Using Long-Term Information to Improve Robustness in Speaker Identification

James G. Lyons, James G. O'Connell, Kuldip K. Paliwal

Signal Processing Laboratory, Griffith School of Engineering
Griffith University, Brisbane Queensland 4111, Australia

j.lyons@griffith.edu.au, james.oconnell@griffithuni.edu.au, k.paliwal@griffith.edu.au

Abstract

In this paper we propose two new methods of improving the robustness of Automatic Speaker Identification systems. These methods rely on using long-term information in the speech signal to improve the robustness of the features. The first method involves averaging filterbank parameters from consecutive short-time frames over a longer window. The second method investigates the use of frame lengths longer than generally assumed stationary. We show that these two methods result in an improvement over standard Mel Frequency Cepstral Coefficients in the presence of additive white Gaussian noise in speaker identification applications. Furthermore, additional improvements are observed at mid-range SNR when the proposed methods are used in combination.

Index Terms: Feature averaging, analysis window duration, long window, speaker recognition, automatic speaker identification

1. Introduction

Automatic speech recognition (ASR) and automatic speaker identification (ASI) have opposing goals: the former attempts to identify the speech content within a signal independent of the speaker-specific vocal tract characteristics, while the latter ultimately attempts to identify the speaker independent of the speech content. Interestingly, modern ASR and ASI systems use the same base features, usually mel-frequency cepstral coefficents (MFCC), derived from the short-time magnitude spectrum to achieve these opposing goals.

The estimation of the magnitude spectrum using the short-time Fourier transform (STFT) is performed due to the non-stationary nature of the speech signal. Since the vocal tract articulatory structures can only move at a finite speed, it is assumed that at sufficiently small time periods the speech signal can be considered stationary. For this reason, frames of 20-30ms in length are typically chosen [1].

This short-term analysis is important in ASR, where features in the speech signal may persist for only short periods of time. In ASI, however, the speaker's identity is constant throughout the utterance. While short frames will capture the dynamics of an individual's vocal tract, it will also make the model more susceptible to noise than a longer term analysis or ensemble average.

Huang et al. proposed averaging STFT, MFCC-based feature vectors over an extended window, which they entitled *short-time frequency with long-time window* (SFLW), and showed that this increased robustness to additive white Gaussian noise (AWGN) in an automatic speaker verification (ASV) application [2]. This study aims to expand on this by proposing two further methods for improving the robustness of ASI. The first method is a modification to that proposed by Huang et al. [2], and averages short-time filterbank parameters over longer windows. The second method investigates the use of STFT frame lengths generally assumed non-stationary, which we will refer to as long-time Fourier transform (LTFT) analysis. Both of these methods are tested in the presence of AWGN at several SNR.

The remainder of the paper investigates the possibility of improving robustness through the fusion of methods using a combination of log-likelihoods from each method, and the effective class separability provided by each method.

2. Experiment

2.1. Database

This study employed the TIMIT database, consisting of spoken sentences sampled at 16kHz with 10 utterances each from 630 speakers [3]. The *si** and *sx** utterances for each speaker were used for training (8 utterances per speaker), while the *sa** utterances were used for testing (2 utterances per speaker). For testing purposes, the utterances were degraded by AWGN at several SNR. Speakers from the *test* subset of TIMIT (168 speakers) were used for training and testing.

2.2. Feature Generation

Two feature generation methods are proposed in this study. In the first method, the LTFT is used to obtain spectral estimates of the framed, pre-emphasised speech signal. The discrete LTFT of a signal is given by:

$$X_k = \sum_{n=0}^{N-1} \omega_n x_n \mathrm{e}^{-\frac{j2\pi kn}{N}}, \quad 0 \leq k < N-1 \qquad (1)$$

where ω_n is an analysis window function. This is, of course, identical to that normally employed to calculate the STFT. However, the LTFT uses frame lengths that exceed that typically assumed stationary for speech signals. From the LTFT spectral analysis, 12 MFCC are generated from the estimated magnitude spectrum. To this, some combination of delta, acceleration and energy coefficients can be appended to generate the final features.

In the second method, the SFLW method employed by [2] is used. Instead of averaging the final MFCC features, however, the Mel-filterbank parameters (prior to taking the logarithm) are averaged across frames within the window, and the MFCC generated from these band averages: the FAVG, or filterbank averaging method. Once again, some combination of delta, accel-

978-1-4244-7907-8/10/$26.00 © 2010 IEEE

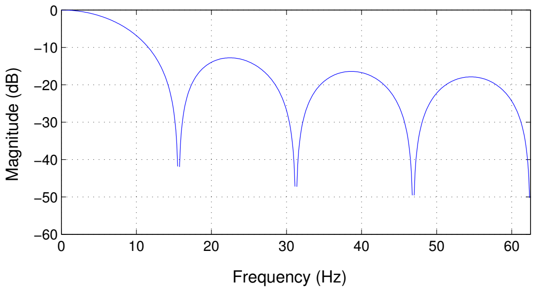

Figure 1: *Magnitude response of a 72ms window (16ms frames with 8ms overlap) FAVG system.*

eration and energy coefficients can be appended to generate the final feature vector for the window.

The optimal SFLW system employed by Huang et al. incorporated non-overlapping windows of 72ms duration, broke each window into 8 frames of 16ms duration with 8ms shift, and averaged the MFCC features from those frames to produce a feature-vector for each window [2]. The 2006 NIST Speaker Recognition Evaluation (SRE) core tests used to evaluate their system contains approximately 5 minutes of training conversation per speaker, and a further five minutes of test conversation [4]. In comparison, TIMIT has an average utterance length of 3.08s. Given that 8 utterances are used to train each speaker, an average of 24.6s of training data is available per speaker. As a result, window overlapping was employed in this study to provide a sufficient number of features to accurately estimate the GMM. In addition, 16ms frames with 8ms shift were employed to allow direct comparison to the SFLW method in [2].

Huang et al. appended delta and acceleration coefficients to the averaged MFCC features [2]. The use of overlapping windows in this study, however, results in high correlation between features from window to window; using a 256ms window with 8ms shift results in adjacent windows having 96.9% of their data in common. This high correlation between features has the effect of reducing the significance of the delta and acceleration coefficients in classification for the proposed methods, although it doesn't affect the SFLW method.

Like the SFLW system, the FAVG method shares similarities with RASTA processing of speech; in which a bandpass filter is applied to frequency channel trajectories [5]. A 72ms FAVG window incorporating 16ms frames with 8ms overlap, for example, effectively applies an FIR low-pass filter with a 9.5Hz cut-off to the filterbank trajectories. Fig. 1 shows the magnitude response (the phase response is linear) of the aforementioned FAVG system.

The two proposed methods will be compared against the optimal SFLW method from [2], and against a baseline STFT, MFCC-based method.

2.3. Speaker Identification

A Gaussian mixture model (GMM) classifier, trained using the expectation maximisation (EM) algorithm, was used in this study for speaker identification [6]. A GMM classifier consists of a linear combination of multivariate Gaussian distributions:

$$p(x|\lambda) = \sum_{k=1}^{K} P_k \mathcal{N}(x|\mu_k, \Sigma_k), \tag{2}$$

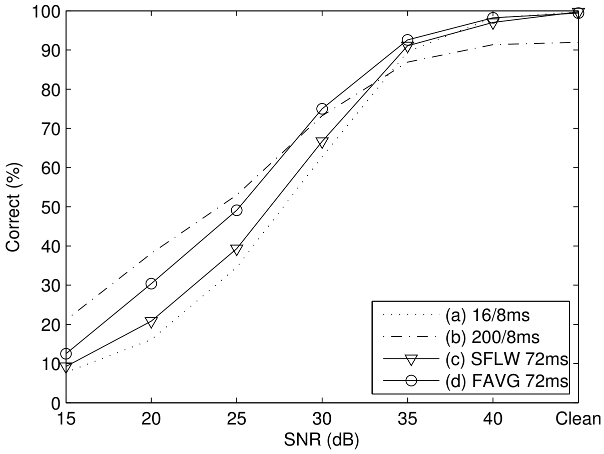

Figure 2: *ASI accuracy with MFCC features. (a) STFT with 16ms frame and 8ms frame-shift; (b) LTFT with 200ms frame and 8ms frame-shift; (c) SFLW with 16ms frame, 8ms frame-shift, averaged over a 72ms window; (d) FAVG with 16ms frame, 8ms frame-shift, averaged over a 72ms window.*

where K is the number of Gaussian components, P_k is the mixing weight of the kth Gaussian component, and:

$$\mathcal{N}(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \mathrm{e}^{\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]}, \tag{3}$$

where μ_k and Σ_k are the mean vector and the covariance matrix, respectively, of the kth Gaussian component. In this study, 32 mixtures with diagonal covariance matrices were used. The speaker identification system was implemented using HTK [7], with only clean utterances used for training. The systems were tested on both clean utterances, and utterances degraded by AWGN at several SNR.

2.4. System Fusion

In order to improve ASI robustness in noisy conditions whilst still retaining high-performance in clean conditions, a fusion of methods was used:

$$\Lambda_s = \sum_{m=1}^{M} \alpha_m \Lambda_m(s), \tag{4}$$

where $\Lambda_m(s)$ is the log-likelihood of speaker s using method m, and α_m is the weighting assigned to method m. In this study, each method was given equal weighting α_m.

3. Results and Discussion

3.1. Performance

The proposed methods were compared to two baselines: the optimal SFLW system proposed by Huang et al. [2], and a baseline MFCC system. The baseline MFCC system and FAVG method incorporate 16ms frames with 8ms overlap – chosen for comparability as it is the base frame size used in the SFLW system.

As previously mentioned, Huang et al. [2] included delta and acceleration coefficients in the feature vector. Their method provides a feature vector for each 16ms frame so, given the 8ms frame overlap, 50% of the speech signal is shared between feature vectors, allowing delta and acceleration coefficients to

Table 1: *ASI accuracy (%) for the various methods using MFCC features. (a) STFT with 16ms frame and 8ms frame-shift; (b) LTFT with 96ms frame and 8ms frame-shift; (c) LTFT with 200ms frame and 8ms frame-shift; (d) SFLW with 16ms frame, 8ms frame-shift, averaged over a 72ms window; (e) SFLW with 16ms frame, 8ms frame-shift, averaged over a 72ms window with delta and accelaration coefficients appended. Please note the inclusion here of the delta and acceleration parameters, as compared to the basic MFCC parameters employed in all other methods. This was included as it was the top-performing method in [2]; (f) FAVG with 16ms frame, 8ms frame-shift, averaged over a 72ms window; (g) FAVG with 128ms frame, 8ms frame-shift, averaged over a 72ms window.*

Method	SNR						
	∞	40	35	30	25	20	15
(a) 16ms frame, 8ms shift	**99.70**	98.21	89.58	62.80	34.52	16.07	7.74
(b) 96ms frame, 8ms shift	98.81	97.02	91.67	73.51	47.92	27.68	11.90
(c) 200ms frame, 8ms shift	91.96	91.37	86.90	73.21	**52.98**	**38.10**	**21.13**
(d) SFLW 72ms	**99.70**	97.02	91.07	66.67	39.29	20.83	9.23
(e) SFLW 72ms (MFCC_DA)	**99.70**	**98.51**	92.26	71.43	42.86	19.35	9.82
(f) FAVG 72ms	99.40	98.21	**92.56**	**75.00**	49.11	30.36	12.50
(g) FAVG 128ms	93.45	91.96	86.31	73.21	51.79	33.04	16.96

Table 2: *ASI accuracy (%) for various fusions using MFCC features. (a) Fusion of STFT and LTFT with 16ms and 256ms frames with 8ms frame-shift; (b) Fusion of FAVG 16ms frame / 8ms frame-shift, averaged over a 72ms window and LTFT with 200ms frame / 8ms frame-shift; (c) Fusion of FAVG 16ms frame / 8ms frame-shift, averaged over a 72ms window and LTFT with 256ms frame / 8ms frame-shift; (d) Fusion of STFT and LTFT with 16ms, 128ms and 256ms frames with 8ms frame-shift; (e) Fusion of STFT and LTFT with 32ms, 64ms, 128ms and 256ms frames with 8ms frame-shift.*

Method	SNR						
	∞	40	35	30	25	20	15
(a) [16+256]/8ms	97.92	95.54	90.77	73.51	51.79	32.14	17.26
(b) FAVG 72ms + 200/8ms	98.21	**96.73**	92.86	77.98	56.85	34.82	18.75
(c) FAVG 72ms + 256/8ms	97.02	95.54	91.67	**80.06**	**58.33**	**36.61**	**19.64**
(d) [16+128+256]/8ms	97.92	95.24	92.56	78.27	55.36	33.33	18.75
(e) [32+64+128+256]/8ms	**99.11**	**96.73**	**93.45**	76.79	54.76	32.44	16.07

provide additional separability. Using LTFT or FAVG, though, results in a single feature vector sharing 96% (given a 200ms frame with 8ms overlap) of the speech signal with adjacent features, reducing the effectiveness of the delta and acceleration coefficients. To allow a fair comparison between methods, both the MFCC and MFCC with delta and acceleration coefficients appended (MFCC_DA) were investigated. Table 1 and Fig. 2 contain the results.

The LTFT and FAVG methods outperform both baselines in low SNR conditions. FAVG outperforms LTFT for 30dB SNR and above, and vice-versa for lower SNR. LTFT outperforms the baseline systems below 35dB – at 20dB SNR, it improves on the baseline MFCC (16/8ms) accuracy by 22.03% (137% relative improvement) and the SFLW system by 17.27% (83% relative improvement). At 15dB SNR, the relative improvement of LTFT over the baseline MFCC and SFLW systems increases to 173% and 129% respectively.

The FAVG method is comparable above 40dB to the baseline systems, with a maximum discrepancy of 0.3%. The FAVG system shows improvement (up to 89% relative) over the MFCC baseline in all but clean conditions, and outperforms the SFLW method below 40dB by a relative improvement of up to 46%.

3.2. Model Variance

The periodogram estimate in Eq. (1) suffers from variance from the ideal spectrum. Bartlett's method, in which spectral estimates from L consecutive segments within a frame are averaged, reduces the variance by a factor of $\frac{1}{L}$ at the expense of spectral resolution [8, p. 974]. Welch's method improves further on this by overlapping the segments [8, pp. 974-977]. In speech processing, spectral estimate variance is generally reduced by combining frequency bins from the same frame according to the Mel-scale, rather than applying Bartlett's or Welch's methods; so, for example, a spectral estimate containing 512 frequency bins might be reduced to 24 Mel-frequency bank parameters. Reduced variance in the spectral estimate should result in reduced variance within the GMM. Given this reduction and since the means of the GMM clusters are unlikely to change significantly, the ratio of inter-speaker to intra-speaker variance should increase.

The inter-class to intra-class variance ratio is a widely used quantitative measure of the discriminability between classes. One such variance ratio metric proposed by Theodoridis and Koutroumbus [9, pp. 280-281] is:

$$J = \frac{|\mathbf{S}_w + \mathbf{S}_b|}{|\mathbf{S}_w|}, \tag{5}$$

where $\mathbf{S}_b$ is the between-class scatter matrix, and $\mathbf{S}_w$ is the within-class scatter matrix. The between-class scatter matrix $\mathbf{S}_b$ is defined as:

$$\mathbf{S}_b = \sum_{c=1}^{C} P_c \left(\boldsymbol{\mu}_c - \boldsymbol{\mu}\right) \left(\boldsymbol{\mu}_c - \boldsymbol{\mu}\right)^t, \tag{6}$$

where C is the total number of classes, $\boldsymbol{\mu}_c$ is the mean for class c, $\boldsymbol{\mu}$ is the global mean, and P_c is the *a priori* probability of class c. The within-class scatter matrix, $\mathbf{S}_w$ from Eq. (5), is defined as:

$$\mathbf{S}_w = \sum_{c=1}^{C} P_c \left(E\left[(\mathbf{x} - \boldsymbol{\mu}_c)(\mathbf{x} - \boldsymbol{\mu}_c)^t \right] \right), \tag{7}$$

where **x** is a feature vector in class c.

As previously mentioned, it is expected that smaller variance within the model should result in wider margins between speakers; thus a second measure, the mean difference in normalised log-likelihood (LLH) between the first and second most-likely speakers, was also included. The log-likelihoods were calculated using 32-mixture GMM trained and tested over the full TIMIT database. The variance ratio metric was calculated using the raw feature vectors from the full TIMIT database.

Table 3: *Model separability by method. (a) SFLW with 16ms frame, 8ms frame-shift, averaged over a 72ms window; (b) FAVG with 16ms frame, 8ms frame-shift, averaged over a 72ms window; (c) LTFT with 96ms frame and 8ms frame-shift; (b) STFT with 16ms frame and 8ms frame-shift.*

Method	Variance ratio metric	LLH margin
(a) SFLW 72ms	8.66×10^5	1.650
(b) FAVG 72ms	8.01×10^5	1.617
(c) 96/8ms	4.14×10^5	1.506
(d) 16/8ms	2.65×10^5	1.265

The calculated measures are shown in Table 3. The longer frame (96ms frame, 8ms shift) shows an increase in separability over the baseline 16/8ms system. This is due to greater numbers of frequency bins averaged in each Mel-frequency bank reducing the spectral estimate variance. By averaging the filter-banks, the FAVG method shows further separability. The greatest separability is evident in the SFLW system. This is to be expected, as the SFLW system averages the final feature vector – rather than an intermediary step in production of the feature vector as does FAVG – which is used directly to estimate the GMM. In addition, a clear increase in mean LLH differences between the two most likely speakers is evident as separability increases.

3.3. Fusion

Table 2 shows ASI accuracy for fusion of various methods, while Fig. 3 compares a FAVG+LTFT fusion with the baseline MFCC system and the top-performing methods described previously in the paper. Fusion (e) in Table 2 performs best for SNR of 35dB and above. It only improves on the other proposed methods, however, at 35dB, while being comparable at higher SNR. Fusion (c) in Table 2 outperforms all other methods examined in this study by approximately 5% for SNR of 25 to 30dB, but is slightly outperformed by the LTFT method with a 200ms window for SNR of 20dB and below.

4. Conclusions

The proposed LTFT and FAVG features show improved robustness to additive white-gaussian noise in automatic speaker identification accuracy, over both baseline MFCC and SFLW features. Furthermore, fusion of the proposed methods through a linear combination of the log-likelihood scores produced further improvements at particular SNR. Thus, features derived from window lengths longer than would generally be considered stationary for speech signals are an effective alternative to the 20-30ms windows currently in widespread use.

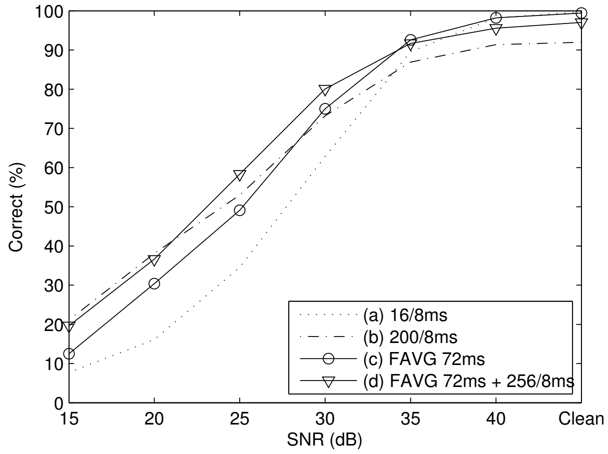

Figure 3: *ASI accuracy for various methods and fusion. (a) STFT with 16ms frame and 8ms frame-shift; (b) LTFT with 200ms frame and 8ms frame-shift; (c) FAVG with 16ms frame, 8ms frame-shift, averaged over a 72ms window; (d) Fusion of FAVG 16ms frame / 8ms frame-shift, averaged over a 72ms window and LTFT with 256ms frame / 8ms frame-shift.*

5. References

[1] B. Gajic and K. K. Paliwal, "Speech parameterization for automatic speech recognition in noisy conditions," *Proc. Norwegian Symp. Signal Processing*, Oct 2001.

[2] C.-L. Huang, B. Ma, C.-H. Wu, B. Mak, and H. Li, "Robust speaker verification using short-time frequency with long-time window and fusion of multi-resolutions," in *Proceedings of Interspeech 2008*, 2008, pp. 1897–1900.

[3] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The darpa speech recognition research database : Specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, feb 1986, pp. 93 – 99.

[4] NIST, "The NIST year 2006 speaker recognition evaluation plan," mar 2006. [Online]. Available: http://www.itl.nist.gov/iad/mig/ /tests/sre/2006/sre-06_evalplan-v9.pdf

[5] H. Hermansky and N. Morgan, "Rasta processing of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 578 –589, oct. 1994.

[6] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, aug 1995.

[7] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[8] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*, 4th ed. Pearson Prentice Hall, 2007.

[9] S. Theodoridis and K. Koutroumbas, *Pattern Recognition (4th Edition)*. Elsevier, 2009.