COMBINING A GAUSSIAN MIXTURE MODEL FRONT END WITH MFCC PARAMETERS

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

Fitting a Gaussian mixture model (GMM) to the smoothed speech spectrum allows an alternative set of features to be extracted from the speech signal. These features have been shown to possess information complementary to the standard MFCC parameterisation. This paper further investigates the use of these GMM features in combination with MFCCs. The extraction and use of a confidence metric to combine GMM features with MFCCs is described. Results using the confidence metric on the WSJ task are presented. In addition, the performance of the GMM features in noise is explored aSnd robustness and compensation techniques are investigated. The system performance is examined on the RM task with additive noise.

1. INTRODUCTION

It is well known that formant frequencies represent the underlying phonetic content of speech and are considered potentially useful in speech recognition. However, there are a number of issues with extracting formants from speech data. Formants cannot always be extracted robustly, especially in the case of fricatives or nasalised sounds. Additionally, formant frequencies alone do not represent the spectral shape and thus cannot distinguish between certain phone types, for instance nasalised phones and voiced vowels. Rather than looking for resonances in the signal, an alternative approach is to model the spectral peaks. Energy gravity centroids [1] are one approach to modelling spectral peaks. However, the locations of the centroids are constrained by the choice of frequency bands. Another approach is the HMM2 system [2]. The HMM2 system uses a second (frequency) HMM to model the speech spectrum. The state transitions in the second (spectral) layer have been shown to roughly follow the peaks in the spectrum. Both of these representations of spectral peaks have been demonstrated to give slight improvements in recognition performance on small tasks in noisy environments when combined with MFCC systems [1] [2].

The approach to representing spectral peaks adopted in this paper is to fit a Gaussian mixture model (GMM) to the smoothed spectrum. This was originally used for speech coding [3] and later to extract features for speech recognition [4]. The GMM features consist of a set of means, variances and component energies. These means, variances and weights are related to the locations, bandwidths and amplitudes of the formants in the signal. The performance of the GMM features for speech recognition is described in [4]. Alone, the GMM features perform significantly worse than standard MFCC parameters on the Resource Management (RM) task. However, combining GMM means with MFCCs by concatenation into a single feature vector improved recognition performance. Combining GMM and MFCC features using a synchronous streaming system also increased performance. These results suggest that the GMM features can provide information complementary to MFCCs.

As previously mentioned, some parts of speech are not well modelled by spectral peak representations or formant frequencies. Hence, it has been proposed to use a measure of confidence when combining MFCCs with formant features [5]. The confidence measure deweights these features in regions where the formants are not strongly defined. A similar measure derived from the GMM features is presented in this paper together with results on a large vocabulary task. A technique for noise compensation during the GMM fitting stage is presented, as well as a method to adapt the clean models to a noisy environment. Results using both noise compensation schemes are presented for a medium vocabulary task.

2. USING CONFIDENCE METRICS FOR FEATURE COMBINATION

This section briefly describes how a GMM may be fitted to a speech spectrum. A more detailed description of the extraction and use of GMM features for speech recognition can be found in [4].

The first step in fitting a GMM to a spectrum is to window the speech into frames 25.6ms wide at a rate of 10ms. A DFT magnitude spectrum is calculated for each window and then smoothed by convolving it with a pitch-based raised cosine window. A histogram $\{x_1(t), \dots, x_K(t)\}$ is obtained from the smoothed DFT by approximating each point in the DFT by a rectangular bin. The EM algorithm is used to estimate in an ML fashion a N component Gaussian mixture model for the smoothed values. The EM algorithm optimises the log likelihood of the histogram at time t with respect to the model parameters $\boldsymbol{\theta}(t)$.

$$l(\boldsymbol{x}(t)|\boldsymbol{\theta}(t)) = \sum_{k=1}^{K} \left(\ln \sum_{n=1}^{N} e_n(t) \mathcal{N}\left(x_k; \mu_n(t), \sigma_n^2(t)\right) \right) \quad (1)$$

where the parameters $\boldsymbol{\theta}(t)$ comprise

$$\boldsymbol{\theta}(t) = \{\mu_1(t), \dots, \mu_N(t), \sigma_1^2(t), \dots, \sigma_N^2(t), e_1(t), \dots, e_N(t)\}$$

A feature vector is formed by taking the means, the standard deviations and the log normalised component energies $\tilde{e}_n(t)$ for each component. The log-normalised energy is generated in a similar fashion to the terms used in HTK [7]. A silence floor set at 50dB

M.N. Stuttle is funded by an ESPRC studentship. He received additional support from the Newton Trust and SoftSound.

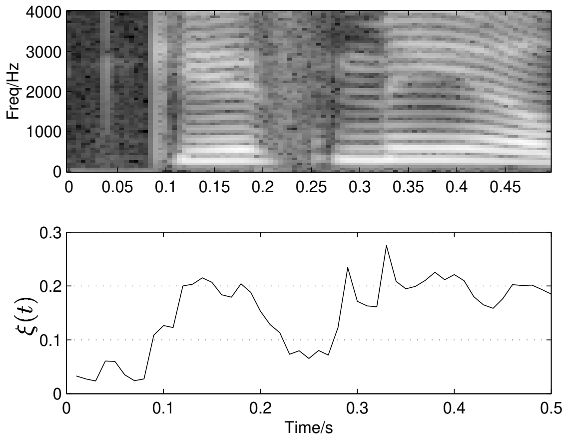

Fig. 1. Confidence metric plot for a test utterance fragment, with $\beta = 0.3$

is applied to the component energies. Thus, $\tilde{e}_n(t)$ lies in the range 1 to -10.53.

In previous work [4] it was found that optimal recognition performance on the RM task was obtained when the speech was band-limited to 4kHz and six Gaussians were fitted to each frame (N=6). Also, no frame to frame constraints were found to improve the system. Therefore, all GMM systems presented in this paper use this parameterisation.

Peak representations of the spectra may be less reliable for certain classes of phone. A measure of confidence in the formant frequencies when incorporating them with MFCCs was proposed in [5]. This confidence metric was based on the amplitude and degree of curvature of each formant or spectral peak. For a component in the GMM, these can be represented by the log normalised component energies $\tilde{e}_n(t)$ and the reciprocal of the the standard deviation $[1/\sigma_n(t)]$. For frame t, the confidence metric $\xi(t)$ is defined as

$$\xi(t) = \beta \left[\prod_{n=1}^{N} \frac{\tilde{e}_n(t) + 10.53}{\sigma_n(t)} \right]^{\frac{1}{N}} \quad (2)$$

where β is a fixed scale factor. The silence floor (10.53) has been added to the log-normalised component energy $\tilde{e}_n(t)$ to constrain $\xi(t)$ to be positive. An example spectragram and associated confidence $\xi(t)$ are shown in Figure 1. The confidence metric is high in regions with strong formant structures and low during unvoiced sounds, as expected.

In previous work [4], a synchronous stream system for combining the features was investigated. For a multiple stream system the output probability distribution $b_j(\boldsymbol{y}(t))$ for input vector $\boldsymbol{y}(t)$ divided into R streams $\{\boldsymbol{y}_1(t), \ldots \boldsymbol{y}_R(t)\}$ is calculated as

$$b_j(\boldsymbol{y}(t)) = \prod_{r=1}^{R} \Big[\sum_{m=1}^{M} c_{jrm} \mathcal{N}(\boldsymbol{y}_r(t); \boldsymbol{\mu}_{jrm}, \boldsymbol{\Sigma}_{jrm}) \Big]^{\gamma_r(t)} \quad (3)$$

where $\gamma_r(t)$ is the stream weight and c_{jrm}, $\boldsymbol{\mu}_{jrm}$ and $\boldsymbol{\Sigma}_{jrm}$ are the weight, mean and variance for component m of stream r for state j. In our systems

$$\boldsymbol{y}_1(t) = [m_1(t), \ldots, m_{12}(t), l(t), \Delta, \Delta^2]^T$$
$$\boldsymbol{y}_2(t) = [\mu_1(t), \ldots, \mu_6(t), \Delta, \Delta^2]^T$$

where m_k is the k^{th} MFCC and $l(t)$ is the log normalised energy of the complete frame. The first and second order derivatives, Δ and Δ^2 respectively, are also included. Usually $\gamma_r(t)$ is constant over all time frames. In this work $\gamma_r(t)$ is made a function of the confidence metric, so the GMM stream is deweighted in regions with poorly defined formants. However, the feature streams have different dynamic ranges and hence different optimal language model scale factors, α_1 and α_2. By scaling the acoustic model scores it is possible to use α_1 as the language model scale factor regardless of $\xi(t)$. In a confidence framework the stream weights are given by

$$\gamma_1(t) = 1 - \xi(t) \quad (4)$$

$$\gamma_2(t) = \left[\frac{\alpha_1}{\alpha_2}\right] \xi(t) \quad (5)$$

and these will be substituted into Eqn. 3 during recognition.

3. COMPENSATING GMM PARAMETERS IN NOISE

Spectral peak features have been demonstrated to give slight improvements when combined with MFCC systems on noisy data [1] [2]. However, these representations of spectral peaks cannot be used to recover the spectrum. So compensation using a noise model is not simple. In contrast, the GMM features can be used to reconstruct the spectrum. Two approaches for compensating the GMM features are outlined, one during the extraction stage, the other operating on the models. Both compensation schemes use a model derived from the average GMM parameters for the noise. Figure 2 shows a GMM plot of a clean spectrum and one with additive "operations room" (OpRoom) noise from the Noisex database at 18dB signal to noise ratio. Previous work has shown that this form of noise severely corrupts MFCC parameterisations [6]. The OpRoom noise is coloured and possesses a strong low frequency spectral peak. A spectral peak representation of the corrupted speech signal will model the noise rather than the speech in this low frequency region.

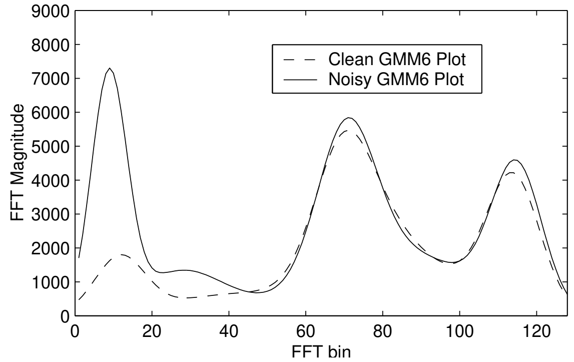

Fig. 2. Plots of GMM with 6 components fitted to a clean spectrum and one with additive OpRoom noise

A noise model can be used when fitting the speech Gaussian components to noise corrupted data. The noise model consists of the average parameters $\hat{\boldsymbol{\theta}}^{(n)}$ of a series of Q-component GMMs fitted to the noise.

$$\hat{\boldsymbol{\theta}}^{(n)} = \{\hat{\mu}_1^{(n)}, \ldots, \hat{\mu}_Q^{(n)}, \hat{\sigma}_1^{(n)2}, \ldots, \hat{\sigma}_Q^{(n)2}, \hat{e}_1^{(n)}, \ldots, \hat{e}_Q^{(n)}\}$$

The following log likelihood is then optimized with respect to the clean speech parameters $\boldsymbol{\theta}(t)$

$$l(\boldsymbol{x}(t)|\boldsymbol{\theta}(t),\hat{\boldsymbol{\theta}}^{(n)}) = \sum_{k=1}^{K} \ln \left(\sum_{q=1}^{Q} \hat{e}_q^{(n)} \mathcal{N}\left(x_k(t); \hat{\mu}_q^{(n)}, \hat{\sigma}_q^{(n)2}\right) + \sum_{n=1}^{N} e_n(t) \mathcal{N}\left(x_k(t); \mu_n(t), \sigma_n^2(t)\right) \right) \quad (6)$$

This technique avoids the problems of underflow that occur with spectral subtraction techniques. When this approach was applied, the observed tracks for the GMM fits were closer to the clean speech, but unfortunately exhibited large discontinuities between certain frames. These were caused by the noise masking the speech during low intensity sounds. To counteract this effect, a moving average filter was applied to smooth the parameters extracted using this noise compensation.

To avoid the problems of masking caused by the front end noise compensation, the HMMs may be adapted with the noise model. Using the static means from the output probability distribution function (PDF) of HMM state j component m, it is possible to estimate the average of the GMM parameters $\hat{\boldsymbol{\theta}}_{jm}$. A spectrum $\{x_{jm1}, \ldots, x_{jmK}\}$ can be generated from the GMM parameters. A noise spectrum $\{q_1, \ldots, q_K\}$ can then be added to the reconstructed spectrum. Thus, by optimising the likelihood

$$l(\boldsymbol{x}_{jm} + \boldsymbol{q}|\hat{\boldsymbol{\theta}}_{jm}) = \sum_{k=1}^{K} \left(\ln \sum_{n=1}^{N} \hat{e}_{jmn} \mathcal{N}\left(x_{jmk} + q_k; \hat{\mu}_{jmn}, \hat{\sigma}_{jmn}^2\right) \right) \quad (7)$$

we can re-estimate the average static GMM parameters for each state in the HMM. The cost of this is the same order of magnitude as compensation using parallel model combination (PMC) with a log-add approximation [6].

4. EXPERIMENTAL RESULTS

Experiments were carried out on two tasks, RM and WSJ. The RM corpus task has roughly a 1000 word vocabulary, and models were built up using the RM recipe [7]. Experiments on the WSJ corpus used the si284 training set as described in [8] with a trigram language model and a 65,000 word vocabulary. The noise source used was the OpRoom noise from the Noisex corpus. The clean speech waveforms were corrupted by adding noise data. Although this is not a realistic noise condition, it does allow more control as required for initial investigations.

For both corpora, cross-word decision tree clustered triphone HMMs were trained. The number of components in the state PDFs was increased until no improvements were observed on a subset of the test data. The language model scale factors were tuned on subsets of the data, except for noise experiments where the noise-free scale factor was used. Δ and Δ^2 parameters were added to all feature vectors.

4.1. Using a Confidence Metric in Recognition

A two stream system was built using the stream definitions specified in section 2. The stream weight of the GMM means $\gamma_2(t)$ was set to zero during training. A set of models were trained on the clean data and tested using the confidence metric. The results are shown in Table 1.

Confidence scale β	Clean	18dB SNR
0.0 (MFCC only)	4.19	32.3
0.1	3.95	29.6
0.2	3.94	29.6
0.3	4.12	30.4
0.4	4.23	31.9
0.5	4.32	33.5
$\xi(t) = 1$ (GMM only)	9.23	58.3

Table 1. Word Error Rates (%) on RM task using confidence metrics, for clean data and with added noise

The confidence metric gives a small improvement over the MFCC baseline system. The same models were used on the test data with additive OpRoom noise at a SNR level of 18dB and the results are also shown in Table 1. Using the confidence metric and the GMM parameters, a 9% reduction in error rates relative to the MFCC system in noise were obtained, although the error rate is still high. No improvements were obtained by retraining the models using the confidence weights in Eqn 3.

Description	% WER
MFCC	9.75
MFCC+6 Means Concatenative	9.56
MFCC+6 Means $\xi(t)$ = 0.2	9.64
MFCC+6 Means β=0.2	9.52

Table 2. Word error rates on the WSJ task

Two comparison systems were built on the WSJ task: one using only MFCCs [8] and another using a concatenative feature vector of the 6 GMM means and the MFCCs in a single stream. The streaming system used a 12 component output PDF for each HMM state, the concatenative system 16. Results for the systems are presented in Table 2. Using the confidence metric produces a small improvement on the performance of a system with $\xi(t)$ fixed at 0.2. However, the performance increase over the concatenative system is not significant. Both the concatenative and confidence streamed systems outperform the MFCC baseline.

4.2. Adding GMM Means to MFCC features in Noise

In order to initially explore the performance of GMM features in noise, a single stream systems was built by concatenating the MFCCs and the GMM mean features into a single feature vector. An MFCC system was also built for comparison. The systems were tested on data with additive OpRoom noise and results are shown in Table 3 and Figure 3.

The concatenative MFCC + GMM means system performs slightly better than the MFCC system in additive noise, suggesting that the GMM means supply complementary information to MFCCs in noisy environments. The OpRoom noise corrupts speech badly even at a relatively high SNR.

A noise matched system was built using single pass retraining from the models on data with with added OpRoom noise at 18dB SNR. In these noise matched conditions, adding GMM mean features to the MFCCs reduces the WER from 8.2% to 7.1% at 18dB SNR. This improvement suggests that if the GMM means can be compensated, then they possess complementary information in noise-corrupted environments.

Description	WER /%
MFCC (UC)	32.3
MFCC (UC) + GMM (UC)	30.6
+ Confidence	29.6
MFCC (UC) + GMM (FC)	28.3
MFCC (UC) + GMM (MC)	22.1
MFCC (MC)	14.0
MFCC (MC) + GMM (MC)	13.1
+ Confidence	12.6

Table 3. Word Error Rates (%) on RM task with additive OpRoom noise at 18dB SNR with uncompensated, smoothed front end compensation (FC) and static mean model compensation (MC)

The front-end noise compensation technique presented in Section 3 was implemented on noisy test data for the RM task. A four-component model of the noise was obtained by taking the average values of a GMM fit to the noise spectra. The compensated GMM means were combined with the uncompensated MFCCs and were tested with the clean system. A moving average filter of length 3 was applied over the GMM mean features after the front end compensation as mentioned previously. The filter was applied prior to the calculation of dynamic parameters. Without smoothing, the system performed slightly worse than an uncompensated system. Note that when a moving average was applied to the clean system in [4], performance was degraded. However, when used with this compensation scheme, a small improvement was observed. The model compensation technique outlined in Section 3 was also

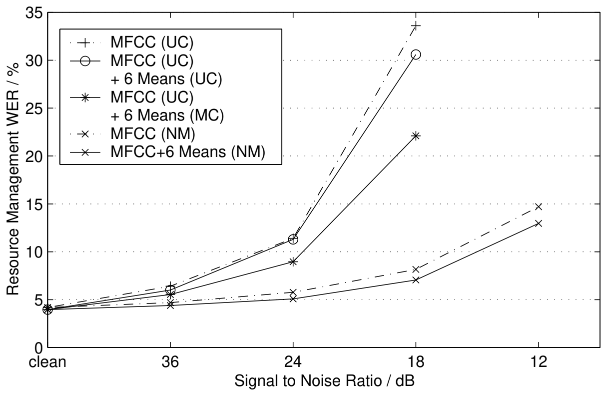

Fig. 3. Word Error Rates on the RM task with additive OpRoom noise for uncompensated (UC), model compensated of static means (MC) and noise matched data (NM) systems

tested and the results are shown in Table 3 and Figure 3. Model compensation of the static mean MFCC parameters was simulated by replacing the values in a clean model by the those from a noise matched system. A relative improvement of 31% was achieved when only the six static means were compensated in the feature vector. Adding GMM mean features to a static mean compensated MFCC system reduces error rates by 7%. The GMM parameters allow a compensation technique to work directly in the spectral domain, thus reducing the complexity of mapping linear cepstral domain and the log-add approximations that are made with PMC on MFCC features [6]. As is the case with MFCC features, adapting the GMM model parameters yields better performance than compensation at the front-end level.

The model compensated systems were also tested using the confidence metric to combine the MFCC and GMM features, and the result is in Table 3. Using the confidence metric gives a 4% reduction in WER relative to a single stream system. However, using the confidence metric with a noise matched system gave no reduction in error rate.

5. CONCLUSIONS AND FUTURE WORK

This paper has described two extensions to the use of GMM features in combination with the standard MFCC parameters. A technique to combine GMM features using a confidence metric has been presented. Using this system gives a slight improvement on the WSJ task relative to a standard streaming system or an MFCC parameterisation. However, it does not give significant improvements over a single-stream, concatenative system on clean data. In addition, the behaviour of GMM features in a noisy environment have investigated and novel schemes proposed to rapidly adapt the GMM features. Combining MFCC features with GMM features in noise led to improved performance of noise matched and static mean model compensated systems. In future work, experiments will be performed on speech recorded in a noisy environment rather than using artificially added noise.

6. REFERENCES

[1] R. De Mori, L. Moisa, R. Gemello, F. Mana, and D. Albensano, "Augmenting Standard Speech Recognition Features with Energy Gravity Centres," *Computer Speech and Language*, vol. 15, pp. 341–354, 2001.

[2] K. Weber, S. Bengio, and H. Bourlard, "HMM2- Extraction of Formant Features and their Use for Robust ASR," in *Proceedings Eurospeech*, 2001.

[3] P. Zolfaghari and T. Robinson, "Formant Analysis Using Mixtures of Gaussians," in *Proceedings ICSLP*, 1996.

[4] M. Stuttle and M.J.F. Gales, "A Mixture of Gaussians Front End for Speech Recognition," in *Proceedings Eurospeech*, 2001.

[5] N. Wilkinson and M.J. Russell, "Progress Towards Improved Speech Modelling Using Asynchronous Sub-bands and Formant Frequencies," in *Proceedings WISP*, 2001.

[6] M.J.F. Gales and S.J. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 352–359, 1996.

[7] S. Young, G Evermann, D. Kershaw, G. Moore, J.J. Odell, D. Ollason, V. Valtchev, and P.C. Woodland, *"The HTK Book, Version 3.1"*, Cambridge University Engineering Department, 2002.

[8] P.C Woodland, C.J Legetter, J.J. Odell, V. Valtchev, and S.J. Young, "The 1994 Large Vocabulary Speech Recognition System," in *Proceedings ICASSP*, 1994.