MODELLING ASYNCHRONY IN SPEECH USING ELEMENTARY SINGLE-SIGNAL DECOMPOSITION

M J Tomlinson, M J Russell, R K Moore, A P Buckland and M A Fawley

Speech Research Unit, DRA Malvern
St Andrews Road, Malvern, Worcs. WR14 3PS, UK
phone: (+44) 1684 894105, fax: (+44) 1684 895103
e-mail: mjt@signal.dra.hmg.gb

ABSTRACT

Although the possibility of asynchrony between different components of the speech spectrum has been acknowledged, its potential effect on automatic speech recogniser performance has only recently been studied. This paper presents the results of continuous speech recognition experiments in which such asynchrony is accommodated using a variant of HMM decomposition. The paper begins with an investigation of the effects of partitioning the speech spectrum explicitly into sub-bands. Asynchrony between these sub-bands is then accommodated, resulting in a significant decrease in word errors. The same decomposition technique has previously been used successfully to compensate for asynchrony between the two input streams in an audio-visual speech recognition system.

1. INTRODUCTION

A factor which has been noted by some researchers in the acoustic-phonetic field is the possibility of asynchrony between the different components of the acoustic parameterisations which are typically used for automatic speech recognition. This asynchrony may be due to a number of factors, for example underlying asynchrony between different parts of the human speech production system, or asynchrony introduced by a communication channel. Conventional HMM-based approaches to automatic speech recognition use model states which are associated with complete synchronous feature vectors and are therefore unable to accommodate these effects. One possible solution is to construct independent sub-band HMMs, to conduct separate classification experiments by applying these HMMs to the appropriate sub-band representation of the speech pattern, and to combine the separate results [1]. This paper presents the results of an investigation into an alternative strategy to determine the effects of accommodating asynchrony between the upper and lower frequency bands of the speech spectrum using *integrated* models based on HMM decomposition [2].

This approach can be viewed as an elementary form of *single-signal decomposition* (SSD) [3] - the decomposition of a speech signal into parallel asynchronous components, corresponding, for example, to the states of the principal articulators in the vocal tract. Such an approach extends the conventional HMM framework by attempting to model the underlying speech production process, rather than simply represent detailed surface behaviour [4]. The proposed method has the potential to offer significant advantages for automatic speech recognition by reducing the reliance on an explanation of variations in speech production in terms of random behaviour [5].

2. EXPERIMENTAL METHOD AND SPEECH DATA

The speech data used for training and testing was derived from an in-house speaker-dependent 500 word ARM (Airborne Reconnaissance Mission) task [6]. This was chosen because a corpus of more than 15,000 results already existed with which any new results could be compared. Each speaker (2 male, 1 female) provided 36 ARM reports for training and 10 for testing, giving totals of 1991 and 541 words respectively.

The speech, sampled at 20 ks/sec, was analysed using a 400 point DFT. A Hamming window was used with a 50% overlap and the bottom 160 DFT coefficients were selected. This provided spectral estimates of the d.c. to 8 kHz band at a frame rate of 100 frames/second. The speech recognition system was based on 3-state monophone (context-independent) HMMs, plus four single-state, non-speech HMMs. Single component, continuous density, Gaussian states were used and a citation form pronunciation dictionary was employed for training, recognition and scoring.

All data was annotated orthographically, with timing markers provided at the boundaries of breath groups for training data only. The recognition system used was a conventional one-pass Viterbi decoder with beam

pruning and partial traceback [6, 7, 8]. Neither language model nor syntactic constraints were employed. The resulting recognised output strings were scored using a phone-mediated word alignment algorithm, with all word errors being reported.

3. SUB-BAND REPRESENTATIONS FOR SPEECH RECOGNITION

In order to investigate asynchrony between the upper and lower frequency bands of the speech spectrum, it is necessary to specify the frequency at which the spectrum should be split. Also, since the spectrum is typically cosine transformed prior to speech recognition, splitting into two sub-bands allows the cosine transform to be applied separately to each band. The resulting feature vector is the concatenation of the two vectors of sub-banded cosine coefficients. In view of the non-linear nature of speech with respect to frequency, this in itself may lead to performance improvements. It has been shown elsewhere [1] that this multi-band approach can also provide a degree of robustness to narrow-band interfering noise.

The composite feature vector $\mathbf{o}^c_t$ at time t is a K dimensional vector obtained from the cross-product of the vector of upper-band cosine coefficients $\mathbf{o}^u_t$ with the vector of lower-band cosine coefficients $\mathbf{o}^l_t$ as follows:

$$\mathbf{o}^c_t = \mathbf{o}^u_t \otimes \mathbf{o}^l_t$$

where the cross-product ⊗ is defined by:

$$o^c_{tk} = o^u_{tk},\ k = 1, ..., U$$
$$o^c_{t,U+k} = o^l_{tk},\ k = 1, ..., L$$

where U and L are the numbers of dimensions of the *upper-* and *lower-band* data respectively.

Sets of sub-band phoneme-level HMMs were constructed in which the underlying Markov model topology for a particular sub-band HMM was the same as that for the corresponding conventional HMM. A left-to-right topology with no state skipping was used.

4. ASYNCHRONY BETWEEN FREQUENCY BANDS

Asynchrony between the upper and lower frequency bands can be accommodated using either HMM decomposition [9, 10] or parallel model combination (PMC) [11] which is computationally more efficient. The latter approach was used at the phoneme-level in this study. The principle disadvantage of PMC applied in this way is that it only allows within-phoneme asynchrony, since model entry and exit is synchronised between the two streams. Apart from this PMC and HMM decomposition are functionally equivalent, since the combination function is simply vector concatenation. Previously this approach has been used successfully to accommodate asynchrony between the two components of an audio-visual speech recognition system [12].

Separate upper- and lower-band phoneme-level HMMs were constructed from the corresponding components of the synchronous sub-band HMMs. The underlying Markov model topologies were the same as for the corresponding sub-band HMMs, and for each state of the upper-band HMM, the mean vector and diagonal (co)variance matrix were set to be equal to the upper-band of the mean vector and covariance vector of the corresponding synchronous sub-band HMM. The lower-band HMMs were constructed analogously. The upper- and lower-band HMMs for each phoneme were then compiled into single PMC models [11]. More formally, the *asynchronous* model mean vector μ^p_{ir} and variance vector $\mathbf{d}^p_{ir}$ and the elements $a^p_{ir,\,js}$ of the state transition matrix $\mathbf{A}^p$ were computed from the relevant components of the *synchronous* model as follows:

$$\mu^p_{ir} = \mu^u_i \otimes \mu^l_r$$
$$\mathbf{d}^p_{ir} = \mathbf{d}^u_i \otimes \mathbf{d}^l_r$$
$$a^p_{ir,\,js} = a^u_{ij}\, a^l_{rs}$$

where ⊗ represents the cross-product as described earlier and state *ir* corresponds to state *i* of the *upper-band* component in the synchronous model and state *r* of the *lower-band* component.

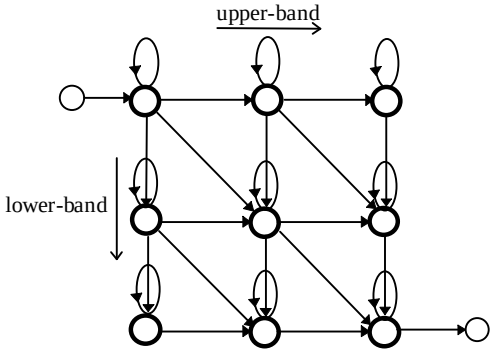

Figure 1: Asynchronous 9-state PMC model

The allowed state sequences of the asynchronous 9-state model are shown in Figure 1, where the horizontal and vertical time axes represents respectively transitions in the states of the upper- and lower-bands. Diagonal movement indicates simultaneous transitions in the two bands. Figures 2 and 3 illustrate the relaxation obtained

when an asynchronous HMM topology is employed to model a single phone.

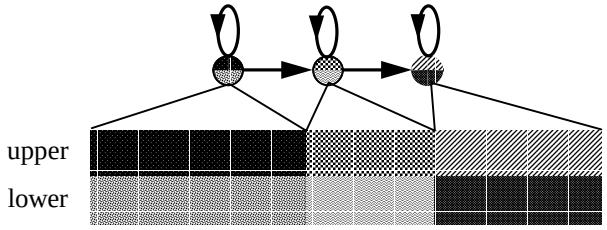

Figure 2: Conventional modelling using synchronous sub-band structure

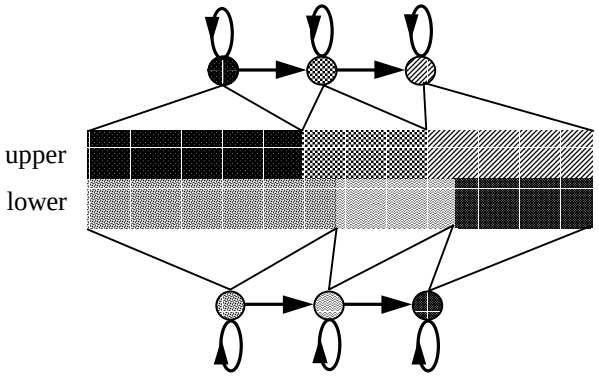

Figure 3: Decomposition using asynchronous PMC model

5. EXPERIMENTS AND RESULTS

A large set of experiments was conducted to investigate the effects of the choice of splitting frequency and the numbers of cosine coefficients used to represent the upper- and lower-frequency bands for *synchronous* sub-band HMMs. Reference results are provided for the single 8 kHz band and a feature vector of 30 cosine coefficients plus one energy term. Not surprisingly, the optimal frequency which separates the upper- and lower-bands varies between speakers. Experimental results, shown in Table 1, indicate an improvement in recognition accuracy for the two male speakers when the spectrum is split into two bands at 4 kHz. The system used 25 and 4 cosine coefficients plus two energy terms to describe the lower and upper bands respectively.

Increasing the split frequency improved the performance of the female speaker at the expense of the two males, however *all sub-band* results shown here are for a 4 kHz split. Experiments have also been conducted in which the frequency axis is partitioned into three bands, but as yet performance for clean speech is no better than that achieved with two bands.

The investigation of the asynchronous models was made more complex by the interaction of the PMC process with HMM parameter training, since it can be applied to the initial HMMs before sub-banding, the sub-band HMMs, the individual upper- and lower-band HMMs or to the combined PMC HMMs. In terms of synchrony, these correspond, respectively, to (i) synchronous training of the upper- and lower-band HMMs, (ii) asynchronous training of these models and (iii) a compromise between these two extremes.

All of these factors have been investigated through an extensive programme of experiments using a range of two-band representations and 9-state monophone HMMs. The best results from these experiments are between 2% and 4% better than the corresponding synchronous results in absolute terms. Again there was variation between speakers - due in part to the numbers of cosine coefficients representing each band and also due to the different training schemes employed. It was found to be detrimental to re-estimate the parameters of the upper- and, to a lesser extent, lower-band HMMs separately. This is because the upper-band HMMs do not contain sufficient information in themselves to guarantee a meaningful alignment of the HMMs and speech data during training. By contrast it was found to be beneficial to train the PMC HMMs, indicating that while the PMC HMMs allow some asynchrony, they still enforce sufficient synchrony to ensure a good alignment between the HMMs and training data during parameter re-estimation. The results in Table 1 indicate word error rate reductions for asynchronous models of 29% and 22% for the male speakers and 19% for the female, in relative terms, when compared to the results for the corresponding synchronous models. These improved results are for a 31 dimensional representation and models trained for 10 iterations from initial PMC creation.

Three band asynchronous PMC models have also been investigated, but again they have not performed as well as the two band PMC models to-date. It has already been noted that the asynchronous behaviour in the PMC approach is constrained by the requirement that the component streams are synchronised on model entry and exit (i.e. at phone boundaries). This can be overcome by extending the HMM topology to allow more model entry and exit states. However, benefits in terms of recognition performance have been small.

Data representation	Model configuration	Training	Word error rate, %		
			Male-M	Male-R	Female-S
Single band, 30 CC + 1 E	3-state	30 iterations from initialisation	11.1	16.3	12.8
Two band, 25+4 CC + 2 E	3-state synchronous	30 iterations from initialisation	8.3	15.3	13.5
Two band, 25+4 CC + 2 E	9-state asynchronous	10 iterations after PMC creation	5.9	12.0	10.9

Table 1: Speaker-dependent recognition results for a 500-word continuous-speech task

6. CONCLUSION

The use of two-band analysis of speech can provide an improvement in recognition performance. Further improvement are obtained by accommodating asynchronous behaviour in the speech signal using a technique which decomposes the speech into separate components. This single-signal decomposition approach provides a method which better models the underlying speech production process. More work is required to model the asynchrony across phoneme and word boundaries. In addition, a more principled strategy is required for providing optimal frequency-band partitioning for each speaker or set of speakers.

7. REFERENCES

[1] H Bourlard and S Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands", *Proceedings of ICSLP*, Philadelphia, October 1996.

[2] R K Moore, "Signal decomposition using Markov modelling techniques", *RSRE Memorandum 3931*, 1986.

[3] R K Moore, "Critique: The potential role of speech production models in automatic speech recognition", *JASA vol. 99*, March 1996.

[4] M J Russell, "Advances in speech recognition", *Proceedings of IoA Conf. on Speech and Hearing*, Windermere, November 1996.

[5] R K Moore, "Twenty things we still don't know about speech", *Proceedings of CRIM/FORWISS Workshop on 'Progress and Prospects of Speech Research Technology'*, Munich, September 1994.

[6] M J Russell, K M Ponting, S M Peeling, S R Browning, J S Bridle and R K Moore. "The ARM continuous speech recognition system", *Proceedings of IEEE ICASSP*, Albuquerque, April 1990.

[7] J S Bridle, M D Brown and R M Chamberlain, "A One-Pass Algorithm for Connected Word Recognition", *Proceedings of IEEE ICASSP*, Paris, 1982.

[8] K-F Lee, "Large Vocabulary Speaker Independent Continuous Speech Recognition: the SPHINX System", *PhD thesis*, Carnegie Mellon University, 1988.

[9] A P Varga and R K Moore, "Hidden Markov model decomposition of Speech and Noise", *Proceedings of IEEE ICASSP*, Albuquerque, April 1990.

[10] M Kadirkamanathan, "Hidden Markov Model Decomposition Recognition of Speech in Noise: A comprehensive experimental study", *Proceedings of ESCA Workshop on Speech Processing in Adverse Conditions*, Nice, November 1992.

[11] M J F Gales and S J Young, "HMM recognition in noise using Parallel Model Combination", *Proceedings of EUROSPEECH*, Berlin, September 1993.

[12] M J Tomlinson, M J Russell and N M Brooke, "Integrating audio and visual information to provide highly robust speech recognition", *Proceedings IEEE ICASSP*, Atlanta, May 1996.