TOWARDS IMPROVED SPEECH RECOGNITION USING A SPEECH PRODUCTION MODEL

C. S. Blackburn　　*S. J. Young*

Cambridge University Engineering Department
Trumpington St, Cambridge
CB2 1PZ England

email: csb@eng.cam.ac.uk

ABSTRACT

Considerable improvement in the performance of continuous speech recognition systems, particularly those based on Hidden Markov Models (HMMs), has been shown in recent years. Nevertheless a number of unsolved problems remain which limit this progress, including the successful modelling of co-articulation and the identification of out of vocabulary utterances. One possible solution is to re-synthesise speech from the N-best time-aligned phonemic transcriptions produced by an HMM, and re-score this list based on a spectral comparison between the original and re-synthesised speech frames. In this paper a novel speech production model (SPM) suitable for use in such a system is introduced, and preliminary re-scoring results are presented.

1. INTRODUCTION

The application of speech production models to the task of automatic speech recognition is a relatively new area of research which has attracted increasing interest over the past few years [10]. The basic operation of such a combined system is illustrated in figure 1.

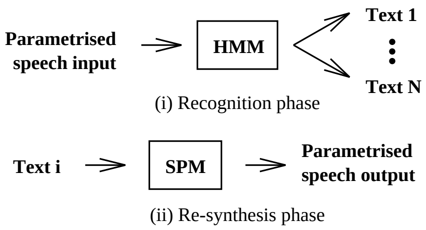

Figure 1. System overview.

A conventional speech recogniser such as an HMM is used to provide a list of hypothesised speech transcriptions, along with a time-alignment at the phonemic level. Each of these time-aligned phoneme strings is then used as input to the speech production model and the speech is re-synthesised. By comparing the synthetic speech corresponding to each of the N-best phonemic transcriptions with the original speech, it is possible to re-score the transcriptions, thus potentially offering improved recognition results.

In order to successfully re-order the HMM output transcriptions, the combined system must make use of information which the HMM alone does not. The key advantage of using a production model in this context is that it allows an explicit representation of co-articulation to be used. In statistically based systems such as HMMs, variability in the speech signal is typically accounted for by the inclusion of multiple models for each phoneme according to its phonetic context. Thus instead of maintaining a single model for each phoneme, either biphone or triphone models are used, in which a phoneme's left or right context (biphone) or both (triphone) are incorporated into the model. The success of this approach is usually limited by the very large amount of training data that would be required in order to adequately model each phoneme in all of its contexts.

By using an articulatory speech production model it is possible to provide an explicit model of co-articulation as shown in figure 2. If a suitable set of articulator targets and co-articulation parameters can be extracted for each of the phonemes, plausible articulator trajectories can be generated for any phoneme string. Thus whereas the majority of speech synthesisers are currently rule-based, this requirement has generated renewed interest in articulatory speech synthesis [11, 13, 9].

The performance of many of these systems has however, been limited by an inability to model the excitation sources and propagation characteristics of the human vocal tract sufficiently accurately. An alternative approach, and that which is pursued here, is to relax the constraint of exactly mimicking human physiology and instead to construct a self-organising model which learns an appropriate representation from speech data. A brief overview of this speech production system will now be presented [1], along with the results of a preliminary evaluation.

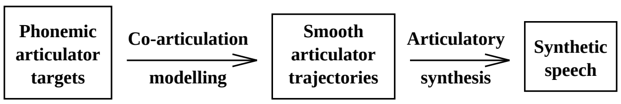

Figure 2. Co-articulation modelling in articulatory speech synthesis.

2. SYSTEM INITIALISATION

An appropriate initialisation is essential for any complex learning system. In this model initialisation data consists of parametrised human speech along with the associated pseudo-articulator trajectories. The term "pseudo-articulator" refers to a variable which is used to parametrise the vocal tract shape, but which is not constrained to model an actual human articulator [9]. These data were generated by using a Kelly–Lochbaum synthesiser [8, 2, 3, 11] to produce a codebook of [pseudo-articulator, parametrised speech vector] pairs. The speech vectors were parametrised using 12 mel-frequency liftered cepstral coefficients excluding log energy.

The codebook generated in this way contains 102488 entries, and unlike other such codebooks described in the literature [12, 9] variable vocal tract length is explicitly incorporated via a parameter which controls the number of vocal tract sections used in the Kelly–Lochbaum synthesiser. Approximately 55% of the codebook entries correspond to vowels semi-vowels and glides, 30% to unvoiced and voiced fricatives, and 15% comprise the nasals. In addition a small number of silence fames are included for modelling stops and inter-word spacing.

Data with which to initialise the system are generated by inverting this codebook to obtain approximate pseudo-articulator trajectories corresponding to a corpus of human speech. Since the inverse mapping from speech spectral vectors to pseudo-articulatory vectors is non-unique, quite different articulator positions can produce approximately the same speech output; hence the choice of articulator positions must also be constrained to follow a smooth geometric path.

In addition, the combination of this non-uniqueness with the fact that the mapping is also non-linear means that it is possible to have non-convex target regions in pseudo-articulator space [5], so the inverse model must also be constrained to a *particular* solution.

Both of these constraints are satisfied in our system by using a dynamic programming algorithm, incorporating both geometric and acoustic cost functions, to map speech spectral vector sequences into corresponding pseudo-articulator sequences. To reduce the computational load, a sub-optimal search was used in the dynamic programming algorithm, which considered only the 500 codebook vectors with the best acoustic match at each step. Figure 3 shows an example of a pseudo-articulator trajectory generated in this way. Phonemic boundaries are shown as dotted lines, and the pseudo-articulator takes on steady values during phonemes, with transitions occurring at phoneme boundaries.

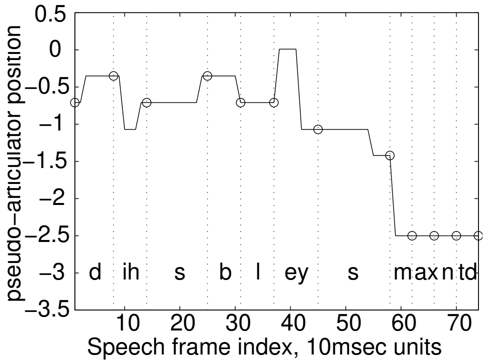

Figure 3. Pseudo-articulator trajectory for "displacement".

We applied this technique to 600 sentences of speech from one male speaker taken from the speaker-dependent portion of the Defence Advanced Research Projects Agency (DARPA) Resource Management (RM) corpus. From the pseudo-articulator trajectories generated, statistics describing the position of each pseudo-articulator at the midpoint of each phoneme were computed.

These statistics can then be used to generate approximate pseudo-articulator trajectories corresponding to arbitrary time-aligned phoneme sequences. At the phoneme midpoints, the pseudo-articulators take on the mean values computed for the phoneme concerned. Trajectories are then generated by a piece-wise linear interpolation between the means, constrained to pass through the average of two adjacent pseudo-articulator targets at the phonemic boundary, as shown in Figure 4.

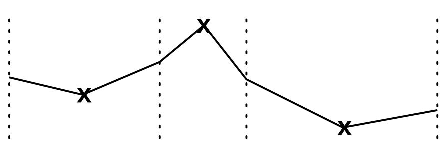

Figure 4. Generation of pseudo-articulator trajectories using constrained linear interpolation.

In this figure, phonemic boundaries are shown by dotted lines, and the target positions for the pseudo-articulator at the phoneme midpoints are shown by crosses.

3. CO-ARTICULATION

3.1. Models of co-articulation

During speech production, considerable flexibility is possible in the positions of the various articulators without causing excessive degradation of the acoustic speech signal. As a result, any particular sound will be produced differently according to the context in which it appears, due to two types of co-articulatory effect [4]:

- Anticipatory co-articulation: occurs from right to left and is due to timing effects – the articulator is "looking ahead" to its next target position and has started to move towards it.
- Carryover co-articulation: occurs from left to right and is due to mechanical inertia – articulators can only move at a finite speed.

This suggests the use of a model in which each pseudo-articulator has a specific target position which would ideally be achieved during the production of a given phoneme, but in which the actual path followed is governed by the dynamics of the particular articulator.

The drawback of such a model is that both the single target position for each phoneme and the articulator-specific dynamics are independent of context. In reality however, the positions of the various articulators will be more or less highly specified according to the phoneme being produced [6]. For example, during the production of the phoneme /p/, the lips are constrained to form a closure, while the tongue body is free to take up the position required for production of the following vowel.

A more accurate model therefore, would be one in which the articulators follow a path of "least resistance", subject to the constraint that their positions at any given time must be sufficiently accurate to produce the desired acoustic output.

3.2. Implementation

One approach to this problem [7] is to specify a *region* through which an articulator must pass in order to produce acceptable acoustic output, with the actual path taken being chosen to minimise articulatory effort. This is effectively what is achieved in our system, in which the position of each pseudo-articulator is specified as a Gaussian distribution about a mean (or "target") value.

The statistics describing these distributions are found by sampling the initial pseudo-articulator trajectories at the phonemic mid-points for each of the 600 training sentences. Co-articulation is then modelled by modifying the pseudo-articulator positions about these means, according to an approximation to the curvature of the pseudo-articulator trajectory at each phonemic midpoint.

These curvatures were approximations to the second derivative of the trajectory, computed as the difference between the gradients of the two linear interpolants on either side of each target. The distribution of this curvature measure for each phoneme was also modelled as a Gaussian distribution estimated from the 600 sentences of training data.

Co-articulation was then implemented by modifying the initial positions of the pseudo-articulator targets up or down according to the local trajectory curvature. For example, if the curvature at a particular phoneme was one standard deviation higher than the mean curvature for that phoneme, then that target was under-shot by one standard deviation from the target mean.

Once the pseudo-articulator target positions have been modified according to their context in this way, the constrained linear interpolation described above is applied to give the actual pseudo-articulator trajectories used.

4. SYSTEM TRAINING

4.1. Acoustic mapping

Our aim when training the production system was to jointly optimise both the pseudo-articulator trajectory shapes and the mapping from these to acoustic vectors. For this reason we replaced the Kelly-Lochbaum vocal tract model used in the generation of the acoustic codebook with an assembly of neural networks, in which each network was trained to approximate the mapping from pseudo-articulator positions to output speech for a particular phoneme. We parametrised the speech using 24 mel-scaled log-spectral coefficients computed every 10msec, with a 25msec Hamming window applied to the input speech.

To generate a sequence of these acoustic vectors corresponding to a given time-aligned phonemic string, the pseudo-articulator trajectories are divided in time at the phoneme boundaries. The trajectories corresponding to each individual phoneme are then passed to the appropriate neural network, and the resulting acoustic vectors are concatenated to yield the parametrised speech vector sequence corresponding to the original utterance.

4.2. Re-estimation

During the training phase, the neural network mappings and the pseudo-articulator statistics are iteratively re-estimated so as to more accurately reproduce the input speech.

The target positions of the pseudo-articulators for each phoneme are re-estimated using the pre-trained neural networks and linearised Kalman filtering. The Jacobian matrices H of the network mappings are derived using an extension of the standard back-propagation formulae to compute the derivative of each network output with respect to each input. An updated estimate $\mathbf{x}$ of the pseudo-articulator position at any point in time can then be found by:

$$\mathbf{x} = \hat{\mathbf{x}} + \hat{P}H^T(H\hat{P}H^T + R)^{-1}(\mathbf{z} - h(\hat{\mathbf{x}}))$$

where $h()$ is the network mapping, R is the network's output error covariance matrix, $\hat{P}$ is the covariance matrix associated with the original estimate $\hat{\mathbf{x}}$ and $\mathbf{z}$ is the observed acoustic vector [1].

This technique was used to re-estimate the positions of each of the pseudo-articulator target positions at each phoneme midpoint, and from these updated values a new set of statistics describing means and standard deviations of pseudo-articulator positions and curvatures were derived. From these, the pseudo-articulator trajectories were re-computed, and the neural network mappings retrained. This process was iterated to determine an optimum set of pseudo-articulator statistics.

5. EVALUATION PROCEDURE

5.1. Generating N-best lists

The system was evaluated on the DARPA RM speaker-dependent evaluation data which comprises 100 sentences read by the same speaker used in the training data set. In order to test the ability of the system to accurately re-synthesise parametrised speech for recognition purposes, we require a set of word-level N-best hypotheses as to the text corresponding to these evaluation sentences.

These hypotheses were generated using the HTK Hidden Markov Model toolkit. A set of speaker-dependent single Gaussian mixture triphone HMMs was trained on the 600 training sentences, and these models were then used to generate up to 100 most likely word transcriptions of the input sentences. While it was possible to generate all 100 transcriptions for most sentences, some extremely short (e.g. 2-word) sentences generated less than 100 different transcriptions.

Since the standard word-pair grammar for the RM corpus is extremely simple and the vocabulary size relatively small (997 words), recognition rates are typically high. Our aim however, is to assess the *acoustic* modelling achieved by the speech production model, hence the word N-best transcriptions were generated using a null grammar.

Time-aligned phonemic transcriptions corresponding to each word-level N-best hypothesis were generated using a forced Viterbi alignment of the word transcriptions to the parametrised speech vectors. This algorithm uses dynamic time-warping to find the most likely phoneme sequence consistent with a given word sequence. The result is a set of phoneme strings together with a time alignment, suitable as input to the speech production model.

5.2. Re-scoring N-best lists

For each time-aligned phonemic transcription, a corresponding set of co-articulated pseudo-articulator trajectories was derived and passed to the appropriate neural networks, which produce 24-dimensional acoustic vectors at their outputs. For each transcription in the N-best list the resulting sequence of acoustic vectors is compared with the sequence of acoustic vectors corresponding to the original speech being recognised. The entry in the N-best list which gives the closest match to the original is selected as the most likely transcription.

When computing the distance between a given synthetic acoustic vector and the actual acoustic vector corresponding to it, we use a Mahalanobis distance measure which incorporates our confidence in the synthetic vector coefficients. Since we have previously computed the error covariance matrix R associated with each neural network mapping, the standard deviation associated with each coefficient in the synthetic vector can be used to scale the difference between this vector and that corresponding to the actual speech. The total distance between the actual and synthetic vector sequences is therefore determined by summing the Mahalanobis distances between each of the individual vectors.

Since we also have prior information as to the probability of a given entry in the N-best list representing the correct transcription of the input speech, this knowledge is incorporated into the scoring process. This probability decreases roughly exponentially with depth in the N-best list, so we weighted the scores with a function which decreases exponentially over the first 15 N-best scores, and is constant thereafter.

6. RESULTS

6.1. Re-scoring example

An example of the successful re-scoring of a sentence is shown in Figures 5, 6 and 7. These show smoothed acoustic vector sequences corresponding to a section of an utterance in which the speaker said the word "conventional". In this case the $N = 1$ hypothesis incorrectly transcribed this section as "can change no", whereas the $N = 6$ hypothesis contained the correct transcription.

Figure 5 shows the sequence corresponding to the original speech, and Figures 6 and 7 show the sequences synthesised by the SPM from the $N = 6$ and $N = 1$ transcriptions respectively. As can be seen from these diagrams, Figure 6 gives both a better alignment and a closer spectral match to the original speech than Figure 7, and hence is the transcription chosen as corresponding to the utterance.

While the gross spectral characteristics of the synthesised vector sequences correspond roughly with those in the original speech, the undesirable discontinuities introduced by the concatenation of the outputs of the phoneme-specific neural networks are clearly visible.

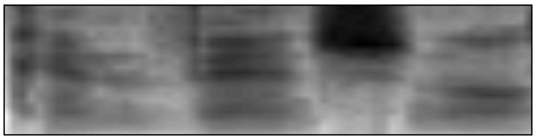

Figure 5. Acoustic output for original speech fragment: "conventional".

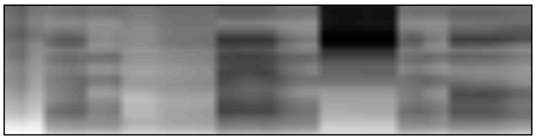

Figure 6. Acoustic output for synthetic speech from $N = 6$ transcription: "conventional".

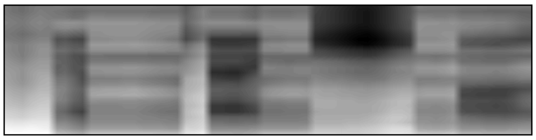

Figure 7. Acoustic output for synthetic speech from $N = 1$ transcription: "can change no".

6.2. Overall results

The performances of HTK and the SPM are measured in terms of:

- H, the number of correct word labels.
- D, the number of deleted labels.
- S, the number of substituted labels.
- I, the number of inserted labels.

From these, the percentage of labels correctly recognised is given by:

$$Percent\ Correct = \frac{N - D - S}{N} * 100\%$$

where N is the total number of labels. The overall accuracy also incorporates the number of insertions, and is computed as:

$$Percent\ Accuracy = \frac{N - D - S - I}{N} * 100\%$$

Over the 100 test sentences, HTK and the SPM both score 82.07% of words correctly identified, with the same number of deletions and substitutions. However, the SPM has many more insertions than does HTK, and hence a lower accuracy score. This indicates that the SPM is favouring transcriptions in which additional words have been inserted which give a good short-term match to the original speech, yielding slightly improved scores.

This is supported by the observation that when a threshold is applied to the re-scoring process, such that an N-best list is only re-ordered if the difference between the $N = 1$ score and the $N = k$ score is greater than this value, the SPM accuracy increases to 75.61% as compared with 75.95% for HTK.

While these results demonstrate that it is possible to achieve results comparable to those of a simple HMM by using a SPM to re-score N-best lists, it is clear that greatly improved acoustic modelling will be necessary if this technique is to be successfully applied to more complex tasks.

7. CONCLUSIONS

The field of speech recognition using speech production models is still in its infancy, yet appears to show great promise. In this paper a pseudo-articulatory speech production system has been presented which is self-organising and has the potential to address the problem of co-articulation modelling encountered by current continuous speech recognition systems. The co-articulation scheme proposed not only models geometric constraints, but is also consistent with "window" models in which the relative degrees of specification or under-specification of pseudo-articulators are explicitly modelled.

By using a set of pseudo-articulators which are not constrained to human physiology and an assembly of neural networks to learn the mapping from the pseudo-articulatory trajectories to output speech, both the trajectories and the acoustic mapping can be jointly and iteratively optimised from an initialisation state, thus avoiding the problems of accurately modelling the human speech production process.

When evaluated on speech drawn from the DARPA RM speaker-dependent corpus, the system described exhibits an acoustic modelling performance which is comparable to that of a simple HMM, however significant improvement is clearly necessary before the system could be applied to more difficult problems.

Considerable work will be required before production-based systems such as that presented in this paper will be able to improve upon the results obtained by state-of-the-art recognisers. However, at the present time phonemic variation due to co-articulatory effects is the significant factor limiting the performance of large vocabulary continuous speech recognition systems, due to the inability to provide sufficient training data to train context-dependent models. Speech production models such as that described in this paper may prove a viable technique for addressing this problem.

REFERENCES

[1] C. S. Blackburn and S. J. Young. "A novel self-organising speech production system using pseudo-articulators". *Int. Congr. Phon. Sc.*, 1995. Accepted for publication.

[2] G. Fant. *Acoustic theory of speech production.* Mouton & Co., The Hague, 1970. First published 1960.

[3] J. L. Flanagan. *Speech analysis synthesis and perception.* Springer-Verlag, 2 edition, 1972.

[4] T. Gay. "Articulatory movements in VCV sequences". *J. Acoust. Soc. Am.*, 62(1):183–193, July 1977.

[5] M. I. Jordan and D. E. Rumelhart. "Forward models: Supervised learning with a distal teacher". *Cog. Sc.*, 16:307–354, 1992.

[6] P. A. Keating. "Underspecification in phonetics". In *Coarticulation*, pages 30–50. UCLA phonetics laboratory, 1988.

[7] 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.

[8] J. L. Kelly Jr. and C. Lochbaum. "Speech synthesis". In *Sp. Comm. Sem.*, Stockholm, 1962.

[9] P. Meyer, R. Wilhelms, and H. W. Strube. "A quasiarticulatory speech synthesizer for German language running in real time". *J. Acoust. Soc. Am.*, 86(2):523–539, 1989.

[10] R. C. Rose, J. Schroeter, and M. M. Sondhi. "An investigation of the potential role of speech production models in automatic speech recognition". In *Proc. Int. Conf. Sp. Lang. Proc.*, volume 2, pages 575–578, 1994.

[11] P. Rubin, T. Baer, and P. Mermelstein. "An articulatory synthesizer for perceptual research". *J. Acoust. Soc. Am.*, 70(2):321–328, Aug. 1981.

[12] J. Schroeter and M. M. Sondhi. "Techniques for Estimating Vocal-Tract Shapes from the Speech Signal". *IEEE Trans. Sp. Aud. Proc.*, 2(1):133–150, Jan. 1994.

[13] M. M. Sondhi and J. Schroeter. "A hybrid time-frequency domain articulatory speech synthesizer". *IEEE Trans. Acoust. Sp. Sig. Proc.*, ASSP-35(7):955–967, July 1987.