FILTER BANK FEATURE EXTRACTION FOR GAUSSIAN MIXTURE MODEL SPEAKER RECOGNITION

James H. Nealand, Alan B. Bradley, & Margaret Lech

School of Electrical and Computer Systems Engineering,
RMIT University, Melbourne, Australia

ABSTRACT: Speaker Recognition is the task of identifying an individual from their voice. Typically this task is performed in two consecutive stages: feature extraction and classification. Using a Gaussian Mixture Model (GMM) classifier different filter-bank configurations were compared as feature extraction techniques for speaker recognition. The filter-banks were also compared to the popular Mel-Frequency Cepstral Coefficients (MFCC) with respect to speaker recognition performance on the CSLU Speaker Recognition Corpus. The empirical results show that a uniform filter-bank outperforms both the mel-scale filter bank and the MFCC as a feature extraction technique. These results challenge the notion that the mel-scale is an appropriate division of the spectrum for speaker recognition.

INTRODUCTION

Speaker recognition is the task of establishing personal identity from a spoken utterance. Speaker recognition encompasses the tasks of speaker identification and verification. Speaker identification is the task of identifying a target speaker from a group of possible speakers, whereas speaker verification is the task of accepting or rejecting a claim of identity from a speaker.

Speaker identification and verification is typically performed in two stages: feature extraction and classification. The feature extraction process reduces the speech signal to finite feature vectors that convey speaker identifying information. The classification process is typically stochastic and compares the observed feature vectors to a pre-built model of a speaker.

Feature extraction is typically performed on short overlapping frames of speech (< 30ms), during which the speech is assumed to be quasi-stationary. Popular feature extraction techniques include the Mel-Frequency Cepstral Coefficients (MFCC) and Linear Prediction (LP) based techniques (Reynolds 1994). One of the key aspects of the MFCC for speech feature extraction is that the mel-frequency scale resembles human auditory perception. There is no theoretical or empirical evidence to suggest that the mel-scale is in any way an optimal division of the frequency spectrum for speaker separability. Despite MFCC having been used extensively for speaker recognition, there is no evidence to suggest that it is optimal for feature extraction for speaker recognition.

Filter-banks are common in signal processing and have been used as a feature extraction technique for speech recognition (Biem, Katagiri, McDermott & Juang 2001). A filter-bank in the context of feature extraction divides the spectrum into bands. For a single frame of speech each band becomes one dimension of the feature vector. A filter-bank is defined by the number of filters, the shape, centre frequency and bandwidth of each filter.

This paper reports on experiments comparing both mel-scale and uniform filter-banks to the MFCC metric for speaker recognition using a Gaussian Mixture Model (GMM) classifier. The GMM is a standard classifier for speaker recognition having demonstrated robust speaker identification and verification performance (Reynolds 1995).

The experiments were performed on the CSLU Speaker Recognition Corpus; a database of telephone quality speech collected over a period of two years (Cole, Noel & Noel 1998). The CSLU Speaker Recognition corpus provides a realistic speaker recognition task, although to date there have been little published results using this corpus.

The experiments show that the uniform scale filter-bank outperforms the mel-scale filter-bank for GMM based speaker recognition on the CLSU Speaker Recognition corpus. Furthermore the uniform filter-bank outperforms the MFCC as a feature extraction technique for speaker recognitions. These results although limited, challenge the notion that the mel-scale division of the spectrum is appropriate for speaker recognitions.

Proceedings of the 9th Australian International Conference on Speech Science & Technology
Melbourne, December 2 to 5, 2002.

Accepted after full review

THE GAUSSIAN MIXTURE MODEL

The GMM is a specific configuration of a radial basis function artificial neural network, and has shown robust text independent results for both speaker identification and verification applications (Reynolds 1994; Reynolds 1995; Reynolds & Rose 1995; Reynolds, Rose & Smith 1992).

The GMM models the observed feature vectors as a weighted sum of *M* Gaussian components.

$$p\left(x_t \middle| \lambda_s\right) = \sum_{i=1}^{M} w_{si} b_{si}\left(x_t\right) \tag{1}$$

Where each Gaussian component $b_{si}\left(\bullet\right)$ is a normal probability density function, w_{si} is the *priori* probability of the *i*'th Gaussian component, x_t is the observed feature vector for frame *t* and λ_s is the GMM for speaker *s*. Each Gaussian component is given by Equation (2).

$$b_{si}\left(x_t\right) = \frac{1}{\left(2\pi\right)^{D/2}\left|\Sigma_{si}\right|^{1/2}} e^{-\frac{1}{2}\left\{\left(x_t-\mu_{si}\right)^T\left(\Sigma_{si}\right)^{-1}\left(x_t-\mu_{si}\right)\right\}} \tag{2}$$

The parameters μ_{si} and Σ_{si} are the mean and covariance parameters of the *i*'th Gaussian component for speaker *s* respectively, while *D* is the dimension of the feature vector. The number of Gaussian components used was 16, which is common in text-independent speaker recognition applications (Reynolds 1995). The covariance matrices were constrained to be diagonal. Reynolds (1995) claims that empirical evidence suggests that diagonal covariance matrices outperform full order matrices.

The likelihood of a speaker having generated the utterance of length *T* frames $X = \left\{x_t, 1 \le t \le T\right\}$ is the multiplication of the speaker having generated each feature vector x_t in the utterance. The logarithm of the likelihood is taken to make the multiplication an additive process as shown in Equation (3).

$$\log\left(p\left(X\middle|\lambda_s\right)\right) = \sum_{t=1}^{T} \log\left(p\left(x_t\middle|\lambda_s\right)\right) \tag{3}$$

A GMM is constructed independently for each speaker using training or enrolment data provided from each speaker. Although a number of approaches can be used to construct the models, a conventional two-stage approach used in these experiments. The models were first initialised using the K-means clustering algorithm, and then trained using the Expectation Maximisation (EM) algorithm (Dempster, Laird & Rubin 1977).

FILTER-BANK BASED FEATURE EXTRACTION

Filter-banks have previously been applied in both speech and speaker recognition although the comparison between types of filter-banks for speaker recognition has not been reported extensively in the literature.

In the context of feature extraction the output of each filter is one dimension of the feature vector and represents the energy in a certain region of the speech spectrum. The filter-banks used in the experiments reported herein were emulated using a Fourier based approach identical to that used by Biem (Biem, Katagiri, McDermott & Juang 2001). The output of the *i*'th filter for frame *t* is given by Equation (4).

$$y_{it} = \log_{10}\left(w_i^T x_t\right) \tag{4}$$

The parameter x_t is the FFT of the windowed frame of samples, and w_i is the vector of spectral weightings for the *i*'th filter as calculated by Equation (5). For all of the experiments reported herein a hamming window of length 160 samples was applied giving the frame duration of 20ms. There was a 50% overlap between successive frames. Prior to the FFT the samples were zero padded to 256 samples so that a faster FFT routine could be used.

$$w_i[n] = \alpha_i e^{-\beta_i (n-\gamma_i)^2} \tag{5}$$

For the uniform filter-bank the centre frequencies of the filters were distributed evenly over the useable frequency range. The data was collected over digital telephone lines and sampled at 8kHz. The bandwidth of the filters in the uniform filter-bank was chosen such that adjacent filters intersect at the point of 3dB attenuation for both filters.

For the mel-scale filter-bank the centre frequencies of the filters were distributed evenly over the mel-frequency scale. The mel-scale is approximated in (Picone 1993) as Equation (6).

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \tag{6}$$

The bandwidths of the filters in the mel-scale filter bank were calculated using the expression for critical bandwidth given in (Picone 1993) and shown in Equation (7).

$$BW = 25 + 75\left[1 + 1.4\left(\frac{f}{1000}\right)^2\right]^{0.69} \tag{7}$$

The mel-scale filter bank used in the experiments reported herein may otherwise be known as a critical band filter-bank (Picone 1993).

THE MEL-FREQUENCY CEPSTRAL COEFFICIENTS

The MFCC are a standard feature extraction metric for speech and speaker recognition. The MFCC are calculated by taking the Discrete Cosine Transform (DCT) of the log energy of the output of a mel-scale filter bank proposed by (Davis & Mermelstein 1980). As is typically done in speaker recognition the first Cepstral coefficient was discarded from each feature vector (Reynolds 1995). This means that for a 23-dimension MFCC feature vector a 24-dimension mel-scale filter-bank was applied.

The MFCC was chosen because it is a standard feature extraction technique and has been reported to show robust speaker recognition performance in the past (Reynolds 1994; Reynolds 1995). Comparing results obtained with a filter-bank based feature extraction to that obtained with the MFCC is useful in assessing the suitability of a filter-bank structure for speaker recognition feature extraction.

SPEECH DATA

A subset of the CSLU Speaker Recognition Corpus was selected for the training and testing data. The Speaker Recognition Corpus is a database of telephone quality speech collected over digital telephone lines. Each speaker contributed 12 sessions of speech data over a period of 2 years. Four sessions of speech data were designated as training sessions. The following 4 sessions were designated as testing data. Five sentences of speech were chosen from each of the training sessions from each speaker for training data. A total of approximately 60s of speech from each speaker was used for training.

DECISION CRITERIA

The speaker recognition decision criteria are different for speaker verification and identification. For identification the most likely speaker is chosen from the group of ten speakers. This decision criterion

is based on Bayes minimum error rule (Fukunaga 1990) and is given by Equation (8) where C_k represents the *k*'th speaker.

$$X \in C_k \text{ if } k = \arg\max_j \left\{ \log\left(P\left(X \middle| \lambda_j \right) \right) \right\} \tag{8}$$

The decision rule for the speaker verification experiments is binary. The claim of identity is either accepted or rejected. This leads to two types of possible errors: false acceptance and false rejection errors. A false acceptance error occurs when an impostor speaker is falsely accepted as the claimant speaker. A false rejection error occurs when a legitimate claim of identity is rejected. The same sets of 10 speakers were used in both the verification experiments. For any claimant speaker the remaining 9 speakers in the set were designated as background speakers. The likelihood of the claimant speaker having produced the utterance was compared to the average likelihood of the background speakers having produced the utterance. The result was compared to a threshold *K*, which controls the ratio between false rejection and false acceptance errors.

$$X \in C_k \text{ if } \left[\log\left(P\left(X \middle| \lambda_k \right) \right) - \frac{1}{9} \sum_{j \neq k} \log\left(P\left(X \middle| \lambda_j \right) \right) \right] \geq K \tag{9}$$

It is standard in speaker verification experiments to quote the error rate as the point where the rate of false acceptance errors is equal to the rate of false rejection errors. The threshold is varied to determine this rate, otherwise known as the Equal Error Rate (EER).

EXPERIMENT

Both speaker identification and verification experiments were performed. For the speaker identification experiments 5 groups of ten speakers were randomly selected. Each group of speakers were evaluated independently and the recognition results averaged. For the verification experiments 50 speakers were evaluated as claimant speakers, and sample impostor speakers were chosen so that no impostor speaker was among a claimants background speaker models. This precaution ensures a valid speaker verification experiment.

Both the speaker identification and verification performance were evaluated with respect to utterance length. Tests were generated for longer utterances by connecting different sentences in the same manner suggested by (Reynolds & Rose 1995). Both the performance over the four training sessions and the performance over the 4 testing sessions were evaluated.

RESULTS

Table 1 and Table 2 show the results of the speaker identification and verification experiments respectively for both the 12 and 23 dimension feature vectors. Both Tables show the recognition results with respect to utterance length in frames. Since the window length of each frame was 20ms and the hopping rate was 10ms, 1000 frames represent an utterance length of approximately 10s.

For the training sessions it is observed that the recognition results are high, with the uniform filter bank results narrowly outperforming the mel-scale filter bank and both filter-banks outperforming the MFCC feature vectors.

For the testing sessions the uniform filter-bank consistently outperforms both the mel-scale filter-bank and the MFCC feature vectors. The mel-scale filter-bank outperformed the MFCC feature vector in the first test session, however the MFCC features outperformed the mel-scale filter-bank in the later test session.

Table 1. Speaker identification results with respect to utterance length

Test Conditions		12-D Results v Utterance Length (Frames)					23-D Results v Utterance Length (Frames)				
Session	FB	1	50	200	500	1000	1	50	200	500	1000
Training	Uniform	40.2%	86.8%	97.5%	99.5%	99.7%	44.5%	91.2%	98.6%	99.8%	100.0%
Training	Mel	41.4%	86.6%	96.9%	98.8%	99.3%	43.5%	88.5%	97.8%	99.3%	99.7%
Training	MFCC	19.9%	69.5%	86.1%	93.0%	95.3%	22.6%	77.7%	91.0%	95.7%	96.5%
Test 1	Uniform	32.8%	70.4%	82.5%	87.8%	90.1%	35.2%	73.0%	83.4%	88.1%	90.7%
Test 1	Mel	30.8%	59.7%	70.4%	73.5%	80.4%	31.8%	59.9%	70.3%	75.1%	81.5%
Test 1	MFCC	16.7%	46.6%	57.3%	62.9%	66.8%	18.0%	54.9%	67.3%	72.9%	74.0%
Test 2	Uniform	24.6%	50.3%	61.4%	64.1%	63.9%	26.2%	52.4%	61.5%	65.2%	65.6%
Test 2	Mel	22.7%	40.7%	46.6%	51.2%	50.7%	22.9%	41.5%	48.6%	53.2%	53.6%
Test 2	MFCC	14.5%	34.2%	41.4%	46.0%	51.1%	15.8%	39.3%	49.1%	54.4%	56.4%
Test 3	Uniform	20.6%	39.9%	46.4%	49.2%	48.5%	22.4%	41.3%	47.7%	49.3%	50.2%
Test 3	Mel	19.4%	30.3%	36.5%	38.7%	41.1%	19.9%	31.5%	36.8%	39.9%	41.9%
Test 3	MFCC	13.4%	27.4%	32.7%	37.0%	42.7%	14.4%	32.6%	36.6%	41.9%	46.2%
Test 4	Uniform	21.0%	41.6%	49.0%	51.4%	53.4%	23.3%	44.6%	50.3%	51.7%	53.4%
Test 4	Mel	21.2%	35.9%	42.2%	46.2%	48.7%	21.1%	34.6%	40.0%	42.6%	44.1%
Test 4	MFCC	13.5%	29.2%	36.2%	42.6%	48.3%	14.6%	33.4%	40.7%	44.2%	45.8%

Table 2. Equal Error Rates v Utterance Length for Speaker Verification Experiments

Test Conditions		12-D Results v Utterance Length (frames)						23-D Results v Utterance Length (frames)					
Sessions	FB	1	50	100	200	400	500	1	50	100	200	400	500
Training	Uniform	29.82%	10.16%	7.03%	4.47%	2.41%	1.98%	28.08%	8.70%	5.95%	3.64%	3.22%	3.32%
Training	Mel	29.28%	10.18%	7.04%	4.65%	2.72%	2.05%	28.66%	9.83%	6.86%	4.18%	3.58%	3.72%
Training	MFCC	40.97%	17.99%	14.26%	11.04%	9.73%	10.26%	36.24%	13.09%	10.27%	7.52%	6.79%	7.27%
Test 1	Uniform	33.26%	16.35%	14.31%	12.15%	8.45%	8.34%	32.34%	16.22%	14.06%	11.71%	10.12%	9.24%
Test 1	Mel	34.89%	21.93%	19.63%	16.52%	13.15%	12.61%	34.52%	21.75%	19.33%	17.60%	15.41%	14.06%
Test 1	MFCC	43.84%	27.77%	26.47%	23.68%	22.51%	21.58%	42.80%	24.22%	21.19%	16.89%	14.12%	13.03%
Test 2	Uniform	38.96%	27.21%	24.59%	23.63%	21.54%	20.88%	37.97%	27.43%	25.37%	24.44%	22.89%	22.64%
Test 2	Mel	40.21%	33.68%	33.44%	34.15%	34.01%	33.77%	40.47%	34.08%	34.04%	34.51%	34.47%	34.36%
Test 2	MFCC	45.61%	34.98%	33.94%	34.06%	33.70%	33.45%	44.57%	32.24%	30.39%	29.63%	32.39%	31.99%
Test 3	Uniform	41.56%	32.98%	31.16%	29.85%	29.91%	30.47%	41.13%	33.43%	31.67%	31.04%	29.73%	30.11%
Test 3	Mel	43.11%	37.64%	36.77%	36.73%	38.63%	38.37%	42.52%	37.80%	37.22%	36.91%	38.42%	38.12%
Test 3	MFCC	46.56%	39.29%	37.79%	36.55%	36.49%	35.68%	45.96%	35.32%	33.15%	32.06%	30.99%	31.54%
Test 4	Uniform	41.81%	35.00%	34.45%	32.93%	32.95%	32.78%	40.14%	33.78%	33.89%	34.98%	34.99%	34.97%
Test 4	Mel	42.09%	36.03%	34.31%	34.68%	35.50%	36.97%	42.19%	37.24%	37.55%	37.73%	38.76%	39.42%
Test 4	MFCC	46.70%	38.00%	37.52%	37.38%	37.00%	36.37%	46.01%	35.63%	34.84%	33.44%	33.28%	32.55%

DISCUSSION

In both speaker verification and identification experiments the uniform filter-bank consistently outperformed the mel-scale filter-bank. This result is consistent for both 12 and 23 dimension feature vector experiments, although there is no consistent evidence to suggest whether the 12-dimension or 23-dimension feature vectors are superior.

For the first test session in both identification and verification experiments the filter-banks both outperform the MFCC feature vectors. In the later test sessions and for longer utterances the MFCC feature vectors outperform the mel-scale filter bank but not the uniform filter-bank performance. This finding indicates that the MFCC feature vectors may be more resilient to the variation in a speakers voice over time than the filter-bank feature vectors. It is observed that this variation of a speakers voice over time otherwise known as ageing effects have a significant impact on all feature vectors. For recognition of shorter utterances and in single frames, both filter-banks outperform the MFCC feature vectors for all test sessions.

The observations from Table 1 and Table 2 challenge the notion that the mel-scale is an appropriate division of the spectrum for speaker recognition. It is not suggested that the uniform filter-bank is in anyway optimal for speaker recognition. Further experiments are necessary to determine an optimal approach to feature extraction for speaker recognition. Data-driven approaches to feature extraction optimisation are currently being investigated (Nealand, Bradley & Lech 2002).

The CSLU Speaker Recognition Corpus is a practical, real-world environment for speaker recognition testing in the presence of background noise, channel noise, linguistic variation and ageing effects. As such the recognition rates are not as high as those reported on less practical speech databases.

A possible criticism of the experiments is that no attempt at channel normalisation or noise removal was applied prior to the feature extraction. Either of these techniques may offer substantial improvements to recognition performance as shown by Reynolds (Reynolds 1994). Furthermore MFCC features are known to be highly susceptible to noise. Background and channel noise is a real and practical problem in speaker recognition. The experiments consider the robustness of feature extraction techniques in the presence of background and channel noise. Future work will consider the data-driven development of noise and channel robust feature extraction for speaker recognition.

CONCLUSIONS

The performance of uniform and mel-scale filter-banks as feature extraction techniques for speaker recognition has been assessed over the CSLU Speaker Recognition corpus using a GMM classifier. The uniform filter-bank consistently outperformed both the mel-scale filter-bank and the MFCC feature set, however there is evidence to suggest that the MFCC features were less prone to ageing effects than the filter-banks. These findings challenge the notion that the mel-scale division of the spectrum is appropriate for speaker recognition.

REFERENCES

A. Biem, S. Katagiri, E. McDermott & B.-H. Juang, (2001). *An Application of Discriminative Feature Extraction to Filter-Bank-Based Speech Recognition*, IEEE Transactions on Speech and Audio Processing 9, 96-110.

R. Cole, M. Noel & V. Noel, (1998). *The CSLU Speaker Recognition Corpus*, International Conference on Spoken Language Processing 7, 3167-3170

S. B. Davis & P. Mermelstein, (1980). *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*, IEEE Transactions on Acoustics Speech and Signal Processing 28, 357-366.

A. P. Dempster, N. M. Laird & D. B. Rubin, (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society 1, 1-38.

K. Fukunaga, (1990). *Introduction to Statistical Pattern Recognition*, Academic Press Inc.

J. H. Nealand, A. B. Bradley & M. Lech, (2002). *Discriminative Feature Extraction Applied to Speaker Identification*, International Conference on Signal Processing 1, 484-487

J. W. Picone, (1993). *Signal Modelling Techniques in Speech Recognition*, Proceedings of the IEEE 81, 1215 -1245.

D. A. Reynolds, (1994). *Experimental evaluation of features for robust speaker identification*, IEEE Transactions on Speech and Audio Processing 2, 639-643.

D. A. Reynolds, (1995). *Speaker identification and verification using Gaussian mixture speaker models*, Speech Communication 17, 91-108.

D. A. Reynolds & R. C. Rose, (1995). *Robust text-independent speaker identification using Gaussian mixture speaker models*, IEEE Transactions on Speech and Audio Processing 3, 72-83.

D. A. Reynolds, R. C. Rose & M. J. T. Smith, (1992). *PC-Based TMS320C30 Implementation of the Gaussian Mixture Model Test-Independent Speaker Recognition System*, International Conference on Signal Processing Applications and Technology, 967-973.