A SELF-LEARNING SPEECH SYNTHESIS SYSTEM

C. S. Blackburn and S. J. Young
Cambridge University Engineering Department (CUED)
Cambridge CB2 1PZ, UK
email: csb@eng.cam.ac.uk

ABSTRACT

We describe a self-organising pseudo-articulatory speech production model (SPM), and present recent results when training the system on an X-ray microbeam database. The SPM extracts statistics describing articulator positions and curvatures during the production of continuous speech, then applies an explicit co-articulation model to generate synthetic articulator trajectories corresponding to time-aligned phonemic strings. A set of artificial neural networks estimates parameterised speech vectors from the synthetic articulator traces. We present an analysis of the articulatory information in the X-ray microbeam database used, and demonstrate the improvements in articulatory and acoustic modelling accuracy obtained using our co-articulation system.

Nous décrivons un modèle auto-organisatif pseudo-articulatoire de la production de la parole (SPM), et présentons des résultats obtenus sur une base de données contenant de micro-faisceaux de rayons X. Le SPM extrait des statistiques décrivant les positions et courbures des articulateurs lors de la production de la parole continue, puis met en application un modèle explicite de la co-articulation pour générer des trajectoires articulatoires synthétiques qui s'accordent avec une séquence de phonèmes aligné dans le temps. Un ensemble de réseaux neuromimétiques estime les vecteurs paramétriques de la parole en fonction des traces synthétiques articulatoires. Nous présentons une analyse du contenu informationnel de la base de données utilisée, et montrons l'amélioration des modèles articulatoires et acoustiques obtenu en utilisant notre système co-articulatoire.

INTRODUCTION

The achievable modelling accuracy of any self-learning system is highly dependent upon the quality of the data from which its parameters are extracted. In previous papers we described a self-organising SPM which was trained on synthetic articulatory data derived using a codebook of (pseudo-articulator vector, acoustic vector) pairs. A Kelly-Lochbaum synthesiser was used to generate the codebook entries, and dynamic programming was used to invert this codebook to obtain pseudo-articulatory traces corresponding to speech taken from the Defence Advanced Research Projects Agency (DARPA) Resource Management (RM) corpus [1].

While this provides a useful starting point in the absence of physical articulatory data, there are two significant sources of inaccuracy in such a model. The first of these is due to the fact that the articulatory system which is used to generate the codebook is synthetic, and usually will not correspond closely to that of the speaker whose speech is used in the inversion process. Secondly, this process itself is sub-optimal due to the significant quantisation errors introduced in order to keep codebook size practical, and the necessary trade-off between acoustic and articulatory accuracy.

Until recently however, some such method for obtaining a data set was needed, due to the lack of a physical database providing measured articulator traces accompanied by synchronous speech recordings. In this paper we now present the results of applying our articulatory parameter extraction techniques and subsequent parameterised speech synthesis algorithm to data from the University of Wisconsin (UW) X-ray microbeam (XRMB) speech production database [4].

DATABASE FORMAT

The UW XRMB database contains sampled articulator position traces along with synchronously recorded speech waveforms for 57 speakers of American English, comprising 32 females and 25 males. The corpus contains sentences (40%), citation words and sound sequences (33%), prose passages (13%), oral motor tasks (8%) as well as counting and sequences of number names (6%). For each of these a nominal word-level transcription is provided, although subjects occasionally deviated from this text.

Acoustic data

The speech signal was recorded using a directional microphone in the presence of machine noise at a sampling period of 46μs (approximately 21739 Hz). A fixed recording period was used for each task, which occasionally resulted in truncated recordings for slower speakers. In addition, a short tone is played at the start of each task, and background comments such as "good" and "rep" are present at the end of many utterances.

Articulatory data

The articulator positions were determined using a narrow X-ray beam to track the movements of gold pellets glued to the tongue, jaw and lips of a subject while reading from the set corpus. Three reference pellets were attached to the subject's head, and a total of eight articulator pellets were tracked relative to these, with the subject's head viewed in profile by the apparatus. The eight pellets are denoted UL (upper lip), LL (lower lip), $T1$ to $T4$, (tongue positions 1 to 4 where 1 is closest to the tip), MNI (mandible incisor), and MNM (mandible molar).

The x and y positions of each of these pellets were recorded at sample rates which varied according to the relative observed accelerations of the various articulators, and were then interpolated and re-sampled at a uniform sampling period of 6.866ms (approximately 146 Hz).

Due to limitations imposed by the experimental set-up, the XRMB system occasionally mistracks pellets during recording due to the loss of a pellet trace, or as a result of confusion between two 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 articulators. These tracking errors have been identified and marked by hand for a subset of the speakers, and for the speaker used in this work (jw18), 21% of the records used were found to contain a tracking error in one or more articulator pellets.

DATABASE PREPARATION

Data preprocessing

The raw acoustic signal contained significant noise at half the Nyquist frequency – approximately 5435Hz – which was removed using a notch filter. The resulting signal was down-sampled to 16kHz by interpolating 92 times up to 2MHz, then down-sampling 125 times. The 16kHz speech was parameterised using the CUED HTK hidden Markov model (HMM) toolkit into both 24-dimensional log Mel-frequency log filter bank (FBANK) coefficients and 12-dimensional Mel-frequency cepstral (MFCC) coefficients. 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.

The 16 articulator waveforms (x and y positions of 8 pellets) were also interpolated, and re-sampled at intervals of 10ms starting from 12.5ms to give values corresponding to the centres of the parameterised speech frames. The word-level transcriptions supplied with the database were hand-edited to correspond with the actual text spoken, including nonsense transcriptions corresponding to truncated words at the ends of utterances.

Generation of alignments

A phonetic dictionary for the words in the XRMB database using the RM phone set was constructed by merging and editing relevant entries from the RM and LIMSI-ICSI dictionaries, and adding entries corresponding to truncated utterance endings.

A set of monophone HMMs with two emitting states for stop and diphthong phonemes and three for the remaining phonemes was trained using HTK on MFCC parameterised speech from the RM speaker-independent corpus. Separate three-state monophone HMMs corresponding to the tone played at the start of each utterance and the "good" and "rep" background comments found after many utterances were trained on parameterised speech vectors extracted by hand from 21, 16 and 5 examples of each sound respectively.

These model sets were then combined and used with the hand-edited transcriptions and dictionaries to train a set of speaker-dependent 5-mixture monophone HMMs on the speech of one speaker (jw18). Sentences, citation words and number sequences were used as training data, and the prose passages were segmented into individual sentences by hand for use as test data. A state-level forced Viterbi alignment of the data sets to the transcriptions was then performed, to yield a data set labelled at the subphoneme level.

In the case of stop phonemes the two states align to the occlusion and burst sections of the phoneme respectively, where the burst state is optional; in the case of diphthongs these align to the initial and final voiced sections. In both instances each state is treated as a separate phoneme, giving an expanded set of 60 "phonemes".

ARTICULATORY MODEL

Using these phonemic alignments, each articulator trace is sampled at points corresponding to the centres of the various phonemes, to give a set of statistics describing mid-phonemic articulator positions for each phoneme. The resulting positional variations are modelled by single Gaussian distributions, where deviations from the mean positions are due both to random variations in articulator positioning, and to anticipatory and carryover co-articulation.

To synthesise articulator trajectories corresponding to an arbitrary phonemic string, we must predict the direction and magnitude of the co-articulatory movement away from these mean positions from knowledge of the time-aligned phonemic string alone. This is achieved by observing that the variation in articulator position in most cases is strongly correlated with the curvature of the trajectory at the point concerned: relatively high and low curvatures tend to give undershoot and overshoot of the mean position respectively.

For each instance of each phoneme we therefore also compute a measure of the approximate curvature of each articulator trajectory at the phonemic midpoint, computed as the difference between the gradients leading out of and into the midpoint when linear interpolation is used between successive phonemic means.

These curvature statistics are also modelled as single Gaussian distributions, so that the position and curvature statistics describing articulator behaviour during the production of a given phoneme form a bi-normal distribution. By computing the correlation coefficients between the curvature and positional statistics we can then predict the articulator's deviation from the mean using only a knowledge of the phoneme sequence.

Given such a set of co-articulated time-aligned mid-phonemic positions, complete articulator trajectories are synthesised by using linear interpolation constrained to pass through the average of two adjacent co-articulated target positions at the phonemic boundary. To enhance the system's robustness to unusual contexts given the small size of the training data set, a low-order low pass filter is applied to the resulting trajectories to remove very sharp articulator movements which are otherwise observed in approximately 0.3% of phonemes.

ACOUSTIC MODEL

The non-linear mapping from synthetic articulator trajectories to FBANK parameterised speech vectors is approximated by a set of artificial neural networks (ANNs), where a single ANN can be used for each phoneme, or else data sets from similar phonemes can be combined into a single mapping.

The articulator inputs to the ANNs were scaled by the mean and standard deviation for each individual articulator computed over the entire training set to yield a majority of inputs in the range [-1,1]. The target vectors were chosen as FBANK vectors as these result in a less compact but simpler acoustic mapping than do MFCC parameters, due to the absence of the cosine transform.

In all cases the ANNs were trained using resilient back-propagation (RPROP), which gives much faster training times than simple back-propagation. The number of hidden nodes was varied to give optimum results, and in all cases cross-validation was used to prevent over-training.

RESULTS

Articulatory modelling

Statistics describing articulator positions and curvatures at phonemic midpoints were computed for the training data set, along with their corresponding correlation coefficients. Figure 1 shows an example of the correlation coefficients for the x and y co-ordinates of 8 articulators for the phoneme /s/.

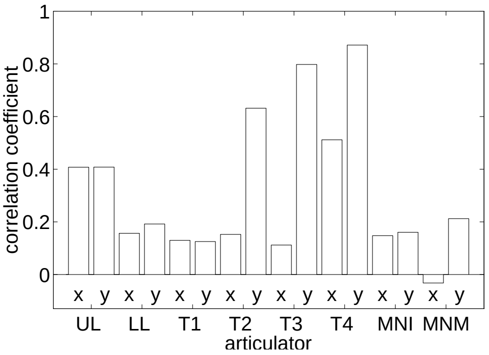

Figure 1: Correlation coefficients for phoneme /s/.

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

Synthetic articulator trajectories were then constructed both with and without co-articulation from time-aligned phonemic strings produced by forced alignment of the transcriptions to the training and test data sets. The errors between the synthetic and X-ray trajectories were computed at all points, and the use of co-articulation gave a reduction in error for all training sentences, and in 24 of 25 test sentences.

The mean error for each articulator scaled by its mean and standard deviation over the entire training set was computed, and each articulator's error was again found to decrease with co-articulation on the training set. The results for the test set are shown in Figure 2.

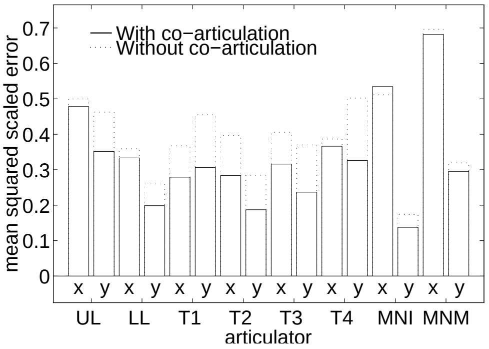

Figure 2: Test set errors by articulator.

The errors for all articulators except the x position of the mandibular incisor decrease with co-articulation. In general the errors in tongue position are less than for lip and jaw position, with the x position of the front and back of the jaw being most poorly modelled. This is as expected since the extension of the jaw has relatively little effect on the acoustic signal, and we expect much of the variation in this articulator to be random movement.

An example of the effects of co-articulation on a synthetic articulator trajectory is given in Figure 3, for the articulator which is is most affected by the co-articulation model, $T4y$.

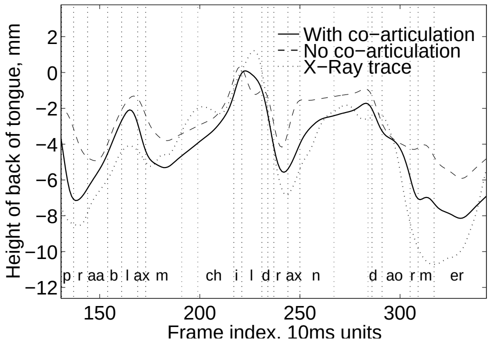

Figure 3: Movement in height of tongue back ($T4y$) for utterance: “problem, children, dormer”.

As expected, the back of the tongue is relatively high during the phoneme /l/ and relatively low during /r/, and the co-articulation model has resulted in a closer approximation to the X-ray trajectory.

Acoustic modelling

Separate sets of ANNs were trained to learn the mappings to FBANK parameterised speech vectors from synthetic articulator trajectories with and without co-articulation, as well as from the original X-ray data.

The mean output errors for the networks trained on synthetic co-articulated input trajectories were lower than the corresponding errors without co-articulation for 80% of phonemes. Of the remaining 20%, half of these showed an error increase of less than 3%; of those with significant error increases, the majority were for uncommon phonemes, for which very little training data was available, resulting in less accurate co-articulation statistics.

Similarly, the networks trained on synthetic co-articulated input gave lower output errors than those trained on the original X-ray articulator data in 77% of phonemes, despite the inherently less accurate inputs used by the former. This result is probably due to the fact that there is significantly more random variation in articulator positions for the X-ray data by comparison with the synthetic traces, which deviate from their mean positions in a systematic way.

FUTURE WORK

In a previous paper [3] we applied a SPM trained on a synthetic articulatory database to the task of word recognition for the RM corpus. We now intend to use the SPM trained on data from the UW XRMB database to the tasks of both phoneme and word recognition, by re-scoring N-best lists produced by a traditional recognition system such as HTK.

It would also be instructive to examine the effects of combining data sets from different phonemes before training the acoustic mappings, and to investigate modifications to the articulator set. These could take the form of dimension reduction using either principal component or linear discriminant analysis, or else an increase in dimensionality via the generation of additional articulatory inputs, as demonstrated for the SPM trained on RM data [2].

CONCLUSION

In this paper we presented a speech production model which extracts its parameters automatically from an X-ray articulatory database. Statistics describing articulator positions and curvatures during the production of 60 phonemes were extracted and modelled with single Gaussian distributions. Significant correlations were found between these two sets of statistics for many articulators and phonemes, indicating that curvature may be a useful predictor of co-articulatory variation in articulator positions.

We demonstrated a method for synthesising articulator trajectories from time-aligned phonemic strings, and showed that the use of our simple co-articulation model significantly improves modelling accuracy. Sets of artificial neural networks were used to approximate the mappings from articulator trajectories to parameterised speech vectors, and the networks supplied with synthetic co-articulated articulator traces trained with lower output errors than did both those without co-articulation and those using the original X-ray trajectories as inputs.

ACKNOWLEDGEMENTS

The authors would like to thank Professor John Westbury and his team at the University of Wisconsin for preparing and making publicly available the invaluable resource that is the UW XRMB database.

Acknowledgements are also due to LIMSI-CNRS and ICSI for access to their phonetic dictionaries.

REFERENCES

[1] C. S. Blackburn and S. J. Young. “A novel self-organising speech production system using pseudo-articulators”. *Int. Congr. Phon. Sc.*, 2:238–241, 1995.

[2] C. S. Blackburn and S. J. Young. “Learning new articulator trajectories for a speech production model using artificial neural networks”. *IEEE Int. Conf. Neural Net.*, 4:2046–2051, 1995.

[3] C. S. Blackburn and S. J. Young. “Towards improved speech recognition using a speech production model”. *Europ. Conf. Sp. Comm. Tech.*, 3:1623–1626, 1995.

[4] J. R. Westbury. “X-Ray microbeam speech production database user's handbook”. Personal communication, University of Wisconsin, 1994.