A Mixture of Gaussians Front End for Speech Recognition

M.N. Stuttle and M.J.F. Gales

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, United Kingdom
{mns25, mjfg}@eng.cam.ac.uk

Abstract

This paper describes a feature extraction technique based on fitting a Gaussian mixture model (GMM) to the speech spectral envelope. The features obtained (the component means, variances and priors) represent both the the general shape of the spectrum and provide information on the position of the spectral peaks. As the features select peaks in the spectrum they are related to the formant amplitudes, locations and bandwidths. Results using the Resource Management corpus, a medium vocabulary task are presented. Although by themselves the GMM features do not outperform MFCC features, systems combining the GMM systems with a standard frontend are shown to give a reduction in word error rate.

1. Introduction

Formants are considered to be representative of the underlying phonetic content of speech and potentially useful for speech recognition applications, particularly in noisy or band-limited environments [1]. A number of techniques for extracting formants from speech data have been investigated, including analysis by synthesis [2], linear prediction analysis incorporating using N-best lists [3], and dynamic template matching of hand-labelled spectra [4]. However, there are a number of problems associated with extracting formants from speech data. First, formants are not always well defined in the spectrum, especially in the case of fricatives of nasalised sounds, and may lie between harmonic peaks. Second, the standard LPC formant analysis extraction scheme does not yield the amplitude information which is required to distinguish certain phone types, for instance between nasalised phones and voiced vowels, and formants positions alone cannot describe the general shape of the spectra.

Zolfaghari and Robinson [5] proposed a technique applying the Expectation-Maximisation (EM) algorithm [6] to fit a set of Gaussian mixtures to the smoothed magnitude spectra of a speech signal. The target application was a low bit rate vocoder. A set of means, variances and mixture weights are extracted. From these parameters it is possible to reconstruct the envelop of the spectral envelope. This scheme is not a formant detector in terms of looking for in resonances in the speech signal. However it does find the spectral peaks characteristic of formants, and hence the features are 'formant-like'. The information extracted, the component means, weights and standard deviations can be related to the formant locations, amplitudes and bandwidths of the formants respectively. Consequently it has some attributes in common with the work done by Padmanabhan [7] which used a series of band-pass filters to locate and detect the amplitudes of spectral peaks and found that they provided incremental information to the MFCC features. The proposed scheme is a general spectral estimation method. Hence, it represents the shape of the spectrum in addition to the locations of spectral peaks. It is able to represent speech sounds such as nasals and fricatives that formant frequencies would not typically be able to discriminate between. This work examines the use of this Gaussian mixture model front end for speech recognition.

As the proposed feature extraction system works directly in the spectral domain there are a number of advantages for both speaker adaption and noise compensation schemes. One currently popular scheme for speaker adaptation is vocal tract length normalisation (VTLN) [8], which warps the frequency spectrum to reduce inter-speaker variability. This would be easily achieved with the GMM frontend. A simple shifting and scaling of the component means could approximate a VTLN transform. This is considerably easier than the nonlinear process required for mel-cepstral features. Furthermore, it would be simple to apply an additive noise model since the mixture components are additive in the spectral domain.

This paper explores the application of a GMM frontend to continuous speech recognition and the paper is organised as follows. The mixture of Gaussians feature extraction technique is outlined, in the next section. Experimental results on the Resource Management task using the mixture of Gaussians features and standard Mel-cepstral and PLP features separately and in combination using streaming systems are then presented and reviewed.

2. Fitting Gaussian Mixtures To Speech Spectra

2.1. Technique for fitting

The general method for fitting a Gaussian mixture to a speech signal is as follows. The speech is windowed into separate frames and a DFT magnitude spectrum is calculated for each. To prevent the EM algorithm from fitting Gaussians to pitch peaks, the envelope of the spectrum is smoothed prior to its application. When the smoothed spectrum has been obtained, a histogram is estimated from it using EM to maximise the log likelihood of the data, as in eq 1, for a histogram $\{x_1, \ldots, x_k\}$ using M mixtures with model parameters θ.

$$l(\theta) = \sum_{k=1}^{n} \ln \Big[\sum_{m=1}^{M} p(x_k|m)c_m \Big] \qquad (1)$$

The approximation made in this work is that each point in the DFT translates to a rectangular bin centred at the DFT point. Note that for a rectangular bin from $x_0 + \frac{1}{2}$ to $x_0 + \frac{1}{2}$:

With many thanks to Dr. A.J. Robinson for help during the initial formulation of ideas. M.N. Stuttle is funded by an ESPRC studentship. He received additional support from the Newton Trust and SoftSound.

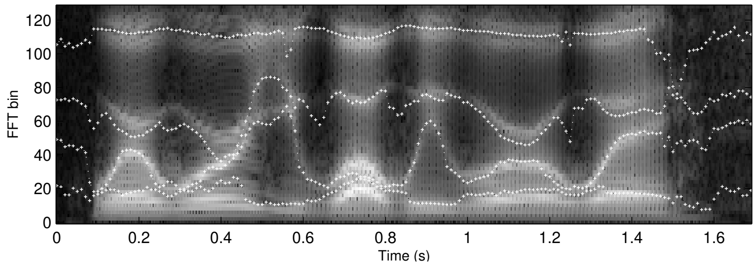

Figure 1: Gaussian mixture component mean positions for the utterance "Where were you while we were away?", with four Gaussian components fitted to each frame.

$$\mathcal{E}(x^2) = \int_{x_0-\frac{1}{2}}^{x_0+\frac{1}{2}} x^2 dx = {x_0}^2 + \frac{1}{12} \quad (2)$$

thus, each data point has an effective variance of $\frac{1}{12}$ resulting from the histogram. This must be included when calculating likelihoods and accumulating statistics. The Gaussian mixtures are initialised uniformly across the histogram with equal component weights. Multiple iterations of the EM algorithm are then perform in order to fit the mixture of Gaussians to a the histogram. An example of the fit of the means to a spectrogram is shown in figure 1 for the all-voiced sentence "*Where were you while we were away*". The GMM component means track the spectral peaks. Note that since no frame to frame information is propagated, several noisy points can be observed.

2.2. Issues in fitting

Normally, speech is sampled at 16kHz, giving a 8kHz spectrum. However, in this work, the the spectrum was low-pass band limited to 4kHz before the mixtures were fitted, as this yielded more consistent parameters within phone classes and hence lower error rates.

The choice of smoothing algorithm used to remove the pitch is also important. Work by Zolfaghari and Robinson [9] achieved the best perceptual results from a GMM vocoder by the using a spectral estimate envelop vocoder (SEEVOC) window which interpolates between the pitch peaks. However, initial results obtained on the Resource Management (RM) task showed that improved performance was obtained by convolving the magnitude spectrum with a raised cosine window centred on the fundamental frequency.

Psychoacoustically motivated methods of treating the spectra prior to fitting Gaussians were also explored. These included applying pre-emphasis, equal loudness auditory transforms and Mel-scale warping of the frequency spectra. However, all of these approaches increased the word error rates on the RM task and were not implemented in the final systems.

2.3. Applying smoothness constraints

As can be observed in figure 1 some of the component trajectories are not smooth. Since the EM algorithm is an iterative parameter estimation technique the features it extracts can be noisy. In order to introduce an element of temporal smoothness in the features the inter-frame information could be applied before or after the fits were made.

To introduce smoothing at the level of the EM fit, several spectral frames were taken on either side of the input frame and appended to the central frame. This yields a two dimensional histogram. A two dimensional mixture of Gaussians was then fitted to this histogram using the EM algorithm. The positions of the Gaussian means were constrained to be on the input time frame.

A simpler smoothness constraint was also investigated using a moving-average filter directly on the input parameters after the algorithm has been applied.

3. Experimental results

All experiments were performed on the Resource Management (RM) task. This consists of 3990 training sentences with roughly a 1000 word vocabulary. There are 109 speakers training speakers in corpus and 1200 test sentences from 40 subjects. The data is sampled at 16kHz. All recognition results were formed using hidden Markov models (HMMs) and using the HMM Toolkit (HTK) RM recipe [10].

Cross-word triphone context-dependent HMMs were made using a phonetic decision class tree. The number of components in the state output probability density functions (PDFs) was increased until no further recognition improvements were observed on a subset of the in the test data. A word-pair grammar was used for recognition; the grammar scale factor was tuned on the 300 sentence 'feb89' subset of data.

3.1. Feature Extraction

The basic generation of Gaussian Mixture Model (GMM) features from speech data is outlined below. Sample windows 25ms long were taken every 10ms, and a Hamming window was applied. A 512-point FFT was used to obtain the magnitude spectrum, which was convolved with a pitch-dependent filter. The EM algorithm was initialised uniformly over the data and iterated twelve times. The parameters extracted were the mel-scaled locations of the Gaussian means, the standard deviations and the logarithms of the component weights.

Two main GMM systems were built and are used in the following experiments. The first used six mixture components to model the spectrum (GMM6). The second used four (GMM4). Mel-frequency cepstral coefficients were used as a baseline performance measure. Perceptual Linear Prediction [11] was also used as it is a popular alternative to MFCCs. All static features were appended with log spectral energy terms and dynamic and acceleration coefficients were taken.

3.2. Baseline results

Initial experiments were run on the RM task to compare the new set of GMM features extracted from speech with conventional mel-cepstral features and PLP features. Six Gaussian mixture components were used for the MFCC, PLP and GMM6 systems, and eight mixture components were used in the GMM4 system. Results are presented in table 1.

Description	Total Features	% Err.
MFCC	39	4.12
PLP	39	3.89
GMM4 (6 o/p mixes)	39	6.60
GMM4 (8 o/p mixes)	39	6.10
GMM6	57	5.36

Table 1: Word error rates for the RM task comparing baseline systems made with Gaussian mixture models, PLP coefficients and MFCC feature vectors

Although the best result obtained for a system using the GMM features alone was 29% worse than the MFCC baseline, this result compares favourably with other work using formant features alone [2]. The PLP features outperformed the MFCC result on this task. The GMM4 system gave poorer results than the GMM6 system. This seems reasonable since the GMM6 system can describe more detail in the spectrum and provides more consistent plots.

Examining the results on a phone class basis, the GMM system recognition scores were consistently down from the MFCC scores. No phone classes were being represented significantly better or worse by the GMM features. A global full covariance matrix was constructed for the features. From this, it was observed that there are high degrees of correlation between the features for the GMM systems, most notably between the log component energy terms. However, a system built using a full covariance matrix gave poor results. This is probably due to the size of the RM training task and a larger task would yield a full covariance system which could generalise well.

3.3. Mean Position Ratios

Work in phonetics suggests that the ratios of formant positions may be informative in distinguishing phone types . To investigate this, the GMM systems were appended with the ratios $\mu_1 : \mu_2$, $\mu_1 : \mu_3$ and $\mu_2 : \mu_3$ and systems were built and tested as before. The inclusion of mean ratios into the GMM4 system reduced error rates to 5.57%, a relative reduction of 10% in WER. Appending the mean ratios to the GMM6 system increased the WER to 5.87%.

3.4. Temporal Smoothing

As described in section 2.3, temporal feature smoothing schemes were applied. Results using the 2D EM fit are shown in table 2. From the 2D Gaussian parameters, the only covariance term coded was that in the frequency direction, so the same number of parameters are used for each frame as before. A moving average (MA) filter was also applied to the 1D feature vectors after the EM fits and was used as a comparative result. All systems use six Gaussian mixture components in the output PDF.

The only form of temporal smoothing which yielded any

Description	% Err.
2D GMM4, 3 frames	7.06
2D GMM4, 5 frames	8.62
2D GMM4, 7 frames	10.10
2D GMM4, 5 frames, windowed	6.72
GMM4 MA, 3 frames	6.13
GMM4 MA, 3 frames	7.14
GMM6 MA, 3 frames	6.80

Table 2: RM word error rates for different temporal smoothing arrangements on the GMM4 and GMM6 system

improvements was the MA scheme with a filter length 3 applied to the GMM4 system. When a longer MA filter was used, the features are excessively smoothed and discriminative detail was lost. Applying a MA filter to the GMM6 system degraded the performance.

Frontends using the 2D fits failed to reduce the error rates. It appears that taking fits over the extra frames smears the spectral information across a larger time period and makes the EM fits more generalised. To alleviate this, a triangular window was applied to the data such that the outermost spectral frames were deweighted. This yielded a performance similar to that of the 1D system.

3.5. Concatenated Feature Vectors

In common with other formant-like front ends, the performance of the formant information alone performed worse than a MFCC based system. However, formant information is considered to be complementary to MFCC features [4]. The standard MFCC feature vector was then augmented with the four mel-scaled mean positions from a four mixture GMM spectral fit. For a comparison, the experiments were also run with a 16 mel-cepstral representation and system using 12 MFCCs combined with the first four PLP coefficients. The MFCC plus GMM means system uses seven Gaussian mixture components whilst the 16 MFCC system and the 12 MFCC plus 4 PLP systems performed optimally with six.

Description	Total Features	% Err
16 MFCCs	51	4.29
12 MFCCs + 4 GMM means	51	4.00
12 MFCCs + 4 PLP	51	4.52
8 MFCCs + 4 GMM means	59	4.14
8 MFCCs	29	4.34

Table 3: RM Results for MFCC feature vectors augmented with additional information

The results for the concatenated feature vectors are presented in table 3. Appending the GMM means gave a relative performance increase of 4.0%, whilst all other features appended worsened the results, indicating that there is some additional information complimentary to MFCCs in the GMM features. Adding PLPs or MFCCs to the system was uninformative. Substituting the last four MFCCs for GMM means gives a similar result to the MFCC baseline with a similar number of features, and is an improvement over using only 8 MFCCs.

3.6. Multiple Streams

As an alternative way to combine two information sources, multiple stream systems were investigated. For a multiple stream system the output probability distribution $b_j(\boldsymbol{y})$ for input vector $\boldsymbol{y}$ divided into S streams $\{\boldsymbol{y}_1, \ldots \boldsymbol{y}_S\}$ is calculated as

$$b_j(\boldsymbol{y}) = \prod_{s=1}^{S} \Big[\sum_{m=1}^{M} c_{jsm} \mathcal{N}(\boldsymbol{y}_s; \boldsymbol{\mu}_{jsm}, \boldsymbol{\Sigma}_{jsm} \Big]^{\gamma_s} \tag{3}$$

Where $\boldsymbol{\gamma}_s$ is the stream weight and c_{jsm}, $\boldsymbol{\mu}_{jsm}$ and $\boldsymbol{\Sigma}_{jsm}$ are the component weight, mean and variance for component m of stream s. For all experiments quoted, stream weights were constrained to sum to one.

Two-stream synchronous stream models, the standard HTK concept [10] were built and tested using a MFCC stream and an additional stream of feature vectors. Three additional parameterisations of the speech were considered: a GMM4 system; GMM6 system and a PLP system for comparison purposes. The systems were built with the second stream (non-MFCC) weight set to zero. Hence the systems are built using only the MFCC features in the Baum-Welch algorithm. Hence, all three systems are built upon identical decision trees. All systems were trained to have six mixture components in the state output PDFs. Testing was performed by varying the MFCC stream weights from zero to one in the trained systems and running recognition experiments.

The results of these experiments are shown in figure 2. Note that the streaming result of MFCC+PLP with a MFCC weight of zero gives slightly worse results than that of the baseline PLP system. This is because the streaming system uses the MFCC alignments to train the observation probability density functions and transition matrices, which are not the ML solution for the other streamed features.

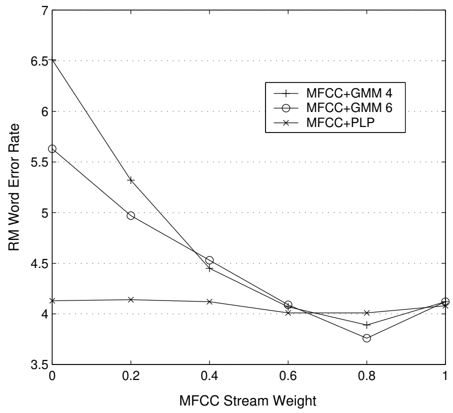

Figure 2: RM Results for a two-stream system using MFCC features and either PLPs, or a four or six component GMM system for the second stream,

The best performance of any system was gained by streaming the MFCC system with the GMM6 system with stream weights 0.8 and 0.2 respectively. This gave a word error rate of 3.76%, or a relative improvement of 8.7% over the baseline MFCC performance. The GMM4 with an MFCC stream weight of 0.8 system gave a performance of 3.89%, for a relative drop of 5.5%. Streaming MFCC and PLP features gave an error rate of 4.00% or 2.9% relative to the MFCC baseline, and this result suggests that there is a great deal of mutual information between the two parameterisations. The greater reduction in the word error rate when streaming with GMM features supports the hypothesis that these systems contain incremental information not present in the MFCC features.

4. Conclusions and Future Work

These experiments show that fitting a GMM to a speech spectrum is able to provide features with information complementary to MFCCs, and give better results than MFCCs alone on the RM task. The next task is to apply the GMM system to a larger vocabulary corpus such as the Wall Street Journal task.

Future work will examine the use of the GMM for rapid speaker adaption transforms, since a simple vocal tract length normalisation approach could be approximated by a linear shift in the positions of the means in the GMM. The use of an additive noise models within the GMM framework is also to be investigated. In addition, the use of linear and non-linear transforms is to be explored to remove the correlations observed in the features.

5. References

[1] M.J. Hunt. ”Delayed Decisions in Speech Recognition - the Case for Formants”. *Pattern Recognition Letters*, 6:121–137, 1987.

[2] L. Welling and H. Ney. ”Formant Estimation For Speech Recognition”. *IEEE Transactions on Speech and Audio Processing*, 6(1):36–48, 1998.

[3] P. Schmid and E. Barnard. ”Explicit, N-Best Formant Tracking”. In *Proceedings ICASSP 97*, 1997.

[4] W.J. Holmes and P.N. Garner. ”On The Robust Incorporation of Formant Frequencies into Hidden Markov Models for Automatic Speech Recognition”. In *IEEE Proceedings ICASSP*, 1998.

[5] P. Zolfaghari and T. Robinson. ”Formant Analysis Using Mixtures of Gaussians”. In *Proceedings ICSLP*, 1996.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin. ”Maximum Likelihood From Incomplete Data Via The EM Algorithm”. *Journal of Royal Statistical Society Series B*, 39:1–38, 1977.

[7] M. Padmanabhan. ”Spectral Peak Tracking and its Use In Speech Recognition”. In *Proceedings ICSLP*, 2000.

[8] L Lee and R C Rose. Speaker normalisation using efficient frequency warping procedures. In *Proceedings ICASSP*, volume 1, pages 353–356, 1996.

[9] P. Zolfaghari and T. Robinson. ”A Formant Vocoder Based on Mixtures of Gaussians”. In *Proceedings ICASSP*, 1997.

[10] S. Young, D. Kershaw, J.J. Odell, D. Ollason, V. Valtchev, and P.C. Woodland. *”The HTK Book, Version 2.2”*. Entropic, 1995.

[11] H. Hermansky. ”Perceptual Linear Prediction (PLP) of Speech”. *Journal of the Acoustic Society of America*, 87(4):1738–1752, 1990.