes. A further 48 parameters per phonetic data set were also used to model raw log spectral vector magnitudes. The total number of parameters in the SPM trained on the UW corpus was therefore approximately 67,000, which represents an increase of approximately 10% over the $\approx 60,000$ parameters used in the corresponding HTK system.

The articulatory model used in the SPM for the RM corpus contained the same number of parameters as the UW SPM system, but due to the larger average size of the MLP hidden layers used on the RM data, approximately 131,000 parameters were used in the RM acoustic models, to yield a total of approximately 135,000 parameters. This represents almost a 90% reduction compared with the 1.2 million parameters used in the HTK RM system.

9.5 Future Work

The application of articulatory speech production models to the task of automatic speech recognition is a field which is far from mature. A review of many attempts to incorporate articulatory information into speech recognition algorithms was given in Chapter 2, and much research remains to be carried out before the true potential of approaches such as these can be determined. In this section some areas are identified which should prove fruitful in the further development of the self-organising system described in detail in this dissertation.

9.5.1 Articulatory Model

While the articulatory model presented in Chapter 3 has been shown to be successful in capturing some of the systematic variability in articulatory positions due to co-articulatory effects, this model is limited in many ways. Firstly, the assumption of independence of articulatory movements is clearly implausible and there is considerable scope for the development of alternative articulatory parameterisations[6]. The potential advantages to be gained in this area include:

- The ability to model the inter-dependencies of articulators and the resulting compensatory articulations.

[5]The reduced number of distinct phonetic sets in the acoustic models for the UW corpus are due to the merging of 4 phonetic data sets with other acoustically similar sets.

[6]Relatively straightforward approaches to deriving articulatory parameterisations of reduced dimension include principal components analysis and linear discriminant analysis.

- The identification of a transformation of the basic Cartesian articulatory parameter space to a co-ordinate system in which co-articulatory effects are more naturally represented.

Secondly, the current model explicitly predicts articulatory parameter values only at the midpoints of phonemes, while a more complete description of articulatory positional variability would specify articulatory positions at all times during the production of an utterance. Future research into biologically-plausible articulatory models may have much to offer in this area, since current biomechanical systems are typically limited to descriptive models of articulatory movement and generally rely on the hand-tuning of parameters. In the absence of a predictive, trainable biological model of articulation, the system described in this dissertation seeks to predict co-articulatory variation from a statistically-based descriptive model of articulatory positions. Refinements which could be made to this positional model include:

- The use of more complex distribution shapes to model non-Gaussian articulatory positional samples.
- The use of multi-modal distributions where multiple articulation strategies are possible.
- A systematic investigation of the validity of using the midpoints of phonemes to characterise articulatory positions, and the possible identification of phoneme-specific sampling points.

Given a descriptive model, the SPM attempts to predict deviations from mean articulatory positions from a knowledge of the time-aligned phonetic sequence alone. In the current model this is achieved by using a simple approximation to the local curvature of an articulatory trajectory. Further work in this area could include:

- A more accurate measure of articulatory effort than the current linear gradient-based approach.
- The use of a larger phonetic context than the immediately neighbouring phonemes when determining the most probable articulatory movements.
- The identification of additional features which are correlated with positional variations, such as articulatory velocities or prosodic effects.

Once features which are correlated with systematic articulatory positional variation have been identified—such as the simple curvature-based estimate of articulatory effort currently used—a more sophisticated technique for predicting co-articulated positions would be likely to result in increased modelling accuracy. Specifically:

- The degree of co-articulatory deviation from mean articulatory positions need not be symmetrical with regard to the preceding and following contexts, so that differing emphases could be attached to carryover and anticipatory co-articulatory effects, respectively.

- Similarly, physiological limitations in the vocal tract lead to asymmetrical deviations from mean articulatory positions in terms of lateral or vertical movements, and this should be reflected in the degree of co-articulatory variation predicted by the model.

There is considerable scope for future research into the optimisation of articulatory representations. In Section 7.4.1 a technique for re-estimating articulatory positions at the midpoints of phonemes was briefly described, in which an inversion of the acoustic model was used within a linearised Kalman filtering algorithm to derive updated articulatory trajectory estimates. Future research along these lines could include:

- Iterative re-estimation of the parameters of the articulatory and acoustic models until the optimum joint articulatory-acoustic performance is obtained.

- Re-estimation of articulatory positions at all sample points, rather than simply at the midpoints of phonemes.

- Relaxation of the constraint that the articulatory parameters model physiological structures in the vocal tract, to permit an appropriate articulatory representation (or partial representation) to be *learned*. Preliminary experiments into the use of the acoustic model in the SPM to hypothesise such an additional articulatory variable have been described, although the general utility of this technique has yet to be demonstrated [13].

Finally, the dissimilar acoustic modelling results obtained on the RM corpus when using two different approaches to synthesising articulatory data from an acoustic representation imply that the nature of the articulatory representation used in the absence of X-ray training data is of considerable importance. For the techniques described in this dissertation to have wider applications for speech recognition applications, reliable methods for deriving accurate articulatory representations for speakers for whom X-ray articulatory data are unavailable are required. Possible further research in this area could include:

- The use of improved articulatory-acoustic inversion techniques: eg. by hand tuning the parameters of the vocal tract model used to construct the articulatory-acoustic codebook, so that the resulting acoustic waveforms more closely match those of the target speaker.

- Investigation of alternative techniques for developing speaker-independent articulatory representations from X-ray data sets.

- The further development of techniques for adapting an articulatory representation to a particular speaker, eg. using the iterative re-estimation techniques suggested above.

Despite the considerable advances made to date, there remain many unsolved problems in the area of articulatory modelling, and it appears likely that it will be many years before a reliable, accurate predictive model of articulatory movements will be developed.

9.5.2 Acoustic Model

The acoustic model used in the SPM was chosen on the basis that it offers a self-organising, robust partially non-linear mapping in which acoustic variability is explicitly modelled. The disadvantage of this approach is that it fails to make use of information concerning the nature of the human vocal tract. Future research into this component of the model might therefore include:

- The development of automatically-trainable explicit vocal tract models, both in terms of the derivation of a suitable set of control parameters *and* the automatic adjustment of the parameters controlling the geometry of the tract.
- Investigations into the provision of a technique for characterising acoustic variability in an explicit vocal tract model, either directly in acoustic space or indirectly through the specification of articulatory uncertainty.

Given that a self-organising implicit acoustic model is to be used until such an explicit vocal tract model is available, numerous alternative approaches to providing acoustic vector predictions are possible. Firstly, in terms of the sub-division of the articulatory input space:

- A smaller number of distinct acoustic models could be employed, for example via the use of separate models for broad phonetic categories rather than phoneme-specific models. The combination of data sets would reduce the problem of discontinuities at phonetic boundaries, and should lead to a reduction in the total number of parameters in the system though parameter sharing. The drawback with such an approach is the corresponding increase in the complexity of the individual acoustic models.
- A larger number of distinct models could be used, for example by using a hierarchical mixture of experts to infer a further sub-division of the data for each phoneme, in an analogous manner to the individual states used in HMM-based approaches.

In addition, different models could be used for the acoustic mappings themselves, such as:

- Bayesian ANNs, which use hyper-parameters instead of cross-validation to prevent over-fitting during training.
- Recurrent neural networks, which could be used to model the inter-dependence of successive articulatory vectors presented to each network mapping.
- Linear models which avoid the problems of over-fitting and non-robustness inherent in least mean squared error regression.

Finally, alternative acoustic output vector representations could be used:

- The use of MFCC parameters would provide a more compact spectral representation, in which energy normalisation is automatically provided for all but the lowest-order acoustic parameter. The disadvantages of such an approach are that it leads to a more complex acoustic mapping, and to acoustic errors which are difficult to interpret.

- Linear predictive coding (LPC) coefficients could be employed, since these represent the response of an all-pole filter approximation to a vocal tract transfer function, and hence are closely related to the parameters which directly control the vocal tract shape. This technique is less appropriate for representing unvoiced speech frames however, and does not provide the opportunity for Mel-scaling the frequency range to match human auditory perception.

- Non-static acoustic parameters could be synthesised, such as the "delta" parameters commonly used in HMM-based systems. The aim of the articulatory model in the SPM is to capture the salient characteristics of articulatory motion however, so that dynamic aspects of the spectral representation need not be explicitly modelled by the acoustic model. Nevertheless, the provision of these parameters might serve to partially compensate for any inadequacies in the articulatory model.

As was the case for the articulatory model, there are few aspects of the acoustic model which might not benefit from further research, whether self-organising implicit models are retained or more explicit knowledge-based approaches are implemented.

9.5.3 Recognition System

The defining characteristics of the combined HTK-SPM recognition system are the lengths of the transcription units which are re-synthesised and the techniques used to score the resulting synthetic acoustic signals. Possible modifications to the transcription units include:

- The use of fewer words in each utterance transcription. This would lead to increased efficiency, as a given depth N of the list of hypotheses generated will produce greater linguistic diversity for shorter utterance lengths, and hence an increased chance of observing the correct transcription.

- Alternatively, transcription *lattices* themselves could be re-scored. This would lead to a further increase in efficiency, as the individual sub-lattices typically provide relatively few word confusions[7].

The acoustic scoring technique currently used provides separate models for raw log energy coefficients and energy-normalised spectral shape coefficients. Scaled acoustic differences are converted to log probabilities, and an approximate boundary re-alignment

[7] Consider for example a very simple lattice for a two word utterance in which five different alternatives are postulated for each word. In an N-best approach there are $5 * 5 = 25$ possible transcriptions, but if the initial and final sub-lattices are separately re-scored then only $5 + 5 = 10$ different comparisons need be made, where each of these comparisons is approximately half the length of the 25 utterance-level comparisons.

algorithm is used. Once again, there are numerous alternative strategies which could be investigated. For example:

- An acoustic error based not only on static acoustic vectors but also on dynamic spectral features could be used. As stated previously, this approach is somewhat counter-intuitive in the SPM as dynamic characteristics are intended to be modelled in the articulatory domain.
- The acoustic parameterisation used for acoustic comparisons need not be that directly synthesised by the acoustic model, but may be one derived from it. For example, more complex auditory scene analysis techniques could be used, in which energy concentrations in the acoustic signal are explicitly modelled.
- The boundary re-alignment algorithm could be extended to investigate boundary shifts of greater than one frame, or additional context-dependent shifts where the HTK-produced alignment was frequently observed to be inaccurate.
- The technique used to combine acoustic scores based on energy-normalised log spectral shapes and raw log spectral energies could be refined by optimising phoneme-specific weight parameters.
- The quality of the HTK-generated alignment itself could be improved, by providing a secondary alignment pass at a higher frame rate where the acoustic features are rapidly changing. In this way the duration and dynamics of events such as stop bursts could be more accurately characterised.

Finally, the predicted articulatory representation itself could be used directly in a recognition algorithm, for example by augmenting the acoustic observation vectors presented to an HMM-based system with articulatory parameter vectors[8].

9.6 Conclusions

This dissertation has described the design and implementation of a novel self-organising articulatory speech production model suitable for use in a continuous speech recognition framework. The principal features of the system are:

- The provision of an explicit predictive time-domain model of the articulatory mechanism in general, and co-articulation in particular.
- The synthesis of a probabilistic parameterised acoustic representation of the speech signal.
- The computation of logarithmic probability scores from a comparison of the synthesised acoustic output with that of an input utterance.

[8] Alternatively the articulatory representation alone could be used to decode phonetic labels, as discussed in Chapter 2.

- The ability to extract all of the parameters of the articulatory and acoustic models automatically from a set of training exemplars.

The explicit predictive model of co-articulatory variation yields a significant increase in articulatory modelling accuracy over the use of interpolation between mean articulatory positions alone, and good approximations to X-ray articulatory trajectories can be achieved.

In the absence of X-ray data, synthetic articulatory trajectories suitable for training the system can be generated either by inverting an articulatory-acoustic mapping, or from speaker-independent statistics extracted from the available X-ray data. The results of evaluating these two techniques indicate that the latter approach yields a considerably more useful articulatory representation in terms of the prediction of acoustic vectors.

The model was applied to the task of re-scoring N-best utterance transcriptions generated by HTK. The results obtained using articulatory models trained from X-ray data indicate that given an appropriate articulatory representation, significant improvements in word recognition accuracy are possible. More importantly, statistics extracted from the X-ray data for a small set of speakers provide a suitable starting point for the prediction of articulatory movements from textual transcriptions alone. Preliminary results indicate that recognition performance improvements can be gained by using such a technique for speakers for whom X-ray data are not available.

The fact that these performance improvements are obtained using relatively simple articulatory and acoustic models—the latter production model uses only 10% of the number of parameters of the HTK system—indicates that acoustic modelling accuracies significantly better than those achieved by HMM-based systems should be attainable through further optimisation of the system.

The recognition results obtained using the self-organising articulatory production model described in this dissertation therefore support the hypothesis that the future of computer speech recognition lies not within a purely acoustic paradigm, but rather at the interface between the articulatory and acoustic domains.

Appendix A

Phoneme Set

The basic phoneme set used in this dissertation is derived from that originally described by Lee [93]. This set comprises the 47 distinct phonetic units listed in Table A.1.

This table lists each of the phonemes, along with an example of their use in American English words, and their corresponding IPA symbols. Apart from the standard vowels, diphthongs, stops[1], fricatives and nasals, three additional phonetic categories are represented:

- The alveolar flap `/dx/` is explicitly included. This phoneme is itself a stop, and is articulated by a fast tap of the tongue tip on the alveolar ridge.
- The compound `/ts/` = `/t/`+`/s/` is included as a separate aspirated fricative.
- The unreleased stops `/dd/`, `/kd/`, `/pd/` and `/td/` are also explicitly represented, where these phonemes occur only word-finally.

As described in Section 5.3.2, this basic phoneme set is modified before being used in the self-organising model of speech production:

- Each voiced and unvoiced stop phoneme is modelled as an obligatory closure followed by an optional burst, so that explicit word-final closure phonemes are no longer required.
- The diphthongs, the fricative `/ts/` and the alveolar flap `/dx/` are also each represented by two sub-phones, where both of these are mandatory in every context.

The presence or absence of a stop burst is not specified in the phonetic dictionary, but is determined during the recognition or alignment of the acoustic waveform. For example, if the final stop in an instance of the word "hit" is released, the `/t/` phoneme will be aligned as `/t.1 t.2/` representing the separate closure and burst. If the `/t/` is unreleased however, it is replaced by the single phone `/t.1/` during recognition or alignment.

[1] The phonemes referred to here as "stops" could more specifically be categorised as "plosives", as they comprise both a stop in the oral cavity and its subsequent "burst", or release.

Phoneme	Example	Symbol
	Vowels	
/aa/	cot	/ɑ/
/ae/	bat	/æ/
/ah/	but	/ʌ/
/ao/	bought	/ɔ/
/ax/	about	/ə/
/eh/	bet	/ɛ/
/er/	bird	/ɜr/
/ih/	bit	/ɪ/
/ix/	beating	/ɜ/
/iy/	beat	/i/
/uh/	book	/ʊ/
/uw/	boot	/u/
	Diphthongs	
/aw/	bout	/aʊ/
/ay/	bite	/aɪ/
/ey/	bait	/eɪ/
/ow/	boat	/o/
/oy/	boy	/ɔɪ/
	Liquids	
/l/	led	/l/
/r/	red	/r/
/w/	wed	/w/
/y/	yet	/j/
	Unvoiced Stops	
/k/	cat	/k/
/p/	pat	/p/
/t/	tat	/t/

Phoneme	Example	Symbol
	Voiced Stops[2]	
/b/	bet	/b/
/d/	debt	/d/
/dx/	ladder	/ɾ/
/g/	get	/g/
	Closures[3]	
/dd/	hid	/d/
/kd/	hick	/k/
/pd/	hip	/p/
/td/	hit	/t/
	Strong Fricatives	
/dh/	that	/ð/
/jh/	judge	/ʤ/
/ts/	its	/ts/
/v/	van	/v/
/z/	zoo	/z/
	Weak Fricatives	
/ch/	church	/ʧ/
/f/	fan	/f/
/hh/	hat	/h/
/s/	sue	/s/
/sh/	shoe	/ʃ/
/th/	thin	/θ/
	Nasals	
/m/	met	/m/
/n/	net	/n/
/en/	button	/n/
/ng/	thing	/ŋ/

Table A.1: The RM dictionary phoneme set as defined by Lee [93], along with standard IPA symbols.

[2] The alveolar flap /dx/ is included as a voiced stop for simplicity.

[3] The closures are represented in the IPA alphabet as standard plosive stop phonemes which are unreleased.

Appendix B

Probabilistic Re-scoring Model

In this appendix a simple probabilistic model for selecting amongst N-best transcription hypotheses is presented, which can be used as a first approximation to the re-scoring behaviour which might plausibly be expected to result when re-ordering N-best lists.

Suppose that N different transcriptions are available for a particular utterance, and that the first M of these are to be re-ordered, where $1 \leq M \leq N$. Once these M entries have been re-ordered, the recognition performance of the system will be determined by the identity of that transcription which is selected by the re-ordering algorithm as the most probable. To predict the system's performance it is therefore only necessary to determine which of the M transcriptions will be chosen as this most probable entry.

Assume now that the re-ordering system has identified the $n = k$ entry as the most probable transcription up to a depth M in the list, where $1 \leq k \leq M$. Since the hypothesised re-ordering algorithm is imperfect, this may not in reality represent the most accurate transcription up to depth M, but the argument is not affected. Consider then adding the $(M+1)^{th}$ transcription to the list of entries to be re-ordered, thus extending the depth of the hypothesis list by one. The recognition performance of the re-ordering system based on these $(M+1)$ entries will differ from the performance of the depth-M system only if:

1. The $n = M+1$ entry is chosen by the re-ordering system as more likely than the current best estimate, $n = k$.

2. The number of errors in the $n = M+1$ entry, E_{M+1}, is different to the number of errors in the $n = k$ entry, E_k.

This implies that the change in recognition performance (if any) caused by extending the depth of the list by one entry will be determined by the two probabilities:

$$\begin{aligned} P_{better} &= P\left(\text{choose}(M+1) \mid E_{M+1} < E_k\right) . P(E_{M+1} < E_k) && \text{(B.1)} \\ P_{worse} &= P\left(\text{choose}(M+1) \mid E_{M+1} > E_k\right) . P(E_{M+1} > E_k) && \text{(B.2)} \end{aligned}$$

which are defined in terms of the conditional probabilities of selecting the $n = M+1$ entry over the $n = k$ entry, given that the number of errors in the former entry is less than or greater than the number of errors in the latter entry, respectively. In any practical re-scoring algorithm—such as the SPM described in this dissertation—these probabilities will be a function of the error magnitudes E_{M+1} and E_k. As a first approximation to plausible re-scoring behaviour however, a simplifying assumption can be made by setting each of these conditional probabilities to constant values. For improved recognition performance, the values chosen must satisfy the constraints:

$$P\left(\text{choose}(M+1) \mid E_{M+1} < E_k\right) \gg 0 \tag{B.3}$$

$$P\left(\text{choose}(M+1) \mid E_{M+1} > E_k\right) \ll 1 \tag{B.4}$$

The expected recognition performance of such a re-scoring algorithm can then be predicted by performing multiple random simulations in which the depth of the list of hypotheses is iteratively extended from 1 to N. At each step:

- If the number of errors in the new hypothesis entry is equal to the number of errors in the current best estimate, do nothing.
- Otherwise, determine whether the new entry will be selected as the new best estimate using the pre-determined values for the probabilities in Equations B.3 and B.4.

The values of these conditional probabilities in reality will not be independent of M, E_{M+1} and E_k; as a result, the re-scoring behaviour of a practical re-scoring implementation will deviate from this simplified model. For example, $P\left(\text{choose}(M+1) \mid E_{M+1} > E_k\right)$ will decrease as M increases, since for large M it is likely that $E_{M+1} \gg E_k$ and these very poor transcriptions will seldom be selected in practice. By contrast, a proportion of these transcriptions will continue to be selected by the simplified model described above, and performance will deteriorate as M increases to large values.

Appendix C

Vocal Tract Model

C.1 Introduction

This appendix describes the design and implementation of an explicit vocal tract model which can be used to generate acoustic waveforms from a specification of the vocal tract area function and the corresponding excitation signals.

The task of modelling the human vocal tract is a complex one, but can be greatly simplified if several assumptions are made, the most common of which are [151]:

1. The vocal tract can be straightened out and hence approximated as a variable-area tube.
2. The wave motion in the tract is planar, ie pressure and velocity are constant in a plane perpendicular to the straightened axis.
3. The linear wave equation is valid.

The vocal tract model used here is a variant of the Kelly-Lochbaum model [81], which uses these assumptions to simulate the propagation of one or more excitation signals through a vocal cavity represented by a series of cylindrical tubes of fixed cross-sectional area.

The input excitation signals are explicitly defined in the time domain, and the transfer function of the tract model is computed in terms of the reflection coefficients between adjacent area function segments. The implementation of the model closely follows that described by Rubin [145], with small differences in the treatment of impedances at the glottis and lips, and the excitation signal for voiced fricatives.

C.2 Wave Propagation in Uniform Lossless Tubes

This section provides a brief description of plane wave behaviour at the boundaries of adjacent uniform lossless tubes. More detailed derivations can be found in Flanagan [44] and Rabiner and Schafer [132].

The simplifying assumption is made, that during speech production a plane wave propagates along the axis of a concatenated series of lossless, uniform area tubes. The motion of such a wave within any given tube is then described by the partial differential equations:

$$
\begin{aligned}
-\frac{\partial p(x,t)}{\partial x} &= \frac{\rho}{A} \cdot \frac{\partial u(x,t)}{\partial t} && \text{(C.1)}\\
-\frac{\partial u(x,t)}{\partial x} &= \frac{A}{\rho c^2} \cdot \frac{\partial p(x,t)}{\partial t} && \text{(C.2)}
\end{aligned}
$$

where

$$
\begin{aligned}
A &\equiv \text{Cross-sectional area of the tube}\\
p(x,t) &\equiv \text{Variation in sound pressure in the tube at position } x \text{ and time } t\\
u(x,t) &\equiv \text{Variation in volume velocity flow at position } x \text{ and time } t\\
\rho &\equiv \text{Density of moist air at } 37^oC\\
&= 1.14 * 10^{-3} gm/cm^3\\
c &\equiv \text{Speed of sound in moist air at } 37^oC\\
&= 3.5 * 10^4 cm/sec
\end{aligned}
$$

The solution to these equations for a sound wave travelling in the k^{th} tube with cross-sectional area A_k has the form:

$$
\begin{aligned}
p_k(x,t) &= p_k^+(t - x/c) + p_k^-(t + x/c) && \text{(C.3)}\\
u_k(x,t) &= u_k^+(t - x/c) - u_k^-(t + x/c) &&\\
&= \frac{A_k}{\rho c} \cdot \left\{ p_k^+(t - x/c) - p_k^-(t + x/c) \right\} && \text{(C.4)}
\end{aligned}
$$

where p^+ and p^- denote the magnitudes of the plane progressive waves travelling in the positive and negative directions respectively, and the volume velocity flow is related to the sound pressure magnitude by the characteristic impedance of the tube, which for lossless tubes is real-valued, and given by:

$$
\begin{aligned}
Z_{0_k} &= \frac{p_k(x,t)}{u_k(x,t)}\\
&= \frac{\rho c}{A_k} && \text{(C.5)}
\end{aligned}
$$

Travelling sound waves are hypothesised in the k^{th} and $(k-1)^{th}$ tubes, each of which has length l as shown in Figure C.1. The boundary conditions:

$$
\begin{aligned}
p_k(0,t) &= p_{k-1}(0,t) && \text{(C.6)}\\
u_k(0,t) &= u_{k-1}(0,t) && \text{(C.7)}
\end{aligned}
$$

are then enforced, to obtain the solutions:

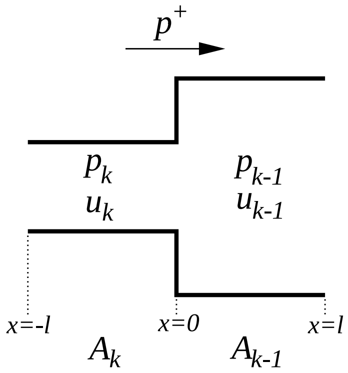

Figure C.1: Propagation of an acoustic wave through cylindrical tubes.

$$p_k^-(t) = r_{k-1} p_k^+(t) + [1 - r_{k-1}] \,.\, p_{k-1}^-(t) \tag{C.8}$$

$$p_{k-1}^+(t) = [1 - r_{k-1}] \,.\, p_k^+(t) + r_{k-1} p_{k-1}^-(t) \tag{C.9}$$

where r_{k-1} is the real-valued reflection coefficient for a pressure wave travelling in the positive direction in A_k encountering the boundary between A_k and A_{k-1} at $x = 0$. Its value is given by:

$$\begin{aligned} r_{k-1} &= \frac{A_k - A_{k-1}}{A_k + A_{k-1}} \\ &= \frac{Z_{0_{k-1}} - Z_{0_k}}{Z_{0_{k-1}} + Z_{0_k}} \end{aligned} \tag{C.10}$$

Now consider the case where the k^{th} tube represents the opening at the lips, nostrils or glottis, and hence is coupled not to an adjacent lossless tube, but to a load with complex impedance Z_L. In this case the *complex* reflection coefficient Γ_k for a wave travelling in the positive direction in the k^{th} tube at the boundary to the load Z_L will be given by:

$$\Gamma_k = \frac{Z_L - Z_{0_k}}{Z_L + Z_{0_k}} \tag{C.11}$$

Expressions can then be obtained for both Z_{in_k}, the impedance seen looking into tube k from tube $(k+1)$, and G_k, the gain across tube k in terms of Γ_k. Initially, observe that if a positive-moving pressure wave $p_k^+(t + l/c)$ enters the left-hand end of tube k and is reflected at the boundary between tube k and the load Z_L, the reflected pressure and volume flow waves at $x = -l$ will be given by:

$$p_k^-(t - \tau) = \Gamma_k \, p_k^+(t - \tau) \tag{C.12}$$

$$u_k^-(t - \tau) = -\Gamma_k \, u_k^+(t - \tau) \tag{C.13}$$

where $\tau = l/c$ is the time taken by the wave to propagate through the tube segment in one direction. Then:

$$Z_{in_k} = \frac{p_k^+(t+\tau) + \Gamma_k\, p_k^+(t-\tau)}{u_k^+(t+\tau) - \Gamma_k\, u_k^+(t-\tau)} \tag{C.14}$$

and using Equation C.5 and taking Z-transforms the following expression is obtained:

$$Z_{in_k} = Z_{0_k} \cdot \left(\frac{1 + z^{-1}\Gamma_k}{1 - z^{-1}\Gamma_k} \right) \tag{C.15}$$

Equations C.10 and C.15 can then be used to define the complex reflection coefficient Γ_{k+1} at the left-hand boundary of tube k for positive-travelling pressure waves in tube $(k+1)$, in terms of Γ_k and r_k as follows:

$$\begin{aligned} \Gamma_{k+1} &= \frac{Z_{in_k} - Z_{0_{k+1}}}{Z_{in_k} + Z_{0_{k+1}}} \\ &= \frac{r_k + z^{-1}\Gamma_k}{1 + r_k z^{-1}\Gamma_k} \end{aligned} \tag{C.16}$$

Similarly, the gain G_k across tube k can now be computed as:

$$\begin{aligned} G_k &= \frac{p_k^+(t) + p_k^-(t)}{p_k^+(t+\tau) + p_k^-(t-\tau)} \\ &= z^{-\frac{1}{2}} \cdot \left(\frac{1 + \Gamma_k}{1 + z^{-1}\Gamma_k} \right) \end{aligned} \tag{C.17}$$

These results greatly simplify the computation of the transfer function for a vocal tract model comprising a concatenation of uniform length tubes, as described in Section C.3. By repeated application of Equations C.15, C.16 and C.17, a vocal tract model consisting of several distinct branches can be reduced to a set of equivalent input impedances and gains, from which the required transfer functions can readily be computed.

C.3 Vocal Tract Transfer Functions

For the purposes of computing the transfer function for a wave travelling from the glottis to the lips and/or nostrils, the vocal tract is divided into pharyngeal, oral and nasal branches as shown in Figure C.2, after Rubin [145].

In this figure, nasals and nasalised vowels are represented by a three-branch vocal tract, where the three branches join at the velum, the physical coupling point of the nasal cavity to the vocal tract. Non-nasalised vowels and voiced and unvoiced non-nasalised fricatives are modelled by a two-branch tract, which is split at the point of maximum constriction in the case of fricatives, and at an arbitrary location for non-nasalised vowels.

The excitation at the glottis is represented by a volume velocity source u_{glot} with source impedance $Z_{S_{glot}}$. The source of frication is assumed to be situated at the point where the air flow emerges from the point of maximum constriction in the vocal tract, and is represented as a pressure source p_{fric}.

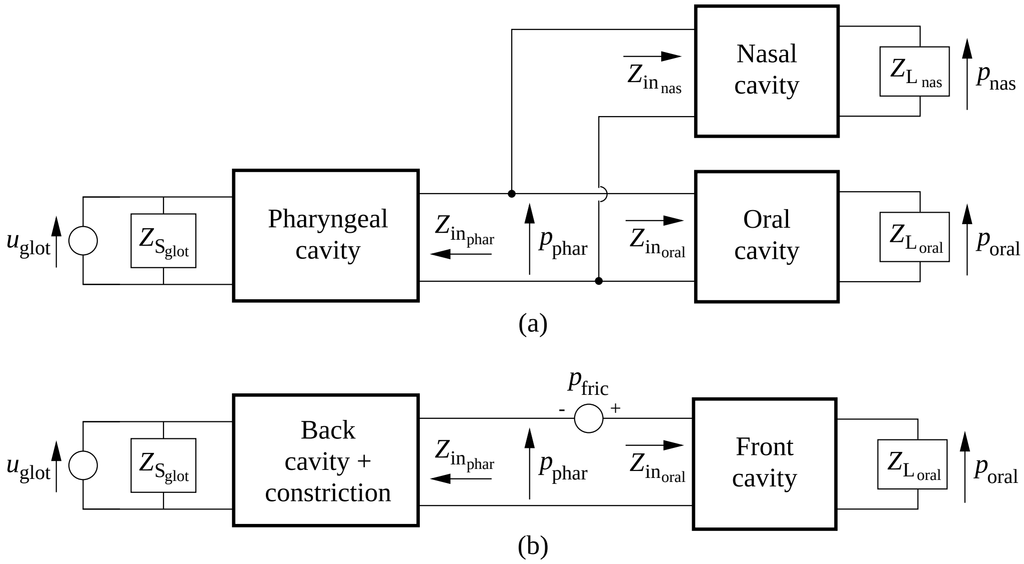

Figure C.2: Schematic diagram of vocal tract model for (a) nasalised sounds (b) non-nasalised sounds.

The output impedances at the lips and nostrils are denoted $Z_{L_{oral}}$ and $Z_{L_{nas}}$ respectively, while the input impedances of the oral, pharyngeal and nasal branches are denoted $Z_{in_{oral}}$, $Z_{in_{phar}}$ and $Z_{in_{nas}}$ respectively. In the case of fricatives and non-nasalised vowels, the first two impedances actually correspond to the input impedances of the "front" and "back" cavities respectively, but the same notation is retained for simplicity.

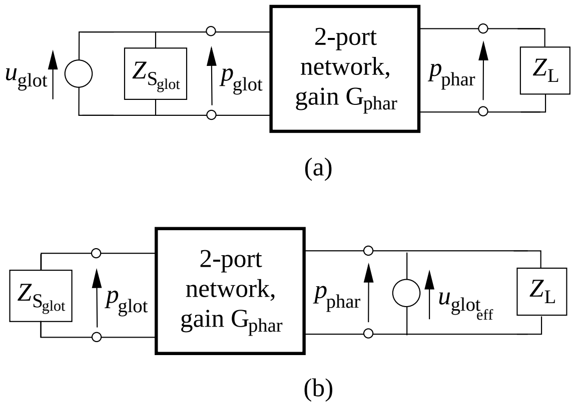

Figure C.3: Transformation of pharyngeal branch (a) to its Norton equivalent circuit (b).

In both models the analysis is greatly simplified by replacing the branch nearest the glottis with its Norton equivalent, as shown in Figure C.3. In this figure the pharyngeal branch is represented as a two-port network with gain:

$$G_{phar} = \left. \frac{p_{glot}}{p_{phar}} \right|_{u_{glot}=0} \tag{C.18}$$

and the oral and nasal branches have been replaced by an equivalent impedance Z_L for simplicity. The source u_{glot} can then be replaced with the effective source $u_{glot_{eff}}$ where:

$$u_{glot_{eff}} = u_{glot} \,.\, G_{phar} \tag{C.19}$$

The transfer function H_{voice} of the three-branch network is then given by[1]:

$$\begin{aligned} H_{voice} &= \frac{p_{oral} + p_{nas}}{u_{glot}} \\ &= \frac{G_{phar}(G_{oral} + G_{nas})}{1/Z_{in_{oral}} + 1/Z_{in_{nas}} + 1/Z_{in_{phar}}} \end{aligned} \tag{C.20}$$

where G_{oral} and G_{nas} are the gains in the oral and nasal sections respectively, given by:

$$G_{oral} = \left.\frac{p_{oral}}{p_{phar}}\right|_{u_{glot}=0} \tag{C.21}$$

$$G_{nas} = \left.\frac{p_{nas}}{p_{phar}}\right|_{u_{glot}=0} \tag{C.22}$$

Similarly, the transfer function H_{fric} for the fricative component is given by:

$$\begin{aligned} H_{fric} &= \frac{p_{oral}}{p_{fric}} \\ &= G_{oral} \cdot \frac{Z_{in_{oral}}}{Z_{in_{oral}} + Z_{in_{phar}}} \end{aligned} \tag{C.23}$$

C.4 Cylindrical Tube Model

The branches of the vocal tract model depicted in Figure C.2 are composed of concatenated co-axial cylindrical tubes. While the cross-sectional areas of the tubes are variable, they are chosen to be of uniform length, in order to simplify the analysis.

C.4.1 Tube Length

The value taken for the fixed tube length is significant, since it determines the maximum sampling rate—and hence the bandwidth—of the synthesised signal. Retaining the notation of Section C.2, if τ is the time taken for the sound wave to propagate through a tube segment, then when an impulse is applied at the input of an N-segment model, the time taken for the signal to reach the output will be $N\tau$, and successive outputs due to reflections will reach the output at multiples of 2τ, which therefore represents the maximum sampling period, T, of the signal.

Since the frequency response of such a lossless tube model is periodic [132], the bandwidth of the output signal will then be limited to:

[1] If the velum is closed, the effects of the nasal tract are neglected by setting $Z_{in_{nas}} = \infty$ and $G_{nas} = 0$ in Equation C.20.

$$|F| < \frac{1}{2T} = \frac{1}{4\tau} = \frac{c}{4l} \tag{C.24}$$

and since throughout this dissertation speech sampled at $16kHz$—corresponding to a bandwidth of $8kHz$—is used, this yields a segment length of:

$$l = 1.09375cm \tag{C.25}$$

C.4.2 Tract Shapes

The nasal branch of the vocal tract is assumed to be of fixed shape, and $\approx 12.5cm$ in length [43]. The nasal model used here therefore comprises 11 tube segments of fixed cross-sectional area, as well as a 12^{th} segment of variable area representing the opening at the velum where the nasal, oral and pharyngeal branches meet.

The shape of the nasal cavity was determined by interpolating and re-sampling Fant's data, and is illustrated in Figure C.4.

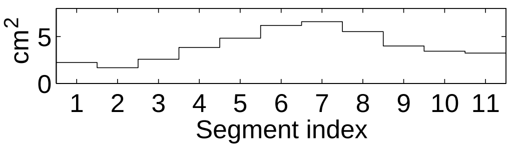

Figure C.4: Area function for fixed nasal tract, from nostrils (1) to velum (11).

By contrast, both the length *and* the shape of the tract between the glottis and the lips vary with time during the production of speech. Variations in length are caused by protrusions of the lips and jaw, as well as by the vertical displacement of the vocal cords, and are difficult to model in a system using fixed-length tube sections. As a simple approximation, a quantised variation in length was therefore implemented, by selecting a variable number of fixed-length tube segments in the model. Specifically, both 15 and 16-section models are used, to give overall tract lengths of $\approx 16.4cm$ and $17.5cm$ respectively, chosen to approximate those of the human vocal tract.

The time-varying shape of this cavity is determined from a set of articulatory parameters as described in Section 6.3.1, and the nasal and pharyngeal-oral tracts are joined at the tube segment boundary which is the minimum number of segments from the glottis, while being at least $8cm$ distant from it.

C.5 Sources, Loads and Losses

The task of developing accurate models of the excitation sources and the excitation and load impedances in the human vocal tract is a complex one, and the implementation of practical vocal tract simulations such as that presented in this chapter relies on the application of a number of simplifying assumptions [43, 44, 132]:

1. The voiced excitation at the glottis is only quasi-periodic and consists of pulses of variable shape, but is approximated using a stable, periodic waveform.

2. The coupling between the glottis and the vocal tract is dependent upon the size of the glottal opening, and hence gives rise to a time-varying source impedance at the glottis, and non-linear system behaviour. The weak nature of this coupling leads to an approximation in which the excitation source and vocal tract model are decoupled and linearised.

3. The noise source associated with frication and stop bursts is spatially distributed, has a pressure which is dependent on the constriction area, has a non-white spectrum, and gives rise to a non-linear turbulent flow. This source is approximated by a white noise source placed at the point at which the air flow emerges from a constriction, and a constant noise source amplitude is used.

4. The opening at the lips resembles an orifice in a spherical baffle of radius $\approx 9cm$, whose radiation impedance is difficult to model. It is therefore approximated as an opening in an infinite planar baffle, as is the opening at the nostrils.

While these assumptions permit the ready implementation of boundary conditions for the vocal tract model, they nevertheless result in a significant degradation in the quality of the resulting synthetic acoustic waveforms.

C.5.1 Voicing and Frication Sources

Flanagan's small-signal equivalent source is adopted as the glottal model [44], which combines viscous and kinetic resistive elements to yield a typical source resistance of:

$$R_{glot} = 100 \text{ acoustic Ohms} \tag{C.26}$$

The glottal waveform for voiced sounds is then explicitly defined in the time domain using Rosenberg's approximation [143]:

$$g(n) = \begin{cases} \frac{1}{2}\left[1 - \cos(\frac{\pi n}{N_1})\right] & 0 \leq n \leq N_1 \\ \cos(\frac{\pi(n-N_1)}{2N_2}) & N_1 \leq n \leq N_1 + N_2 \\ 0 & N_1 + N_2 \leq n \leq N_{max} \end{cases} \tag{C.27}$$

where the waveform is rising for $0 \leq n \leq N_1$, falling for $N_1 \leq n \leq N_1 + N_2$ and zero for $N_1 + N_2 \leq n \leq N_{max}$, where $N_{max} = 1/(F_0.T)$ is the number of sample points in a pitch period, and the sample period of the resulting waveform is as described in Section C.4.1:

$$T = 2\tau = \frac{1}{16kHz} = 62.5\mu sec \tag{C.28}$$

If it is assumed that the pulse has a fundamental frequency $F_0 = 120Hz$, is non-zero for 0.75 of a pitch period, and is rising for 0.7 of this time, then the following example parameter values are obtained:

$$N_1 = 70 \text{ samples} \tag{C.29}$$

$$N_2 = 30 \text{ samples} \tag{C.30}$$

$$N_{max} = 134 \text{ samples} \tag{C.31}$$

Figure C.5 shows the resulting pulse shape, 21 pitch periods of which were used as the input signal at the glottis for all voiced sounds. Two pitch periods are shown in the figure, corresponding to 268 sample points.

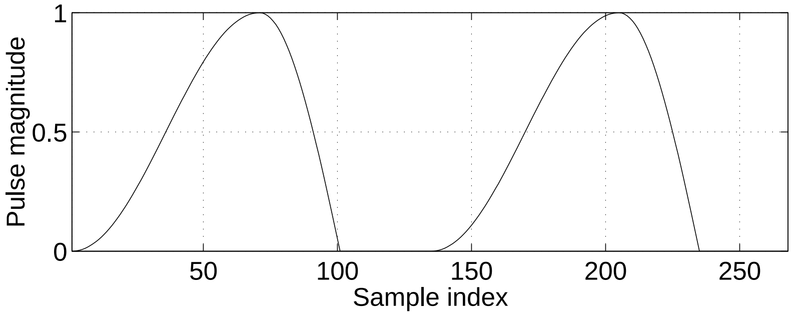

Figure C.5: Glottal pulse shape for voiced sounds.

In the case of fricatives and stop bursts, a white noise source is placed just anterior to the point of maximum constriction as described in Section C.3, and the source impedance used is the characteristic impedance of the tube segment representing the constriction. When a fricative is also voiced, the noise source is not continuously applied, but is modulated by the glottal waveform so that frication occurs only when the glottal pulse is non-zero.

C.5.2 Radiation Impedances

Approximating the openings at the lips and nostrils as circular holes in an infinite plane baffle, the load impedance $Z_L(s)$ can be expressed as a resistance R_L in parallel with an inductance L_L where [132]:

$$R_L = \frac{128}{9\pi^2} \tag{C.32}$$

$$L_L = \frac{8r}{3\pi c} \tag{C.33}$$

where r is the radius of the opening and $c = 3.5 * 10^4 \, cm/sec$ as defined in Section C.2. The bilinear transform is then applied to obtain:

$$\begin{aligned} Z_L(z) &= Z_L(s)\big|_{s=\frac{2}{T}\left(\frac{1-z^{-1}}{1+z^{-1}}\right)} \\ &= \frac{1-z^{-1}}{\frac{T}{L} + (\frac{1}{R} - \frac{T}{2L})(1-z^{-1})} \\ &\approx \frac{1-z^{-1}}{\left(\frac{2.06*10^4 T}{r} + 0.694\right) + \left(\frac{2.06*10^4 T}{r} - 0.694\right) z^{-1}} \end{aligned} \tag{C.34}$$

where the sampling period $T = 62.5\mu sec$ as before[2].

[2]Rubin uses the following expression in place of Equation C.34 [145]:

$$Z_L(z) \approx \frac{1-z^{-1}}{\left(\frac{4.12*10^4 T}{r} + 0.694\right) - 0.694 z^{-1}}$$

C.5.3 Models of Vocal Tract Losses

Following Rubin, losses in the oral and pharyngeal sections of the tract are approximated by introducing an attenuation factor $\alpha_k^{1/2}$ for a pressure wave travelling the length of the k^{th} tube segment with cross-sectional area A_k, given by:

$$\alpha_k^{1/2} = 1 - \frac{0.007}{\sqrt{A_k}} \tag{C.35}$$

so that Equations C.15, C.16 and C.17 become:

$$Z_{in_k} = Z_{0_k} \frac{1 + \alpha_k z^{-1} \Gamma_k}{1 - \alpha_k z^{-1} \Gamma_k} \tag{C.36}$$

$$\Gamma_{k+1} = \frac{r_k + \alpha_k z^{-1} \Gamma_k}{1 + r_k \alpha_k z^{-1} \Gamma_k} \tag{C.37}$$

$$G_k = \alpha_k^{-1/2} z^{-1/2} \frac{1 + \Gamma_k}{1 + \alpha_k z^{-1} \Gamma_k} \tag{C.38}$$

respectively. In the nasal tract, a constant attenuation factor of 0.99 is used, and additional losses are modelled by inserting a lumped component impedance in the middle of the tract, consisting of a series resistance and inductance:

$$\begin{aligned} Z_{mid} &= R_{mid} + sL_{mid} \\ &= 10 + 10.6 * 10^{-3} s \end{aligned} \tag{C.39}$$

where these values are chosen to yield a -3dB frequency of $150Hz$. Finally, a coupling resistance of 50 acoustic Ohms is inserted at the point where the nasal branch joins the oral and pharyngeal branches.

C.6 Vocal Tract Model Performance

Two example vocal tract frequency response curves and their associated time-domain waveforms are shown in Figure C.6. The area functions used were taken from Fant's data for the vowel `/aa/` and the voiced fricative `/z/` respectively [43], and were used as input to the Kelly-Lochbaum vocal tract model. In each case a sampling frequency of $16kHz$ was used, and three pitch periods using the glottal pulse waveform depicted in Figure C.5 are shown. During synthesis of the acoustic signal for `/z/`, the noise source at the point of maximum constriction was modulated by this glottal pulse as described in Section C.5.1. In each case plausible transfer function shapes are obtained, and intelligible acoustic waveforms are produced.

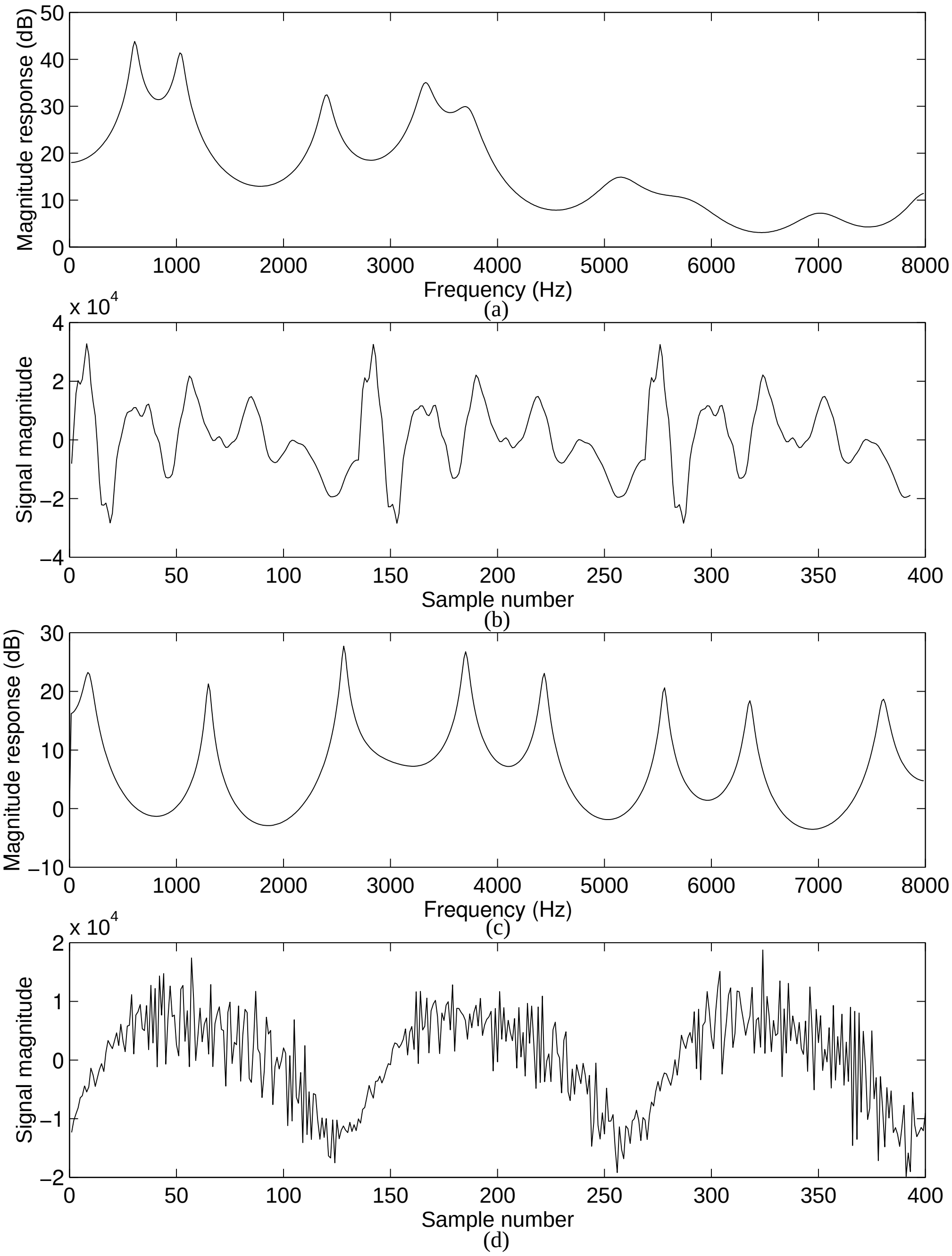

Figure C.6: Example frequency responses and time-domain waveforms synthesised from Fant's area function data using the explicit vocal tract model: (a) frequency response for `/aa/` (b) time-domain waveform for `/aa/` (c) frequency response for `/z/` (d) time-domain waveform for `/z/`.

Bibliography

[1] M. Akagi. "Evaluation of a spectrum target prediction model in speech perception". *Journal of the Acoustical Society of America*, 87(2):858–865, February 1990.

[2] M. Akagi and Y. Tohkura. "Spectrum target prediction model and its application to speech recognition". *Computer Speech and Language*, 4(4):325–344, 1990.

[3] B. S. Atal. "Neural networks for estimating articulatory positions from speech". *Journal of the Acoustical Society of America*, 86:S67, 1989.

[4] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey. "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique". *Journal of the Acoustical Society of America*, 63(5):1535–1555, May 1978.

[5] S. Austin, G. Zavaliagkos, J. Makhoul, and R. Schwartz. "Improving state-of-the-art continuous speech recognition systems using the N-best paradigm with neural networks". In *Proceedings of the DARPA Workshop on Speech and Natural Language*, pages 180–184. Morgan Kaufmann, 1992.

[6] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye. "Analysis of vocal tract shape and dimension using magnetic resonance imaging: Vowels". *Journal of the Acoustical Society of America*, 90(2):799–828, August 1991.

[7] G. Bailly. "Sensory-motor control of speech movements". In *ECSA 4th Tutorial and Workshop on Speech Production Modelling*, pages 145–154, Autrans, France, May 1996. ECSA.

[8] R. Bakis. "Coarticulation Modelling with Continuous-State HMMs". In *IEEE Speech Recognition Workshop*, Arden House, December 1991. IEEE.

[9] Yaakov Bar-Shalom and Xiao-Rong Li. *Estimation and Tracking.* Artech House, Boston, USA, 1993.

[10] J. Bertanstam, J. Beskow, M. Blomberg, R. Carlson, K. Elenius, B. Granström, J. Gustafson, S. Hunnicutt, J. Högberg, R. Lindell, L. Neovius, L. Nord, A. de Serpa-Leitao, and N. Ström. "The Waxholm system - a progress report". In *ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, Denmark, 1995.

[11] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[12] C. S. Blackburn and S. J. Young. “A novel self-organising speech production system using pseudo-articulators”. *International Congress of Phonetic Sciences*, 2:238–241, 1995.

[13] C. S. Blackburn and S. J. Young. “Learning new articulator trajectories for a speech production model using artificial neural networks”. *IEEE International Conference on Neural Networks*, 4:2046–2051, 1995.

[14] C. S. Blackburn and S. J. Young. “Towards improved speech recognition using a speech production model”. In *European Conference on Speech Communication and Technology*, volume 3, pages 1623–1626, 1995.

[15] C. S. Blackburn and S. J. Young. “A self-learning speech synthesis system”. In *ECSA 4th Tutorial and Workshop on Speech Production Modelling*, pages 225–228, Autrans, France, May 1996. ECSA.

[16] C. S. Blackburn and S. J. Young. “Pseudo-articulatory speech synthesis for recognition using automatic feature extraction from X-ray data”. In *Proceedings of the International Conference on Speech and Language Processing*, volume 2, pages 969–972, 1996.

[17] M. Blomberg. “Synthetic phoneme prototypes and dynamic voice source adaptation in speech recognition”. STL-QPSR 4, Dept. of Speech Communication and Music Acoustics, KTH, Stockholm, 1993.

[18] M. Blomberg. “A common phone model representation for speech recognition and synthesis”. In *Proceedings of the International Conference on Speech and Language Processing*, volume 4, pages 1875–1878, 1994.

[19] L. J. Boë, J. L. Schwartz, R. Laboissière, and N. Vallée. “Integrating articulatory-acoustic constraints in the prediction of sound structures”. In *1st ESCA Tutorial and Research Workshop on Speech Production Modeling*, pages 163–166, May 1996.

[20] H. Bourlard. “Towards increasing speech recognition error rates”. In *European Conference on Speech Communication and Technology*, volume 2, pages 883–894, 1995.

[21] H. Bourlard and N. Morgan. “A Continuous Speech Recognition System Embedding MLP into HMM”. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems II*, pages 186–193. Morgan Kaufman, San Mateo, CA, 1990.

[22] H. Bourlard, N. Morgan, and C. J. Wellekens. “Statistical Inference in Multilayer Perceptrons and Hidden Markov Models with Applications in Continuous Speech Recognition”. In F. Fogelman-Soulie and J. Hérault, editors, *Neurocomputing: Algorithms, Architectures and Applications*, NATO ASI Series, pages 215–226. Springer-Verlag, 1990.

[23] H. A. Bourlard and N. Morgan. *Connectionist Speech Recognition.* Kluwer Academic Publishers, USA, 1994.

[24] S. M. Bozic. *Digital and Kalman filtering*. Edward Arnold, 41 Bedfor Square, London, 1979.

[25] H. Brann and M. Riedmiller. "Rprop: A fast and robust backpropagation learning strategy". In *ACNN*, 1993.

[26] J. S. Bridle. "Alpha-Nets: A Recurrent "Neural" Network Architecture with a Hidden Markov Model Interpretation". *Speech Communication*, 9(1):83–92, 1990.

[27] J. S. Bridle. "Probabilistic Interpretation of Feedforward Classification Network Ouputs, with Relationships to Statistical Pattern Recognition". In F. Fogelman-Soulie and J. Hérault, editors, *Neuro-Computing: Algorithms, Architectures and Applications*, NATO ASI Series. Springer-Verlag, 1990.

[28] J. S. Bridle and M. P. Ralls. "An approach to speech recognition using synthesis-by-rule". In F. Fallside and W. A. Woods, editors, *Computer Speech Processing*, chapter 10, pages 277–292. Prentice-Hall International, UK, 1985.

[29] C. Browman and L. Goldstein. "Articulatory phonology: an overview". *Phonetica*, 49:155–180, 1992.

[30] R. G. Brown and P. Y. C. Hwang. *Introduction to random signals and applied Kalman filtering*. John Wiley & Sons, Inc., USA, 2nd edition, 1992.

[31] L. Candille and H. Meloni. "Automatic speech recognition using speech production models". *International Congress of Phonetic Sciences*, 4:256–259, 1995.

[32] S. Chatterjee and B. Price. *Regression Analysis by Example*. John Wiley & Sons, U.S.A., 1977.

[33] C. H. Coker. "A Model of Articulatory Dynamics and Control". *Proceedings of the IEEE*, 64(4):452–460, April 1976.

[34] R. Cole, L. Hirschman, L. Atlas, M. Beckman, A. Bierman, M. Bush, M. Clements, J. Cohen, O. Garcia, B. Hanson, H. Hermansky, S. Levinson, K. McKeown, N. Morgan, D. G. Novick, M. Ostendorf, S. Oviatt, P. Price, H. Silverman, J. Spitz, A. Waibel, C. Weinstein, S. Zahorian, and V. Zue. "The challenge of spoken language systems: research directions for the nineties". *IEEE Transactions on Speech and Audio Processing*, 3(1):1–20, January 1995.

[35] S. B. Davis and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, August 1980.

[36] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society*, 39(B1):1–38, 1977.

[37] L. Deng, G. Ramsay, and D. Sun. "Production models as a structural basis for automatic speech recognition". In *1st ESCA Tutorial and Research Workshop on Speech Production Modeling*, pages 69–80, May 1996.

[38] L. Deng and D. X. Sun. "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features". *Journal of the Acoustical Society of America*, 95(5):2702–2719, May 1994.

[39] V. Digalakis, M. Ostendorf, and J. R. Rohlicek. "Improvements in the stochastic segment model for phoneme recognition". In *Proceedings of the DARPA Workshop on Speech and Natural Language*, pages 332–338, 1989.

[40] V. Digalakis, J. R. Rohlicek, and M. Ostendorf. "A Dynamical System Approach to Continuous Speech Recognition". In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 289–292, May 1991.

[41] R. E. Donovan. *"Trainable speech synthesis"*. PhD thesis, University of Cambridge, 1995.

[42] R. E. Donovan and P. C. Woodland. "Improvements in an HMM-based speech synthesiser". In *European Conference on Speech Communication and Technology*, volume 1, pages 573–576, 1995.

[43] G. Fant. *Acoustic theory of speech production.* Mouton & Co., The Hague, 1970. First published 1960.

[44] J. L. Flanagan. *Speech analysis synthesis and perception.* Springer-Verlag, 2nd edition, 1972.

[45] J. L. Flanagan, K. Ishizaka, and K. L. Shipley. "Synthesis of speech from a dynamic model of the vocal cords and vocal tract". *Bell Systems Technical Journal*, 54(3):485–506, March 1975.

[46] A. Fourcin. "Prospects and Problems in Spoken Language Engineering". In *3rd European Conference on Speech Communication and Technology Plenary Session*, pages 17–20, 1993.

[47] O. Fujimura, H. Ishida, and S. Kiritani. "Computer-controlled dynamic cineradiography". *Annual Bulletin, Research Institute of Logopedics and Phoniatrics, University of Tokyo*, (2):6–10, 1968.

[48] S. Furui. "On the role of spectral transition for speech perception". *Journal of the Acoustical Society of America*, 80(4):1016–1025, October 1986.

[49] B. Gabioud. "Articulatory Models in Speech Synthesis". In E. Keller, editor, *Fundamentals of Speech Synthesis and Speech Recognition*, chapter 10, pages 215–230. John Wiley & Sons, West Sussex, England, 1994.

[50] T. Gay. "Articulatory movements in VCV sequences". *Journal of the Acoustical Society of America*, 62(1):183–193, July 1977.

[51] T. Gay. "Articulatory Units: segments or syllables". In A. Bell and J. B. Hooper, editors, *Syllables and Segments*, pages 121–132. North-Holland Publishing Company, the Netherlands, 1978.

[52] M. George, P. Jospa, and A. Soquet. "Articulatory trajectories generated by the control of the vocal tract by a neural network". In *Proceedings of the International Conference on Speech and Language Processing*, volume 2, pages 583–586, 1994.

[53] Z. Ghahramani. "Solving inverse problems using an EM approach to density estimation". In *Proceedings of the 1993 Connectionist Models Summer School*, Hillsdale, NJ, 1994. Erlbaum.

[54] W. D. Goldenthal and J. R. Glass. "Modelling spectral dynamics for vowel classification". In *European Conference on Speech Communication and Technology*, volume 1, pages 289–292, 1993.

[55] F. H. Guenther. "A modeling framework for speech motor development and kinematic articulator control". *International Congress of Phonetic Sciences*, 2:92–99, 1995.

[56] W. Hardcastle, B. Vaxelaire, F. Gibbon, P. Hoole, and N. Nguyen. "EMA/EPG study of lingual coarticulation in /kl/ clusters". In *1st ESCA Tutorial and Research Workshop on Speech Production Modeling*, pages 53–56, May 1996.

[57] W. J. Hardcastle. *Physiology of Speech Production.* Academic Press, 24-28 Oval Rd, London, 1976.

[58] K. S. Harris. "The Study of Articulatory Organization: Some Negative Progress". In M. Sawashima and F. S. Cooper, editors, *Dynamic Aspects of Speech Production*, pages 71–82. University of Tokyo Press, Japan, 1976.

[59] J. A. Hertz, R. G. Palmer, and A. S. Krogh. *Introduction to the Theory of Neural Computation.* Addison-Wesley, 350 Bridge Parkway, Redwood City, CA 94065, U.S.A, 1991.

[60] G. E. Hinton. "Connectionist learning procedures". Technical Report CMU-CS-87-115, Carnegie Mellon University, Pittsburgh, PA, 1987.

[61] M. Hiraike, S. Shimizu, T. Mizutani, and K. Hashimoto. "Estimation of the lateral shape of a tongue from speech". In *Proceedings of the International Conference on Speech and Language Processing*, volume 2, pages 591–594, 1994.

[62] M. Hirayama, E. Vatikiotis-Bateson, M. Kawato, and M. I. Jordan. "Forward dynamics modeling of speech motor control using physiological data". In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 191–198. Morgan Kaufman, USA, 1992.

[63] J. N. Holmes. *Speech Synthesis and Recognition.* Van Nostrand Reinhold (UK), Molly Millars Lane, Wokingham, Berkshire, England, 1988.

[64] M. Honda and T. Kaburagi. "A dynamical articulatory model using potential task representation". In *Proceedings of the International Conference on Speech and Language Processing*, volume 1, pages 179–182, 1994.

[65] P. Hoole. "Issues in the acquisition, processing, reduction and parameterization of articulographic data". *Forschungsberichte des Instituts für Phonetik and Sprachliche Kommunikation, Universität München*, 34:158–173, 1996.

[66] R. A. Houde. *"A study of tongue body motion during selected speech sounds"*. PhD thesis, University of Michigan, Ann Arbor, 1967.

[67] X. D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition*. Edinburgh Information Technology Series. Edinburgh University Press, Edinburgh, 1990.

[68] A. Hunt. "Utilising prosody to perform syntactic disambiguation". In *European Conference on Speech Communication and Technology*, volume 2, pages 1339–1342, 1993.

[69] K. Ishizaka and J. L. Flanagan. "Synthesis of voiced sounds from a two-mass model of the vocal chords". *Bell system technical journal*, 51:1233–1268, 1972.

[70] K. Iso. "Speech recognition using a dynamical model of speech production". In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 283–286, 1993.

[71] T. T. Jervis. *"Connectionist adaptive control"*. PhD thesis, University of Cambridge, December 1993.

[72] M. I. Jordan and R. A. Jacobs. "Hierarchical Mixtures of Experts and the EM Algorithm". *Neural Computation*, 6:181–214, 1994.

[73] M. I. Jordan and D. E. Rumelhart. "Forward models: Supervised learning with a distal teacher". *Cognitive Science*, 16:307–354, 1992.

[74] P. Jospa and A. Soquet. "The acoustic-articulatory mapping and the variational method". In *Proceedings of the International Conference on Speech and Language Processing*, volume 2, pages 595–598, 1994.

[75] T. Kaburagi and M. Honda. "A trajectory formation model of articulatory movements based on the motor tasks of phoneme-specific vocal tract shapes". In *Proceedings of the International Conference on Speech and Language Processing*, volume 2, pages 579–582, 1994.

[76] T. Kaburagi and M. Honda. "A study on modelling articulator movements based on the task-independent energy criterion". In *ECSA 4th Tutorial and Workshop on Speech Production Modelling*, pages 137–140, Autrans, France, May 1996. ECSA.

[77] A. Kannan, M. Ostendorf, and J. R. Rohlicek. "Weight estimation for N-best rescoring". In *Proceedings of the DARPA Workshop on Speech and Natural Language*, pages 455–456. Morgan Kaufmann, 1992.

[78] P. A. Keating. "Underspecification in phonetics". In *Coarticulation*, pages 30–50. UCLA phonetics laboratory, 1988.

[79] P. A. Keating. "The window model of coarticulation: articulatory evidence". In J. Kingston and M. E. Beckman, editors, *Papers in Laboratory Phonology I*, chapter 26, pages 451–470. Cambridge University Press, Cambridge, 1990.

[80] A. Kehagias. "Optimal Control for Training: The Missing Link Between Hidden Markov Models and Connectionist Networks". Technical report, Division of Applied Mathematics, Brown University, Providence, RI, 1989.

[81] J. L. Kelly Jr. and C. Lochbaum. "Speech synthesis". In *Speech Communication Seminar*, Stockholm, 1962.

[82] R. Kent. "The segmental organization of speech". In P. F. MacNeilage, editor, *The Production of Speech*, chapter 4, pages 57–90. Springer-Verlag, New York, 1983.

[83] D. Kershaw, A. J. Robinson, and M. Hochberg. "Context-dependent classes in a hybrid recurrent network-HMM speech recognition system". In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 750–756. MIT Press, 1996.

[84] D. H. Klatt. "Word Verification in a Speech Understanding System". In D. R. Reddy, editor, *Speech Recognition*, pages 321–341. Academic Press, New York, 1975.

[85] D. H. Klatt. "Review of text-to-speech conversion for English". *Journal of the Acoustical Society of America*, 82(3):737–793, September 1987.

[86] T. Kobayashi, M. Yagyu, and K. Shirai. "Application of neural networks to articulatory motion estimation". In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 489–492, Toronto, 1991.

[87] E. Kreyszig. *Introductory Mathematical Statistics.* John Wiley & Sons, USA, 1970.

[88] B. J. Kröger, G. Schröder, and C. Opgen-Rhein. "A gesture-based dynamic model describing articulatory movement data". *Journal of the Acoustical Society of America*, 98(4):1878–1889, October 1995.

[89] H. Kuwabara. "An approach to normalisation of coarticulation effects for vowels in connected speech". *Journal of the Acoustical Society of America*, 77(2):686–694, February 1985.

[90] R. Laboissière, D. J. Ostry, and P. Perrier. "A model of human jaw and hyoid motion and its implications for speech production". *International Congress of Phonetic Sciences*, 2:60–67, 1995.

[91] P. Ladefoged. "Articulatory parameters". *Language and Speech*, 23(1):25–30, 1980.

[92] C-H. Lee. "On the use of some robust modeling techniques for speech recognition". *Computer Speech and Language*, 3:35–52, 1989.

[93] K-F. Lee. *Automatic Speech Recognition.* Kluwer Academic Publishers, 1989.

[94] S. E. Levinson and C. E. Schmidt. "Adaptive computation of articulatory parameters from the speech signal". *Journal of the Acoustical Society of America*, 74(4):1145–1154, 1983.

[95] B. Lindblom. "Economy of speech gestures". In P. F. MacNeilage, editor, *The Production of Speech*, chapter 10, pages 217–246. Springer-Verlag, New York, 1983.

[96] B. E. F. Lindblom and J. E. F. Sundberg. "Acoustical consequences of lip, tongue, jaw, and larynx movement". *Journal of the Acoustical Society of America*, 50(4 Part 2):1166–1179, 1971.

[97] J. K. Local. "Phonological structure, parametric phonetic interpretation and natural-sounding synthesis". In E. Keller, editor, *Fundamentals of Speech Synthesis and Speech Recognition*, chapter 12, pages 253–270. John Wiley & Sons, West Sussex, England, 1994.

[98] J. K. Local. "Making sense of dynamic, non-segmental phonetics". *International Congress of Phonetic Sciences*, 3:2–9, 1995.

[99] D. MacKay. *Bayesian Methods for Adaptive Models.* PhD thesis, California Institute of Technology, Pasadena, California, 1991.

[100] P. F. Macneilage. "Motor control of serial ordering of speech". *Psychological Review*, 77:182–196, 1970.

[101] S. Maeda. "An articulatory model based on statistical analysis". *Journal of the Acoustical Society of America*, 65(Sup. 1):S22, 1979.

[102] S. Maeda. "A digital simulation method of the vocal tract system". *Speech Communication*, 1:199–229, 1982.

[103] S. Maeda. "Improved articulatory model". *Journal of the Acoustical Society of America*, 84(Sup. 1):S146, 1988.

[104] J. N. Marcus and V. W. Zue. "A variable duration acoustic segment HMM for hard-to-recognize words and phrases". In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 281–284, Toronto, 1991.

[105] P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Number 37 in Monographs on statistics and applied probability. Chapman and Hall, London, 1983.

[106] D. McFarlan, editor. *The Guiness Book of Records 1991.* Guinness Publishing, 1990.

[107] P. Mermelstein. "Articulatory model for the study of speech production". *Journal of the Acoustical Society of America*, 53(4):1070–1082, 1973.

[108] P. Meyer, R. Wilhelms, and H. W. Strube. "A quasiarticulatory speech synthesizer for German language running in real time". *Journal of the Acoustical Society of America*, 86(2):523–539, 1989.

[109] M. Minsky and S. Papert. *Perceptrons.* MIT Press, Cambridge, MA, 1969.

[110] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics.* McGraw-Hill, 3rd edition, 1974.

[111] R. K. Moore. "Whither a theory of speech pattern processing". In *European Conference on Speech Communication and Technology*, volume 1, pages 43–47, 1993.

[112] R. K. Moore. "Computational phonetics". *International Congress of Phonetic Sciences*, 4:68–71, 1995.

[113] D. P. Morgan and C. L. Scofield. *Neural Networks and Speech Processing.* Kluwer academic publishers, USA, 1991.

[114] T. Nakajima. "Identification of a dynamic articulatory model by acoustic analysis". In M. Sawashima and F. S. Cooper, editors, *Dynamic Aspects of Speech Production*, chapter 3, pages 251–275. University of Tokyo Press, Japan, 1976.

[115] T. Nakajima, H. Omura, K. Tanaka, and S. Ishizaki. "Estimation of vocal tract area functions by adaptive inverse filtering methods and identification of articulatory model". In *Speech Communication Seminar*, volume 1, pages 11–20, Stockholm, August 1974.

[116] L. T. Niles and H. F. Silverman. "Combining Hidden Markov Models and Neural Network Classifiers". In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 417–420, Albuquerque, NM, 1990. IEEE.

[117] J. D. O'Connor. *Phonetics.* Pelican, 1973.

[118] J. J. Odell. *"The use of context in large vocabulary speech recognition".* PhD thesis, University of Cambridge, 1995.

[119] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, and J. R. Rohlicek. "Integration of diverse recognition methodologies through reevaluation of N-best sentence hypotheses". In *Proceedings of the DARPA Workshop on Speech and Natural Language*, pages 83–87. Morgan Kaufmann, 1991.

[120] M. Ostendorf and S. Roukos. "A stochastic segment model for phoneme-based continuous speech recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(12):1857–1869, December 1989.

[121] K. K. Paliwal and P. V. S. Rao. "Synthesis-based recognition of continuous speech". *Journal of the Acoustical Society of America*, 71(4):1016–1024, 1982.

[122] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, and J. Zacks. "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data". *Journal of the Acoustical Society of America*, 92(2 Part 1):688–700, August 1992.

[123] S. Parthasarathy and C. H. Coker. "On automatic estimation of articulatory parameters in a text-to-speech system". *Computer Speech and Language*, 6:37–75, 1992.

[124] J. S. Perkell. *Physiology of speech production: results and implications of a quantitative cineradiographic study.* MIT Press, Cambridge, Mass., 1969.

[125] J. S. Perkell, M. H. Cohen, M. A. Suirsky, M. L. Matthies, I. Garabieta, and M. T. T. Jackson. "Electromagnetic midsaggital articulometer systems for transducing speech articulatory movements". *Journal of the Acoustical Society of America*, 92:3078–3096, 1992.

[126] J. S. Perkell and D. H. Klatt, editors. *Invariance and variability in speech processes.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1986.

[127] J. S. Perkell, M. Matthies, R. Wilhelms-Tricarico, H. Lane, and J. Wozniak. "Speech motor control: phonemic goals and the use of feedback". In *ECSA 4th Tutorial and Workshop on Speech Production Modelling*, pages 133–136, Autrans, France, May 1996. ECSA.

[128] P. Perrier and D. J. Ostry. "Dynamic modelling and control of speech articulators: application to vowel reduction". In E. Keller, editor, *Fundamentals of Speech Synthesis and Speech Recognition*, chapter 11, pages 231–252. John Wiley & Sons, West Sussex, England, 1994.

[129] T. Piske. "Phonological organization in early speech production". In *1st ESCA Tutorial and Research Workshop on Speech Production Modeling*, pages 159–162, May 1996.

[130] L. R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". *Proceedings of the IEEE*, 77(2):257–285, February 1989.

[131] L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson. "Considerations in dynamic time warping algorithms for discrete word recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(6):575–582, December 1978.

[132] L. R. Rabiner and R. W. Schafer, editors. *Digital Processing of Speech Signals.* Prentice-Hall, New Jersey, 1978.

[133] M. G. Rahim, C. C. Goodyear, W. B. Kleijn, J. Schroeter, and M. M. Sondhi. "On the use of neural networks in articulatory speech synthesis". *Journal of the Acoustical Society of America*, 93(2):1109–1121, February 1993.

[134] G. Ramsay and L. Deng. "A stochastic framework for articulatory speech recognition". *Journal of the Acoustical Society of America*, 95(5 Part 2):2873, May 1994. Abstract 2pSP19.

[135] G. Ramsay and L. Deng. "Articulatory sysnthesis using a stochastic target model of speech production". *International Congress of Phonetic Sciences*, 2:338–341, 1995.

[136] G. Ramsay and L. Deng. "Maximum-likelihood estimation for articulatory speech recognition using a stochastic target model". In *European Conference on Speech Communication and Technology*, volume 2, pages 1401–1404, 1995.

[137] M. Randolph and J. Schroeter. "Speech decoding using a finite-state model of articulatory behavior". In *IEEE Automatic Speech Recognition Workshop*, Snowbird, Utah, December 1993. IEEE.

[138] M. A. Randolph. "Speech analysis based on a model of articulatory behaviour". *Journal of the Acoustical Society of America*, 95(5 Pt. 2):2818, 1994.

[139] D. L. Rice. "Articulatory tracking of the acoustic speech signal". In *Speech Communication Seminar*, volume 1, pages 21–26, Stockholm, August 1974.

[140] M. Riedmiller and H. Braun. "A direct adaptive method for faster backpropagation learning: the RPROP algorithm". *IEEE International Conference on Neural Networks*, 1993.

[141] A. J. Robinson. "An application of recurrent nets to phone probability estimation". *IEEE Transactions on Neural Networks*, 5(3), 1994.

[142] R. C. Rose, J. Schroeter, and M. M. Sondhi. "An investigation of the potential role of speech production models in automatic speech recognition". In *Proceedings of the International Conference on Speech and Language Processing*, volume 2, pages 575–578, 1994.

[143] A. E. Rosenberg. "Effect of glottal pulse shape on the quality of natural vowels". *Journal of the Acoustical Society of America*, 49:583–590, 1971.

[144] S. Roucos, M. Ostendorf, H. Gish, and A. Derr. "Stochastic segment modelling using the estimate-maximize algorithm". In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 127–130, New York, 1988.

[145] P. Rubin, T. Baer, and P. Mermelstein. "An articulatory synthesizer for perceptual research". *Journal of the Acoustical Society of America*, 70(2):321–328, August 1981.

[146] D. Rumelhart and J. McClelland. *Parallel Distributed Processing: explorations in the microstructure of cognition; Vol. 1: Foundations; Vol. 2: Psychological and biological models*. MIT Press, Cambridge, MA, 1986.

[147] E. L. Saltzman, L. Goldstein, C. Browman, and P. Rubin. "Modeling speech production using dynamic gestural structures". *Journal of the Acoustical Society of America*, 84(Sup. 1):S146, 1988.

[148] E. L. Saltzman and K. G. Munhall. "A dynamical approach to gestural patterning in speech production". *Ecological Psychology*, 1:333–382, 1989.

[149] O. Schmidbauer. "Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations". *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1:616–619, 1989.

[150] J. Schroeter, P. Meyer, and S. Parthasarathy. "Evaluation of improved articulatory codebooks and codebook distance measures". In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 393–396, 1990.

[151] J. Schroeter and M. M. Sondhi. "Techniques for Estimating Vocal-Tract Shapes from the Speech Signal". *IEEE Transactions on Speech and Audio Processing*, 2(1):133–150, January 1994.

[152] R. Schwartz and S. Austin. "Efficient, high-performance algorithms for N-best search". In *Proceedings of the DARPA Workshop on Speech and Natural Language*, pages 6–11. Morgan Kaufmann, 1990.

[153] R. Schwatrz, S. Ausin, F. Kubala, J. Makhoul, L. Nguyen, P. Placeway, and G. Zavaliagkos. "New uses for the N-best sentence hypotheses within the BYBLOS speech recognition system". In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 1–4, March 1992.

[154] T. J. Sejnowski and C. Rosenberg. Nettalk: A parallel network that learns to read aloud. Technical Report JHU/EECS-86/01, Johns Hopkins University, Baltimore, 1986.

[155] K. Shirai and T. Kobayashi. "Estimating articulatory motion from the speech wave". *Speech Communication*, 5(2):159–170, June 1986.

[156] M. M. Sondhi and J. Schroeter. "A hybrid time-frequency domain articulatory speech synthesizer". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(7):955–967, July 1987.

[157] P. Steingrimsson, B. Markussen, O. Andersen, P. Dalsgaard, and W. Barry. "From acoustic signal to phonetic features: a dynamically constrained self-organising neural network". In *International Congress of Phonetic Sciences*, volume 4, pages 316–319, 1995.

[158] M. A. Stephens. *Journal of the Royal Statistical Society*, B 32:115–122, 1970.

[159] K. N. Stevens, C. A. Bickley, and D. R. Williams. "Control of a Klatt synthesizer by articulatory parameters". In *Proceedings of the International Conference on Speech and Language Processing*, volume 1, pages 183–186, 1994.

[160] M. Stone. "A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data". *Journal of the Acoustical Society of America*, 87(5):2207–2217, May 1990.

[161] W. Strange, J. J. Jenkins, and T. L. Johnson. "Dynamic specification of coarticulated vowels". *Journal of the Acoustical Society of America*, 74(3):695–705, September 1983.

[162] H. W. Strube. "Can the Area Function of the Human Vocal Tract be Determined from the Spech Wave?". In M. Sawashima and F. S. Cooper, editors, *Dynamic Aspects of Speech Production*, chapter 3, pages 233–248. University of Tokyo Press, Japan, 1976.

[163] M. Tatham. "Articulatory phonology, task dynamics and computational adequacy". In *ECSA 4th Tutorial and Workshop on Speech Production Modelling*, pages 141–144, Autrans, France, May 1996. ECSA.

[164] R. B. Thosar and P. V. S. Rao. "An approach towards a synthesis-based speech recognition system". *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages 194–196, 1976.

[165] V. Valtchev. *"Discriminative methods in HMM-based speech recognition"*. PhD thesis, University of Cambridge, 1995.

[166] E. Vatikiotis-Bateson, M. Tiede, Y. Wada, V. Gracco, and M. Kawato. "Phoneme extraction using via point estimation of real speech". In *Proceedings of the International Conference on Speech and Language Processing*, volume 2, pages 631–634, 1994.

[167] A. J. Viterbi. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". *IEEE Trans. Information Theory*, IT-3:260–269, 1967.

[168] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. "Considerations in dynamic time warping algorithms for discrete word recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(3):328–339, March 1989.

[169] H. Wakita. "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms". *IEEE Transactions on Audio and Electroacoustics*, AU-21(5):417–427, October 1973.

[170] S. Werner and E. Keller. "Prosodic Aspects of Speech". In E. Keller, editor, *Fundamentals of Speech Synthesis and Speech Recognition*, chapter 2, pages 23–40. John Wiley & Sons, England, 1994.

[171] J. R. Westbury. "X-Ray microbeam speech production database user's handbook". Personal communication, University of Wisconsin, 1994.

[172] R. Wilhelms, P. Meyer, and H. W. Strube. "Estimation of articulatory trajectories by Kalman filtering". In I. T. Young, J. Biemond, R. P. W. Duin, and J. J. Gerbrands, editors, *Signal Processing III: Theories and Applications*, pages 477–480. Elsevier Science Publishers B.V. (North Holland), Amsterdam, 1986.

[173] R. Wilhelms-Tricarico. "Physiological modeling of speech production: methods for modeling soft-tissue articulators". *Journal of the Acoustical Society of America*, 97(5 Part 1):3085–3098, May 1995.

[174] P. C. Woodland, C. J. Leggetter, J. J. Odell, V. Valtchev, and S. J. Young. "The 1994 HTK large vocabulary speech recognition system". In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 73–76, May 1995.

[175] C. Yang and H. Kasuya. "Accurate measurements of vocal tract shapes from magnetic resonance images of child, female and male subjects". In *Proceedings of the International Conference on Speech and Language Processing*, volume 2, pages 623–626, 1994.

[176] S. J. Young, J. Jansen, J. J. Odell, D. Ollason, and P. C. Woodland. *The HTK Book*. Entropic Cambridge Research Laboratory and Univeristy of Cambridge, Cambridge, England, 1995. Version 2.0.

[177] S. J. Young, J. J. Odell, and P. C. Woodland. "Tree-based state tying for high accuracy acoustic modelling". In *ARPA Workshop on Human Language Technology*, pages 307–312. Morgan Kaufmann, March 1994.

[178] J. Zacks and T. R. Thomas. "A new neural network for articulatory speech recognition and its application to vowel identification". *Computer Speech and Language*, 8:189–209, 1994.

[179] I. Zlokarnik. "Experiments with an articulatory speech recognizer". In *European Conference on Speech Communication and Technology*, pages 2215–2218, 1993.

[180] V. Zue, J. Glass, D. Goodine, M. Phillips, and S. Seneff. "The SUMMIT speech recognition system: phonological modelling and lexical analysis". In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 49–52, 1990.