Proceedings of the Institute of Acoustics

PROGRESS TOWARDS IMPROVED SPEECH MODELLING USING ASYNCHRONOUS SUB-BANDS AND FORMANT FREQUENCIES

N Wilkinson School of Electronic and Electrical Engineering, The University of Birmingham, Edgbaston, Birmingham B15 2TT.
M J Russell School of Electronic and Electrical Engineering, The University of Birmingham, Edgbaston, Birmingham B15 2TT.

1. INTRODUCTION

The basic problem in acoustic modelling for automatic speech recognition is that acoustic patterns which correspond to repetitions of the same utterance are not the same. Hidden Markov Models (HMMs) accommodate this variation by assuming that speech is generated by a statistical process. In most systems for recognition of non-tonal languages, the vocal tract configuration, and thus the spectral shape, is considered to be the major cue to phone identity. Hence the feature vectors which are presented to the speech recognition system are typically the results of applying some form of frequency analysis to the speech signal. Contextually induced variations are explicitly modelled by context-sensitive HMMs, which assume that the acoustic realization of a phone depends only on the identities of its immediate neighbours.

Speech pattern dynamics are produced by the motions of the articulators in the vocal tract. Conventional Hidden Markov Models assume that the speech signal is a product of articulators moving in perfect synchrony with one another . However, the articulators have some degree of independence and as such some scope for moving asynchronously. For example, consider the transition between a fricative and a vowel [6]. In this case sound is first generated by a constriction in the vocal tract followed by a "lighting-up" of the vocal-tract formant frequencies by the excitation from the vibrating vocal cords. The constriction generates turbulence with energy in the frequencies above 2.5Khz. The vowel sound is characterized by the strong presence of the first three formant frequencies which range from 300hz to 3Khz. It is plausible that from utterance to utterance the overlap between the onset and conclusion of the fricative can vary. Since the fricative and vowel occur in different halves of the spectrum this effect could be modeled by allowing some asynchrony between them.

Bourlard and Dupont [1] show that separate processing of sub-bands of the spectrum can improve recognition performance. This is particularly apparent when band limited noise is present. Processing the two halves of the spectrum independently has also resulted in improvements in the recognition accuracy of clean speech. In this case the performance increase may be attributed improved representation of the speech due to separate parameterisation of the frequency bands.

Tomlinson et al [4] also show that processing frequency bands separately improves performance and go on to show that by allowing asynchronously between the bands further improvement can be gained. This experiment was done on a 500 word Airborne Reconnaissance Mission Task with 2

male and 1 female speaker using 2 sub-bands. Asynchrony is incorporated into the system through a novel use of Parallel Model Combination (PMC) [10].

It is likely that this asynchrony would be better modeled in the articulatory domain. However, in order to achieve this an articulatory model, as well as a mapping between the acoustic and the articulatory domain, would be required. In the absence of such a model and mapping an intermediate formant representation could be used. The formant frequencies are a direct result of the articulator configuration and can be extracted automatically from the speech signal. Holmes, Holmes and Garner [8] have already shown that the incorporation of formant frequencies can improve recognition accuracy on a small vocabulary task, particularly if the incorporation of formant data includes a measure of confidence in the accuracy of the formant frequency estimates. These confidence measures are required, since in many regions of a speech signal the formants are typically either not present or very weak.

This paper presents results of experiments which investigate these issues using the TIMIT speech corpus. The first part of the paper investigates the benefits of separate signal processing of sub-bands, with and without asynchrony between the bands. The second part repeats the experiments from [7] and [8] on the TIMIT database, using the Holmes formant analyzer, which gives frequency estimates and confidence measures for the first three formants. The paper discusses the issues raised by incorporating confidence measures in the recognition and training algorithms, following Holmes, Holmes and Garner [7], and presents new results on TIMIT.

The longer-term goal of this work is to integrate the formants into an asynchronous speech recognition system. It is hoped that this will improve the modeling of speech production and consequently the recognition performance.

2. EXPERIMENTAL METHOD

2.1 Speech Data

All experiments use the TIMIT database, which comprises phone-labeled speech from speakers from across the USA covering 8 dialect regions. The database is partitioned into a training set (462 speakers) and test set (162 speakers), with mutually independent speakers in each set. The sentences are designed to present phones in as wide a range of contexts as possible. The speech was recorded at 20Khz and then down sampled to 16Khz.

2.2 Speech Analysis

Full-band coding of the speech spectrum is based on mel-frequency cepstral coefficients 0 to 12. To generate these a discrete Fourier transform is applied with a 30ms hamming window. The resulting complex spectrum is replaced by the log power spectrum, followed by mel scale filtering. This provides detailed information about spectrum structure at low frequencies and coarser information at higher frequencies. A cosine transform is then applied to generate the cepstral coefficients. The first- (delta) and second- (acceleration) time-differences are also calculated for each parameter, to give a 39 dimensional feature vector. The samples are taken every 10ms This is the 'MFCC_0_D_A' parameterisation from HTK [9].

2.3 Hidden Markov Models, Training and Recognition

A total of 634 context dependent triphone and biphone HMMs are used to represent the phones in the database. The models were generated by creating all possible triphone combinations then

clustering those which correspond to similar feature vectors. The 634 models give good coverage of the database and allow each model to have sufficient training data. Each model has 3 emitting states comprising 4 Gaussian mixture components. The number of Gaussian components, states and models are the same for every experiment presented in the paper.

The HMM parameters were initialized using the phone-level annotation provided with TIMIT, followed by Viterbi alignment to improve the state-time correspondence. The Baum-Welch algorithm was them applied at the sentence level. This allows the HMMs to refine their start and end times as well as their internal state alignment. All parameters are re-estimated using 13 iterations of the Baum-Welch algorithm, after which it was found that recognition improvement leveled off. Unless stated otherwise, this training procedure was used in all experiments.

The test set comprised 2 speakers from each dialect region, selected from the TIMIT test set. Each sentence was recognised at the phone level using Viterbi decoding to produce a phone sequence. The phones in the recogniser output and in the corresponding correct transcription from the TIMIT corpus were then mapped onto 39 phone equivalence classes, and the system's performance was measured. A statistical phone level-language model, generated from the training data, helped to improve performance.

3. EXPERIMENTS USING SUB-BAND HMMS

3.1 Sub-band and Asynchronous-band HMMs

Figure 1 shows an 'asynchronous-band' HMM, obtained using Parallel Model Combination [10], [4]. Such a model is created by splitting a synchronous HMM into two HMMs each describing a separate frequency band. The two HMMs are then recombined so that the two bands can run together but asynchronously. To make transitions in the upper frequency band one traverses the model from left to right. To make transitions in the lower frequency band one traverses the model from top to bottom. Being in state $S_{m,n}$ of the composite PMC model corresponds to being in state m of the lower band HMM and state n of the upper band HMM. Occupancy of the diagonal states of the asynchronous PMC HMM corresponds to occupancy of (synchronous) states of the original HMM. However, use of off-diagonal states in the PMC model corresponds to asynchrony between the lower and upper bands.

This type of model was used in recognition only in the experiment described in section 3.4, and in both recognition and training in the experiment described in section 3.5.

3.2 Baseline Experiment: Full-Band HMMS

A conventional synchronous full-band model is required as a baseline experiment. This was trained and tested as described in section 2. The resulting phone accuracy was 62.35%.

3.3 Separate Signal Processing of the Upper and Lower Frequency Bands

In this experiment the spectrum was split and the two halves processed separately. A number of split points and allocations of parameters to the upper and lower bands were tested and the optimum chosen. The 1st band ranged from 0 to 3308hz and the 2nd from 2599 to 8000hz. The lower band is modeled by 8 MFCCs plus the 0th MFCC and the upper band is modeled by 3 MFCCs plus the 0th MFCC. The overlap between the bands results from the triangular filters used to generate the mel-spectrum.

3.4 Synchronous Training and Asynchronous Testing

In this experiment, the spectrum was split as described above. After conventional, synchronous training, asynchrony was allowed between the upper and lower bands. This was achieved by splitting each synchronous HMM into two sub-band HMMs, one for each band, and then combining them using Parallel Model Combination, PMC [4]. The PMC model was then tested.

In synchronous training, the alignment between the HMM states and the data is a compromise between the optimal alignment for the upper and lower bands. It is likely that this has a detrimental effect on the upper band, since it has less than half the parameters of its neighbour.

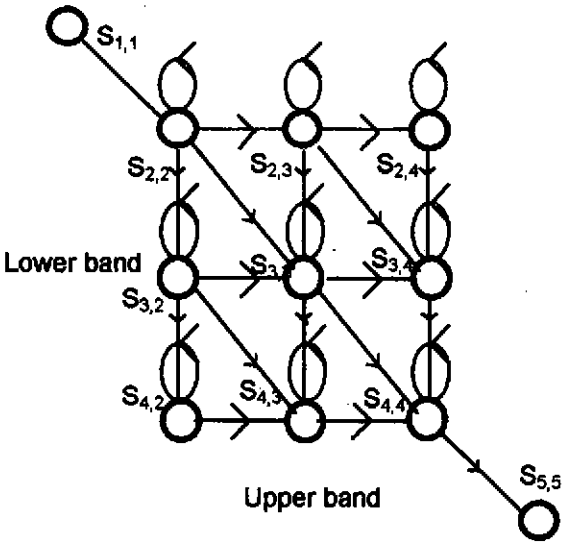

Figure 1: Composite HMM obtained by combining upper- and lower-band HMMs using Parallel Model Combination

3.5 Asynchronous Training and Testing

In this experiment the spectrum was split as described above. The models were trained as synchronous HMMs until the Baum Welch re-estimation stage. Each model was then split into upper and lower band models, combined into a single PMC model, and this was then trained using 13 iterations of the Baum-Welch algorithm.

3.6 Results and Discussion

The results of experiments 3.2 to 3.5 are shown in table 1. The baseline recognition performance is 62.35% phone accuracy. The application of separate signal processing to the upper and lower frequency bands leads to a small increase in recognition accuracy, although there are significant changes in both deletion errors (35% decrease) and insertion errors (30% increase) relative to the baseline system. This result is consistent with previous work ([1],[4]).

Training the sub-bands synchronously but allowing the sub-band HMMs to run asynchronously during testing results in a small drop in performance relative to the previous experiment, which is again consistent with [4]. The best performance (64.03% phone accuracy) is obtained when asynchrony is allowed in both training and recognition. Compared with the baseline system, this results in a 4% increase in the number of phones which are correctly recognised, a 27% decrease in the number of insertion errors, and a 17% increase in the number of deletion errors. Unfortunately it is not clear how much of this improvement is due to exploitation of the model's ability to allow asynchrony between the upper and lower frequency bands, and how much is simply due to the effective increase in the number of model parameters which results from asynchronous training (in the PMC model resulting from synchronous training the model parameters are tied, but this tying is broken by asynchronous training). The extent to which the potential for asynchronous processing is being exploited could be investigated by studying the patterns of state occupancy in the composite model.

Experiment	% Phone Accuracy	Number Correct	Number of Deletions	Number of Subs	Number of Insertions
Full band coding, sync training and testing (baseline)	62.35%	3879	685	1027	394
Sub-band coding,sync training and testing	62.90%	4030	446	1115	513
Sub-band coding, sync training, async testing	62.62%	3997	487	1107	496
Sub-band coding, async training and testing	64.03%	4043	499	1049	463

Table 1: Results of experiments 3.2 to 3.5

4. EXPERIMENTS USING FORMANT FREQUENCY DATA

The second set of experiments investigate the effect of replacing higher order MFCCs with formant frequencies. The experiments use the same formant tracker as in [7] and [8], which provides estimates of the first three formant frequencies plus confidence measures.

4.1 The Formant Analyser

The Holmes formant analyser generates formant frequency estimates by comparing the current 10ms speech spectrum with 129 manually chosen spectra, each with hand labeled formant frequencies. The best six spectra are selected and then warped to obtain the best fit with the current spectrum. Scores are awarded according to the degree of warping which is required. If a spectrum has been recently selected for comparison then this is also taken into account. The formant frequencies associated with the closest fitting spectrum are then adjusted according to the frequency warp and used as formant frequency estimates for the current spectrum. If the closest fitting spectrum has alternative formant frequency estimates, then these are also given.

The confidence measures are determined by comparing the amplitude of the closest spectral frequency to a particular formant to that of the highest spectral frequency in the current frame and to the highest long term spectral frequency. The curvature of the spectrum around the closest spectral frequency is also computed and contributes to the confidence measure. High confidence is given to formants with relatively high spectral frequencies and large curvature measurements. Thus the confidence measure is indicative of the "peakiness" of the spectrum, rather than the accuracy of the format frequency estimates. The confidence weights should be used to indicate how much attention should be given to the formant frequencies. It is reasonable to assume that when the confidence is *0* they should be ignored and should not contribute at all to the recognition process.

4.2 Recognition using Formant Data without Confidence

An initial experiment was conducted to determine the phone recognition performance when the formant frequencies were used without their associated confidence measures. For sounds where one would not expect formants to be present, such as fricatives, the formant estimates would be of little use. However for sounds such as these, one would expect the variances associated with the corresponding HMM states to be large, so that in principle the system would learn to ignore the fomant data for these classes of speech sound.

The formant frequencies are used as a replacement for the 10^{th}, $11^{th}$and 12^{th} MFCCs. Experiments were conducted to assess the performance of a system based on the first nine MFCCs only (i.e without the 10^{th}, $11^{th}$and 12^{th} MFCCs), and with with these last three MFCCs replaced by the formant frequency estimates. The systems were trained and tested in the manner described in section 2.

4.3 Recognition using Formant Data with Confidence

In these experiments, training was again conducted without the use of the confidence measures, but the confidence measure were included in the recognition phase. Two methods for including the confidence measures in decoding were considered. The first, from [7], regards confidence as a measure of uncertainty associated with the corresponding observed formant frequency. The estimated frequency is thought of as the mean of a distribution and the confidence is assumed to indicate the variance of that distribution. Calculation of the probability of a particular format frequency estimate, given a HMM state, then involves the convolution of the state and formant distributions which, because both are Gaussian, amounts to increasing the variance of the state distribution [7]. Thus the variance is increased as the confidence in a particular formant value decreases. Contrary to expectation [7] this method did not give good results in the present experiments. However, from a theoretical stance, the assumption that confidence is inversely related to the 'observation variance' may not be appropriate, since the confidence measure is

really an estimate of how likely a formant is of being observed and not the accuracy of the frequency. With this in mind, a second method for including the confidence measures in the probability calculation was considered whereby the conventional state-conditional probability of the formant value is raised to the power of the confidence. More precisely, if $b(o)$ denotes the state output probability of a particular formant estimate o with associated confidence measure cw, then:

$$b(o) = k * cw * \log(\mathrm{N}(o; f, v))$$

where $N(\,; f, v)$ denotes a Gaussian density with expected formant frequency f and variance v, and k is a constant, refered to as the 'confidence weight scale factor'. This use of the confidence is heuristic, but it has some desirable properties. The higher the confidence value the more note is taken of formant frequency, and if $cw=0$, then the formant data is ignored.

4.4 Recognition and Training using Formant Data with Confidence

Analysis of the distribution of the confidence measure over the training data shows that half of the confidences are below 0.5. However the confidence measure is interpreted, it is likely that the low confidence formant frequency estimates add little useful information to the estimates of the corresponding parameters in the HMMs. Indeed they could be viewed as noise and thus could have a detrimental effect on the HMM state formant frequency parameter estimates. The next experiment involved incorporating confidence measures into the training procedure. For this experiment the HMMs were initialized and the states aligned without using confidence. However at this point a modified Baum-Welch algorithm was used which adjusted the contribution of the formant frequency estimate to the re-estimate of the HMM state formant parameters according to the confidence measures. Many consonants have low or zero confidence associated with their formant frequency estimates, and this implicitly results in an reduction of the size of the training set, which in turn can result in over training. Hence the HMMs were tested after every iteration of the Baum -Welch algorithm. Testing was done with the confidence measures taken into account.

4.5 Results and Discussion

Phone accuracy results for the experiments in section 4 are shown in table 2.

In addition to measuring percentage phone accuracy, confusion matrices were generated for each experiment so that any patterns in the errors could be observed. These matrices showed overall phone confusion, and confusion with and between different phone types. The confusion matrices are summarised in table 3, which shows the percentage correct phone-type classifications, and percentage correct classifications within each phone type.

Table 2 shows that, predictably, reducing the number of MFCCs from 12 to 9 results in a fall in phone accuracy. Less predictable, but consistent with [7], is the fact that addition of the three formant frequencies without confidence leads to a further fall in phone accuracy. Rows 4, 5 and 6 of table 2 show that incorporation of confidence into the recognition process results in an increase in phone accuracy, but that the precise value of the confidence weight scale factor does not appear to be critical. The biggest improvement results from inclusion of confidence in the Baum-Welch algorithm, but the performance still falls short of that obtained with the original 12 MFCCs.

Table 3 gives more insight into the consequences of including formant data in the parameterisation of the speech signal. It is clear that the most significant effects are on vowel recognition, where the inclusion of formants plus confidence in recognition and training raises the percentage of vowels which are recognised correctly from the 'baseline' figure of 73.15% to

76.33%. For the 9 MFCC based parameterisation, the inclusion of formant data plus confidence in training and recognition results in a 6.4% improvement in the percentage of vowels which are correctly classified. Overall, however, the improvements in vowel recognition are eroded by poorer performance for fricatives, affricates and glides. This suggests that in order to obtain a general improvement in performance through the inclusion of formant data, the priority is to prevent this data from influencing the classification process in cases where it is not relevant

Experiment	% Phone accuracy
12 MFCCs (baseline experiment)	62.35%
9 MFCCs	60.42%
9 MFCCs + 3 Formant frequencies	60.15%
9 MFCCs + 3 Formant frequencies * 0.5 * confidence weights	60.92%
9 MFCCs + 3 Formant frequencies * 1.0 * confidence weights	60.70%
9 MFCCs + 3 Formant frequencies * 1.5 * confidence weights	60.54%
9 MFCCs + 3 Formant frequencies * 1.0 * confidence weights 7 iteration of modified Baum-Welch algorithm	61.55%
9 MFCCs + 3 Formant frequencies * 1.0 * confidence weights 15 iterations of modified Baum-Welch algorithm.	61.22%

Table 2: Results of the experiments described in section 4, using formant frequency data

Experiment	Phone Type	Vowel	Nasals	Fric.	Affric.	Glides	Stops
12 MFCCs	91.95	73.15	84.32	89.56	93.88	97.27	90.80
9 MFCCs	91.78	71.71	83.51	87.27	94.00	95.94	90.76
9 MFCCs + 3 ff	91.26	74.64	83.07	86.67	93.02	95.80	89.01
9 MFCCs+3 ff *1* cw	91.28	75.34	83.25	87.14	93.18	96.64	89.93
9 MFCCs+3 ff *1*cw 7 * mB-W algorithm	91.35	76.33	84.31	87.44	89.36	96.94	90.47

Table 3: Percentage correct recognition of phone types, and within phone types

5. CONCLUSIONS AND FURTHER WORK

The results of these experiments confirm that separate signal processing of spectrum sub-bands improves recognition performance. Allowing asynchrony in recognition but not in training leads to a very small increase in performance compared with the baseline system, but poorer performance than synchronous recognition with separate processing of sub-bands . Introducing asynchrony into training improves performance but at the cost of using extra parameters in the HMMs. Thus, the results obtained here on TIMIT are consistent with those presented in [4].

Using formant frequencies without confidence leads to an drop in recognition accuracy. The best result obtained on TIMIT using formants is 61.55% accuracy, using confidence in training and recognition. This is worse than the performance of the baseline system and contrary to the findings of [7] where a 0.9% reduction in error rate is reported, and [8] where a 1.4% reduction in error rate is obtained. However, the current experiments have been conducted on a different corpus, TIMIT, and have involved both male and female speakers. These factors may lead to poorer performance by the formant tracker. An experiment that is closer to that reported in [7] is being run, in which only the male speakers in the TIMIT corpus are used, the speech is band limited to 4Khz and the baseline includes only 8MFCCs.

The inclusion of formant data results in an improvement in vowel recognition, but worse performance for other classes of phone. The fact that consonant classification is worse once formant data is used suggest that the method of exploiting the confidence measures is not yet sufficiently powerful to enable the formant data to be ignored when it is irrelevant. This is an obvious topic for future work.

Eventually is hoped that the formants can be allowed to run asynchronously in a PMC model in an attempt to better model speech production.

ACKNOWLEDGEMENT

The first author, Nick Wilkinson, is jointly sponsored by an EPSRC Project Studentship (Grant GR/M87146 "An integrated multiple-level statistical model for speech pattern processing") and DERA.

REFERENCES

[1] H. Bourlard and S. Dupont. ASR based on independent processing and recombination of partial frequency bands. *Proceedings of ICSLP'96, International Conference on Spoken Language Processing, Philadelphia, October 1996.*

[2] H. Bourlard, S. Dupont, H. Hermansky and N. Morgan. Towards sub-band based speech recognition. *Proceedings of European Signal Processing Conference, Trieste, Italy, pages 1579-1582, September 1996.*

[3] H. Bourlard and S. Dupont. Sub-band based speech recognition. *In Proc. IEEE ICASSP'97, Munich. 1251-1254, 1997*

[4] M.J. Tomlinson, M.J. Russell, R.K. Moore, A. P. Buckland and M. A. Fawley. Modeling asynchrony in speech using elementary single-signal decomposition. *", Proc. ICASSP'97, Munich, 21st-24th April, 1997.*

[5] M. J. Tomlinson, M.J. Russell and N. M. Brooke. Integrating audio and visual information to provide highly robust speech recognition, *Proceedings of EUROSPEECH, Berlin, September 1993.*

[6] S. Fernandez, S. Feijoo, R.Balsa and H. Barros. Perceptual effects of co-articulation in fricatives. *In Proc. IEEE ICASSP'2000, Istanbul, Turkey, vol 3, pages 1347-1350, 2000*

[7] J.N. Holmes, W.J. Holmes and P.N. Garner. Using formant frequencies in speech recognition. *In Proceedings EUROSPEECH '97, volume 4, pages 2083 - 2086, September 1997.*

[8] W.J. Holmes and P.N. Garner. On the robust incorporation of formant features into hidden Markov models for automatic speech recognition. *In Proc. IEEE ICASSP'98, Seattle, pp. 1-4, 1998*

[9] S. Young, J. Odell, D. Ollason, V. Valtchev and P. Woodland. The HTK Book. *Entropic Cambridge Reseach Laboratory.*

[10] M.J.F Gales and S.J. Young, Robust Speech Recognition in Additive and Convolutional Noise using Parallel Model Combination. *Computer Speech and Language, Vol 9, pp289-308, 1995*