PSEUDO-ARTICULATORY SPEECH SYNTHESIS FOR RECOGNITION USING AUTOMATIC FEATURE EXTRACTION FROM X-RAY DATA

C.S.Blackburn and S.J.Young

Cambridge University Engineering Department (CUED), UK
csb@eng.cam.ac.uk

ABSTRACT

We describe a self-organising pseudo-articulatory speech production model (SPM) trained on an X-ray microbeam database, and present results when using the SPM within a speech recognition framework. Given a time-aligned phonemic string, the system uses an explicit statistical model of co-articulation to generate pseudo-articulator trajectories. From these, parametrised speech vectors are synthesised using a set of artificial neural networks (ANNs). We present an analysis of the articulatory information in the database used, and demonstrate the improvements in articulatory modelling accuracy obtained using our co-articulation system. Finally, we give results when using the SPM to re-score N-best utterance transcription lists as produced by the CUED HTK Hidden Markov Model (HMM) speech recognition system. Relative reductions of 18% in the phoneme error rate and 15% in the word error rate are achieved.

1. INTRODUCTION

A framework for the use of our speech production model for improving speech recognition results is illustrated in Figure 1.

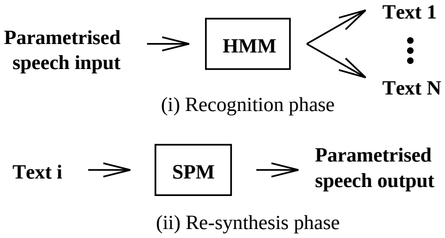

Figure 1: Recognition framework overview.

The HMM recogniser provides an ordered list of N hypothesised transcriptions of an utterance, and from each of these parameterised speech is synthesised by the SPM. The N-best list is then re-ordered according to the errors between the synthesised and original speech. Since one of the limiting factors in the performance of HMM systems is their ability to model contextual variation in the speech signal, we hope to provide more accurate acoustic modelling in the SPM by using an explicit time-domain model of co-articulation.

In previous papers we described a SPM trained on articulator traces synthesised from the Resource Management (RM) corpus. A Kelly-Lochbaum synthesiser was used to generate a codebook of (pseudo-articulator vector, acoustic vector) pairs, and dynamic programming was used to invert this codebook to obtain pseudo-articulatory traces corresponding to speech utterances [2]. The model was then used to re-score N-best word-level transcriptions of these utterances as provided by HTK [3]. Since the system automatically extracts its parameters during training, the achievable modelling accuracy is highly dependent upon the quality of the training data set. In this case, the mis-match between the Kelly-Lochbaum synthesiser and the vocal tract of the speaker, as well as the quantisation error in the codebook lead to significant modelling inaccuracies.

We now present a SPM which was trained on data from the University of Wisconsin (UW) X-ray microbeam (XRMB) speech production database [5], and which incorporates significant improvements in both the articulatory and acoustic models. The system is capable of accurately predicting articulator movements, and when used to re-score N-best utterance transcriptions results in a reduction in recognition error rates over the standard HMM system.

2. DATABASE PRE-PROCESSING

The UW XRMB database contains 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, nominal word-level transcriptions are provided which we hand-edited to correspond to the actual text spoken, including nonsense transcriptions at the ends of some utterances which were truncated.

2.1. 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 was played at the start of each task, and background comments such as “good” and “rep” are present at the end of many utterances.

We filtered this raw acoustic signal using a notch filter to remove background noise at 5435Hz, and down-sampled the resulting signal to 16kHz. The 16kHz speech was then parameterised into 24-dimensional Mel-frequency log spectral 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.

2.2. Articulatory Data

Articulator positions in the UW system are 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 eight articulator pellets were tracked relative to these, with the subject's head viewed in profile by the apparatus: 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 pellet were recorded at sample rates which varied according to the accelerations of the articulators, and were then interpolated and re-sampled at a uniform sampling period of 6.866ms ($\approx$ 146 Hz) before inclusion in the database.

We then re-interpolated the articulator waveforms, and re-sampled them at intervals of 10ms starting from 12.5ms to give values corresponding to the centres of the parameterised speech frames.

2.3. Generation of Alignments

A phonetic dictionary for the 440 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 stops and diphthongs and three for other phonemes was trained using HTK on MFCC parameterised speech from the RM speaker-independent corpus. In the case of stop phonemes the two states align to the occlusion and burst, 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".

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 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, number sequences and prose passages were used, with one quarter of the data ($\approx$ 1500 phonemes) set aside as a test set. A state-level forced Viterbi alignment of the data to the transcriptions was performed, to yield a data set labelled at the sub-phoneme level.

3. ARTICULATORY MODEL

Using these alignments the articulator traces were sampled at the midpoints of each phoneme, and the resulting positional variations were modelled by single Gaussian distributions, where deviations from the mean positions are due both to random positional variations, and to anticipatory and carryover co-articulation [4].

Thus, to synthesise pseudo-articulator trajectories, we must predict co-articulatory movements away from these mean positions from a knowledge of the time-aligned phonemic string alone. We found that variation in articulator midpoint position is strongly correlated with the *curvature* of the trajectory at the point concerned, computed as the difference between the gradients leading out of and into the midpoint when linear interpolation is used between successive phonemic means. Relatively high and low curvatures tend to give undershoot and overshoot of the mean position respectively.

We therefore computed this curvature measure for each instance of each phoneme, again modelling the variations with single Gaussians. The position and curvature statistics describing articulator behaviour during the production of a given phoneme now form a bi-normal distribution, and by computing the correlations between curvature and position we can predict an articulator's deviation from the mean using a knowledge of the phonemic sequence alone. Figure 2 shows an example of the correlation coefficients for the x and y co-ordinates of the 8 articulators for the phoneme /s/.

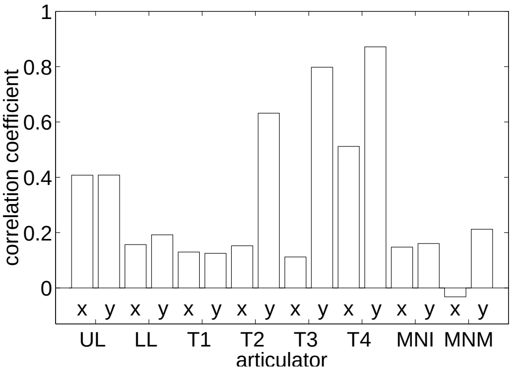

Figure 2: Correlation coefficients for phoneme /s/.

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 and upper lip are relatively free to move to positions dictated by neighbouring phonemes, as evidenced by the larger correlations for UL and $T2$ to $T4$.

Complete articulator trajectories were then generated by linear interpolation between successive co-articulated time-aligned mid-phonemic positions. 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 to remove very sharp articulator movements which are otherwise observed in approximately 0.3% of phonemes.

Synthetic trajectories were 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 and test sentences.

The mean error for each articulator over the training and test sets was computed, and scaled using positional means and standard deviations. Once again, each articulator's error decreased with co-articulation in both cases. The results for the test set are shown in Figure 3, where the errors in tongue position are generally less than for lip and jaw position, with the x position of the upper lip and the front and back of the jaw being most poorly modelled.

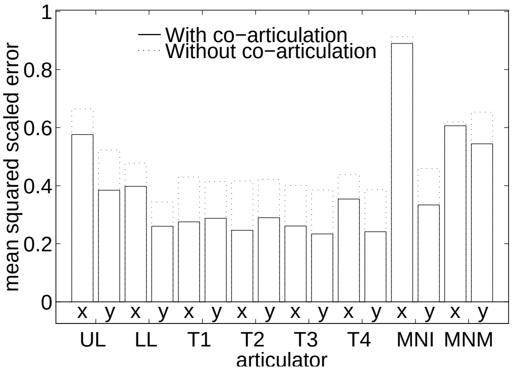

Figure 3: Test set errors by articulator.

This is as expected since the extension of these articulators has relatively little effect on the acoustic signal, and we expect much of the variation in this articulator to be random movement, as evidenced by low correlation coefficients and correspondingly low impact of the co-articulation model on the articulators.

An example of the effects of co-articulation on a synthetic pseudo-articulator trajectory is given in Figure 4, for the articulator most affected by the co-articulation model, $T4y$.

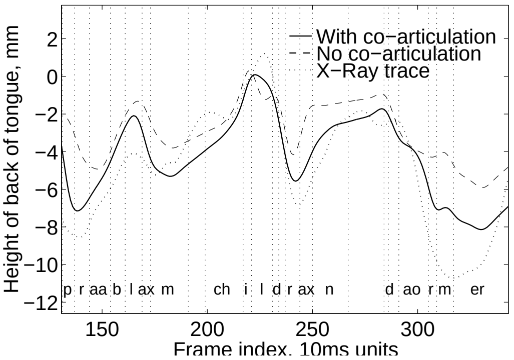

Figure 4: Movement in height of tongue back ($T4y$) for utterance: “problem, children, dormer”.

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.

4. ACOUSTIC MODEL

The non-linear mapping from pseudo-articulator trajectories to parameterised speech vectors was approximated by a separate ANN for each phoneme, pre-scaling the inputs by the mean and standard deviation of each articulator computed over the training set.

The target vectors were chosen as log spectral vectors as these result in a less compact but simpler acoustic mapping than do MFCC parameters, due to the absence of the cosine transform. To provide energy normalisation, the mean of each vector was subtracted before training the networks. As a result, the outputs of the ANNs do not 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 parameter for each phoneme were also computed over the entire training set to provide a separate model of spectral energy levels.

In all cases the networks were trained using resilient back-propagation (RPROP), where the number of hidden nodes was varied to give optimum results, and in all cases cross-validation was used to prevent over-training. The variance of the network output predictions due to noise in the target data was computed over the training set and used as a measure of confidence in the values predicted by the networks [1].

Our acoustic model therefore consists of a set of mean and variance statistics describing the log spectral energy levels of each parameter of each phoneme, along with a separate ANN mapping for each phoneme which can be used to predict spectral shapes (along with associated error variances) from articulator positions.

5. SCORING N-BEST TRANSCRIPTIONS

We generated N=100 word-level transcriptions for each test set utterance using HTK with the speaker-dependent 5-mixture monophone HMMs estimated on the training data. No language model was used, since our aim was to compare the *acoustic* modelling accuracy of HTK and the SPM. For each of these transcriptions a state-level alignment to the acoustic data was performed, and from these alignments synthetic pseudo-articulator trajectories and spectral shape vectors were synthesised.

Since the shape and energy of the spectral vectors are modelled separately, when comparing a sequence of synthetic speech vectors with the corresponding original vectors, two separate error measures are computed.

The first of these compares the energy levels of an actual speech vector with the statistics describing the expected energy levels for the phoneme to which it is hypothesised to correspond. The second measure computes the difference between the spectral shape of each original speech vector and the spectral shape predicted for the corresponding synthetic vector by the ANN concerned. These two error measures are then combined to give an overall error:

$$
\begin{aligned}
E \quad = \quad \sum_{vectors} \Bigg[& K_s \sum_{i=1}^{n} \left(\frac{s_{i,p} - \left(t_i - \frac{1}{n}\sum_{i=1}^{n} t_i\right)}{\sigma_{s_{i,p}}} \right)^2 + \\
& K_e \sum_{i=1}^{n} \left(\frac{e_{i,p} - t_i}{\sigma_{e_{i,p}}} \right)^2 \Bigg]
\end{aligned}
$$

where for each vector p is the phoneme aligned to it in the transcription, $s_{i,p}$ is the spectral shape coefficient synthesised by the ANN corresponding to phoneme p with error standard deviation $\sigma_{s_{i,p}}$, t_i is the actual (target) log spectral coefficient, $e_{i,p}$ is the mean energy coefficient for phoneme p, with standard deviation $\sigma_{e_{i,p}}$ computed over the entire training set, and n is the spectral vector's dimension.

The constants K_s and K_e are weighting factors on the error terms, where $K_s > K_e$ since the shape error is more accurately determined due to the larger number of parameters used in its estimation. The error is also weighted according to the depth of the transcription in the N-best list, to reflect the decreasing prior probability of finding a correct transcription with increasing depth in the list.

Due to small errors in the alignment of the phoneme sequences to the acoustic data, both the shape and energy error terms are dominated by errors within one frame of phoneme 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. To alleviate this problem during the computation of the error measure above, boundaries delimiting stops and nasals are re-aligned by one frame if this results in a reduction in the error.

6. RESULTS

Results for both phoneme and word recognition over 50 test utterances with $K_s = 5K_e$ are given in Figures 5 and 6, up to a depth of N=25. Phoneme recognition results are provided since although the N-best transcriptions were produced at the word level, word-level results can be misleading when assessing acoustic accuracy, since a homonym word error will introduce no phoneme errors.

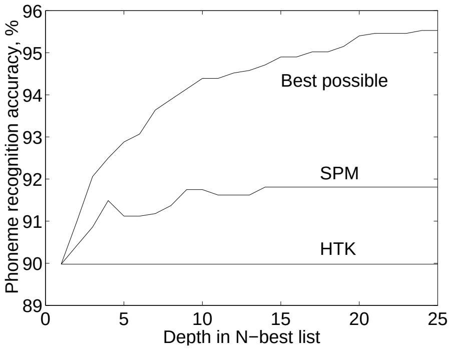

Figure 5: Phoneme recognition results.

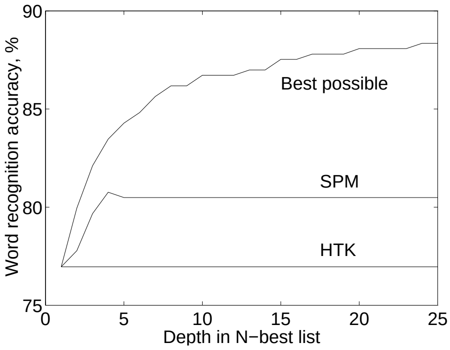

Figure 6: Word recognition results.

The "best possible" curve represents hand-selecting the best transcription available up to a given depth; this curve therefore shows the maximum achievable results when re-scoring the N-best lists. The "HTK" curve represents the baseline performance achieved by always selecting the N=1 transcription, and the "SPM" curve shows results when rescoring the transcriptions up to a given depth with our speech production model. The "SPM" curve becomes flat for large N, as the prior probability of finding a correct transcription becomes very small. Results for large N are summarised in Table 1.

Type	HTK error	SPM error	Improvement
Phoneme	10.02	8.19	18.3%
Word	23.04	19.51	15.3%

Table 1: Recognition results.

7. CONCLUSIONS

We described a speech production model which automatically extracts its parameters from articulatory and acoustic data from the University of Wisconsin X-ray microbeam database.

The articulator traces in the database were aligned and sampled at points corresponding to phonemic midpoints, and trajectory curvatures were estimated at these points. We found significant correlations between positional and curvature variations, and used these as the basis for a statistical model of co-articulation. By applying this model to a set of synthetic pseudo-articulator traces we demonstrated an improvement in articulatory modelling accuracy, as compared with the original X-ray traces.

The HTK speech recognition system was used to generate N=100 ordered hypotheses as to the correct transcription of each of a set of 50 test utterances. From each of these transcriptions pseudo-articulator traces were synthesised, and from these speech spectral vectors were predicted using a set of artificial neural networks optimised using cross-validation on the training data set.

By comparing the synthetic and actual speech spectral vectors for each transcription of each test set utterance, the entries in the N-best list were re-ordered, yielding an 18% decrease in phoneme error rates, and a 15% reduction in word error rates relative to HTK.

8. 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.

9. REFERENCES

1. C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.
2. 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.
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. 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.
5. J. R. Westbury. "X-Ray microbeam speech production database user's handbook". Personal communication, University of Wisconsin, 1994.