JFA based Speaker Recognition using Delta-Phase and MFCC features

Ahilan Kanagasundaram, David Dean, Sridha Sridharan

Speech and Audio Research Laboratory
Queensland University of Technology, Brisbane, Australia
{a.kanagasundaram, d.dean, s.sridharan }@qut.edu.au

Abstract

This paper investigates the use of mel-frequency delta-phase (MFDP) features in comparison to, and in fusion with, traditional mel-frequency cepstral coefficient (MFCC) features within joint factor analysis (JFA) speaker verification. MFCC features, commonly used in speaker recognition systems, are derived purely from the magnitude spectrum, with the phase spectrum completely discarded. In this paper, we investigate if features derived from the phase spectrum can provide additional speaker discriminant information to the traditional MFCC approach in a JFA based speaker verification system. Results are presented which provide a comparison of MFCC-only, MFDP-only and score fusion of the two approaches within a JFA speaker verification approach. Based upon the results presented using the NIST 2008 Speaker Recognition Evaluation (SRE) dataset, we believe that, while MFDP features alone cannot compete with MFCC features, MFDP can provide complementary information that result in improved speaker verification performance when both approaches are combined in score fusion, particularly in the case of shorter utterances.

Index Terms: speaker verification, MFCC features, JFA, Delta-phase

1. Introduction

In recent speaker verification research, the joint factor analysis (JFA) technique has become one of more successful approaches to speaker verification by explicitly modelling enrolment and verification mismatch. This approach is typically based upon acoustic features derived from the magnitude spectrum, with most approaches using mel-frequency cepstral coefficients (MFCC) to represent the acoustic domain for modelling against the universal background model (UBM) in forming the speaker and channel factors. While there have been investigations of phase-based features for speaker verification using simple Gaussian mixture model (GMM) [1] and support vector machine (SVM) [2] approaches, no investigation has yet been performed using phase-based features in the explicit channel and speaker modelling approach taken by JFA speaker verification systems.

In order to make use of the phase spectrum for speaker verification, it needs to be transformed into a meaningful representation that provides adequate discrimination between individual speakers. One of the first attempts at using phase-based features for automatic speaker recognition was through the use of modified group delay function (GDF) by Murthy *et al.* [3], defined as the frequency-domain derivative of the phase spectrum, modified to attenuate the effect of zeros in the z-plane of the frequency representation. This approach was shown to outperform MFCC speaker verification using a GMM-UBM modelling approach. [3].

An alternative approach to constructing phase-based features was introduced by Wang *et al.* by looking at the time-domain derivative of the phase spectrum, termed the instantaneous frequency deviation (IFD) [4]. This work was further extended by McCowan *et al.* to develop the mel-frequency delta-phase (MFDP) representation [2] and demonstrated its performance to be similar to that of MFCC using a modern GMM-supervector-based SVM approach, but only without channel compensation. When feature warping and nuisance attribute projection (NAP) were applied to both the MFCC and MFDP systems, the MFDP system was found lacking in comparison to the MFCC. However, even though the channel-compensated MFDP system was not comparable to the MFCC approach individually it was still shown to provide complementary information in fusion with the MFCC, with a score fusion approach outperforming both individual approaches.

In this paper, we study the use of MFDP features introduced by McCowan *et al.* [2], in a modern JFA-based speaker verification system in order to investigate the ability of the explicit speaker and channel modelling to cope with phase-based features. Initially both MFDP and MFCC features will be studied individually within a JFA speaker verification system with a combination of experimental parameters to determine the best individual approach for both sets of features. Thereafter the best configuration of MFDP and MFCC features will be combined to analyze the fused JFA system.

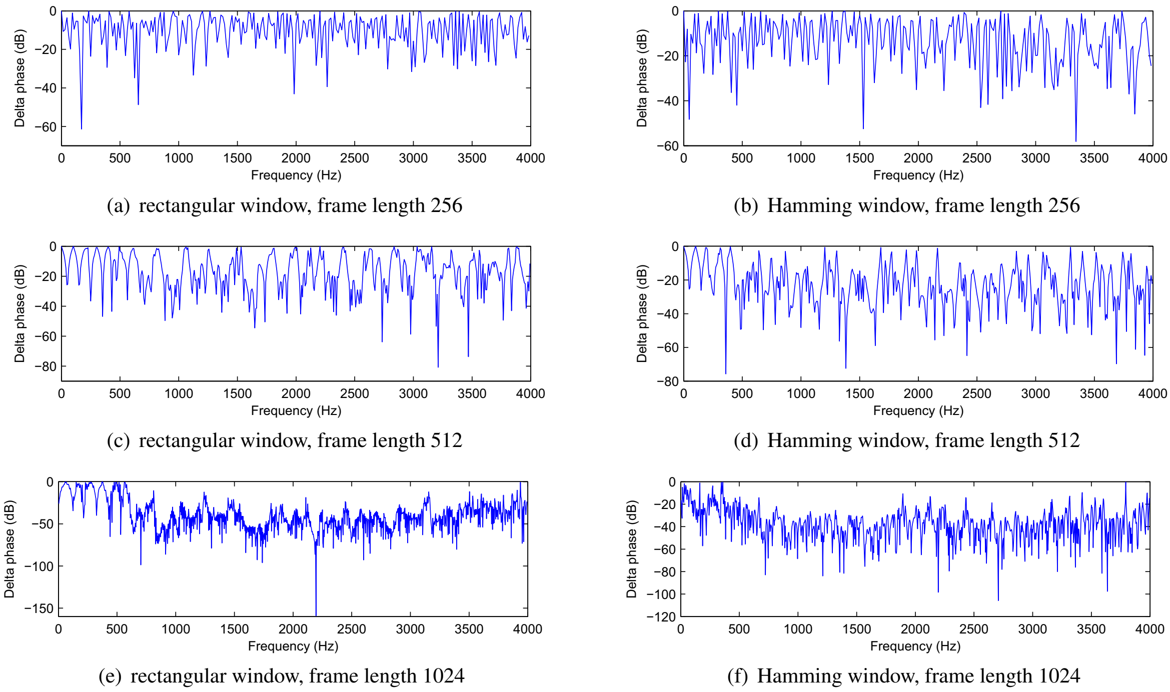
Throughout this paper, both medium length and short utterances will be evaluated to determine if the performance of MFDP features vary according to the amounts of speech available for enrolment and verification. This approach has been taken before for MFCC features in both JFA [5], SVM [6], and i-vector [7] speaker verification systems but no similar studies have been performed on phase-based features.

2. Mel-frequency delta-phase features

The process of extracting MFDP features from the acoustic speech is designed to attempt extract speech information from the phase-domain through calculating a phase difference between successive frames separated by a short time interval [2]. The delta-phase spectrum can be calculated from the Fourier transform of two successive frames $\tilde{X}_m(k)$ and $\tilde{X}_{m-1}(k)$, as follows:

$$\Delta\phi_m(k) = arg\left[\frac{\tilde{X}_m(k)e^{-j\omega_k mD}}{\tilde{X}_{m-1}(k)e^{-j\omega_k(m-1)D}}\right] \tag{1}$$

$$|\Delta\phi_m(k)| = |arg\left[\left(\frac{\tilde{X}_m(k)}{\tilde{X}_{m-1}(k)}\right)e^{-j\omega_k D}\right]| \tag{2}$$



(a) rectangular window, frame length 256

(b) Hamming window, frame length 256

(c) rectangular window, frame length 512

(d) Hamming window, frame length 512

(e) rectangular window, frame length 1024

(f) Hamming window, frame length 1024

Figure 1: *The delta-phase spectrum of a single frame of sample speech file (thbn.wav) from NIST 2008 evaluation, captured at a range of window sizes for both the rectangular and Hamming windowing functions*

where m is the frame index, and D is the number of samples between successive analysis frames.

In order to correctly analyze the delta phase, several factors have to be considered, including the choice of windowing function and the size of the window. Earlier research in phase spectrum has suggested that phase-domain features are generally better segmented using a rectangular windowing function which can provide a higher frequency resolution at a tradeoff against higher spectral leakage, but this spectral leakage does not appear to be as serious a problem as in magnitude-based features [8]. The choice of window-length for phase-based features also appears to generally be much larger than that expected in magnitude-based approaches, with McCowan *et al.* suggesting that higher is generally better [2].

A selection of delta-phase spectrum of a single frame of speech are shown in Figure 1 for both the rectangular and Hamming windowing functions at 256, 512 and 1024 samples/frame.

The process of calculating the MFDP features was modelled closely on a typical MFCC feature extraction process. The acoustic samples were first passed through pre-emphasis filtering, followed by framing using the chosen windowing function and frame size with a small step size resulting in many overlapping frames (particularly for the large frame size of MFDP). Following the windowing, the windows were transferred to the frequency domain for delta-phase calculation, then accumulated using a a triangular filter bank based upon the mel-frequency scale. Finally the output of the filter bank is reduced to the chosen dimensionality by taking the top N features from a DCT transformation.

3. Joint factor analysis

A significant contributor to the degradation of traditional GMM-UBM speaker verification is the presence of session variability between the training and test conditions. One of the more successful approaches to combating this train/test mismatch has been the explicit modelling of speaker and channel factors through JFA. This factor analysis technique introduced by Kenny [9] is based on the decomposition of a speaker-dependent GMM supervector, $\boldsymbol{\mu}$, into separate speaker and channel dependent parts (**S** and **C** respectively):

$$\boldsymbol{\mu} = \mathbf{S} + \mathbf{C}. \tag{3}$$

The speaker dependent and channel dependent components can then be represented by

$$\begin{aligned} \mathbf{S} &= \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z}, & (4)\\ \mathbf{C} &= \mathbf{U}\mathbf{x}. & (5) \end{aligned}$$

In the speaker dependent component, **m** is a session and speaker independent supervector (extracted from a UBM trained on a large development set), **V** is a low rank matrix representing the primary directions of speaker variability, or *eigenvoices*, and **D** is a diagonal matrix modelling the residual variability not captured by the speaker subspace. The speaker factors, **y**, and speaker residuals, **z**, are both independent random vectors having standard normal distributions. Similarly, the channel dependent component contains a low rank matrix, **U**, representing the primary directions of channel variance, or *eigenchannels*, multiplied by the channel factor vector **x**, a normally distributed random vector.

JFA speaker enrolment is performed by calculating the full speaker-dependent GMM supervectors and discarding the channel dependent component. During verification, the channel-dependent component can be estimated directly from the testing utterances, and the entire supervector can be efficiently scored using the linear dot-product approach pioneered by Glembek *et al.* [10].

Table 1: *Comparison of MFDP-based JFA speaker verification performance with different configurations on the common set of the 2008 NIST SRE standard conditions. The best performing systems by both EER and DCF are highlighted within each column.*

(a) frame length 512, rectangular window, with/without feature warping

MFDP features	**short2-short3**		**10sec-10sec**	
	EER	**DCF**	**EER**	**DCF**
feature warping	11.19%	0.0505	31.04%	0.0972
no feature warping	**10.64%**	**0.0479**	**27.56%**	**0.0961**

(b) frame length 512, feature warping, rectangular vs. Hamming window

MFDP features	**short2-short3**		**10sec-10sec**	
	EER	**DCF**	**EER**	**DCF**
rectangular window	**11.19%**	**0.0505**	**31.04%**	**0.0972**
Hamming window	14.00%	0.0611	**31.04%**	0.0999

(c) rectangular window, without feature warping, different frame lengths

MFDP features	**short2-short3**		**10sec-10sec**	
	EER	**DCF**	**EER**	**DCF**
frame length 256	17.13%	0.0676	32.36%	0.0987
frame length 512	10.64%	0.0479	27.56%	0.0961
frame length 1024	**9.32%**	**0.0406**	**26.06%**	**0.0950**

Table 2: *Comparison of MFCC-based JFA speaker verification performance with different configurations on the common set of the NIST 2008 SRE standard conditions. The best performing systems by both EER and DCF are highlighted within each column.*

(a) frame length 256, Hamming window, with/without feature warping

MFCC features	**short2-short3**		**10sec-10sec**	
	EER	**DCF**	**EER**	**DCF**
feature warping	**3.37%**	**0.0149**	**16.69%**	**0.0686**
no feature warping	4.94%	0.0197	17.86%	0.0701

(b) frame length 256, with feature warping, rectangular vs. Hamming window

MFCC features	**short2-short3**		**10sec-10sec**	
	EER	**DCF**	**EER**	**DCF**
rectangular window	**3.05%**	**0.0148**	**16.54%**	**0.0679**
Hamming window	3.37%	0.0149	16.69%	0.0686

(c) hamming window, with feature warping, and different frame-length

MFCC features	**short2-short3**		**10sec-10sec**	
	EER	**DCF**	**EER**	**DCF**
frame length 256	3.37%	0.0149	**16.69%**	**0.0686**
frame length 512	**3.29%**	**0.0146**	16.99%	0.0732
frame length 1024	3.31%	0.0164	21.22%	0.0806

4. Methodology

Both the MFCC and MFDP JFA systems were developed and evaluated identically, with the choice of feature extraction technique being the only difference between the two systems.

For both MFCC and MFDP feature extraction, 13 coefficients were captured for each frame with appended delta coefficients. Two gender dependent GMMs containing 512 Gaussians are used throughout our experiments to represent the UBM, and were trained on the NIST 2004 SRE corpus. Speaker and session variability subspaces of dimension $R_y = 400$ and $R_x = 100$ are applied for JFA experiments. These speaker and channel variability subspaces were also trained on the NIST 2004 SRE, as well as data from the NIST 2005 SRE and Switchboard II.

Enrolment and verification of the MFDP and MFCC systems were performed using NIST 2008 SRE telephone utterances from the *short2-short3*, and *10sec-10sec* conditions. These conditions were chosen to allow the performance of MFDP speaker representations in both a typical utterance length of around 2 minutes (*short2-short3*) and the more difficult task of enroling and verifying speakers in only 10 seconds (*10sec-10sec*).

In addition to investigating the MFDP and MFCC JFA systems individually, this paper will also investigate the performance of both systems combined in a score fusion configuration to determine if there is complementary information in the two speaker representations. For this paper a simple weighted fusion approach will be taken where the final score of combined system, $S_{combined}$, will be calculated as follows,

$$S_{combined} = \alpha * S_{MFCC} + (1 - \alpha) * S_{MFDP} \qquad (6)$$

where S_{MFCC} and S_{MFDP} are the individual systems scores for a given utterance, and α is a weighting coefficient, chosen prior to evaluation.

5. Results and discussions

5.1. Individual performance

The performance of the MFDP and MFCC-based JFA speaker verification on the common set of the NIST SRE 2008 *short2-short3* and *10sec-10sec* conditions are shown over a range of configurations in Tables 1 and 2 respectively.

It can be seen by comparing Tables 1(a) and 2(a), that while feature warping provides an advantage for MFCC extraction as is commonly known [11], no similar effect is found for MFDP. This result is similar to that found by McCowan *et al.* [2] for SVM, where feature warping did not provide a large improvement for SVM-based MFDP speaker verification, although in their application it did still provide a small improvement.

A comparison of Tables 1(b) and 2(b) shows that the choice of a rectangular windowing function provides a clear improvement for MFDP extraction, while the difference is unlikely to be significant for MFCC. This is inline with previous findings that rectangular windowing functions are more suitable for phase-based speech feature extraction as the effects of spectral leaking is less of in an issue than the loss of resolution in the frequency caused by Hamming windowing [8].

Finally, looking at the performance change for different window lengths in Tables 1(c) and 2(c), it can be seen that the MFDP system performs better with longer frame lengths than normally used in MFCC-based feature extraction.

One interesting comparison that can be drawn from these results is that regardless of the choice of configuration, the MFCC features provide a considerable increase in performance over the MFDP features, with the best MFDP system having an EER almost three times as large (9.32%) as the best MFCC

Table 3: *Comparison of score fusion speaker verification performance on the common set of the NIST 2008 SRE standard conditions, as the fusion parameter $alpha$ is varied between 0.0 (MFDP only) and 1.0 (MFCC only). The best performing systems by both EER and DCF are highlighted within each column.*

Fused system	**short2-short3**		**10sec-10sec**	
	EER	DCF	EER	DCF
α = 0.0 (MFDP only)	9.32%	0.0406	26.06%	0.0950
α = 0.1	4.86%	0.0210	21.21%	0.0858
α = 0.2	3.64%	0.0157	17.42%	0.0754
α = 0.3	3.29%	0.0148	15.84%	0.0700
α = 0.4	3.21%	0.0141	14.96%	0.0664
α = 0.5	3.21%	**0.0139**	**14.82%**	0.0659
α = 0.6	**2.96%**	0.0144	15.09%	**0.0650**
α = 0.7	2.98%	0.0145	15.67%	0.0656
α = 0.8	3.03%	0.0146	15.81%	0.0656
α = 0.9	2.96%	0.0146	16.25%	0.0665
α = 1.0 (MFCC only)	3.05%	0.0148	16.54%	0.0679

approach (3.31%). This is particularly interesting in contrast to the performance of MFDP matching the MFCC in the GMM supervector SVM approach of McCowan *et al.* [2], at least prior to channel compensation. From this comparison, it appears that the MFDP features are not well suited to the explicit speaker and session modelling approach of the JFA.

5.2. Fusion performance

While the previous set of experiments demonstrated that the MFDP features do not perform well in comparison to traditional MFCC features for JFA-based speaker verification, it is quite possible that both features can work together to provide complementary information in a fusion configuration. Indeed, this was found to be the case in McCowan *et al.*'s SVM supervector based system [2]. We wish to investigate in this paper if similar complementary performance can be demonstrated in the JFA speaker verification framework.

In order to test this, a simple weighted score fusion system was set up with the two best performing JFA systems: no feature warping, rectangular window of length 1024 for MFDP; and feature warping, rectangular window of length 256 for MFCC. The output scores of these two systems were then fused with a weighting parameter α, that can vary from 0.0 (MFDP only) to 1.0 (MFCC only). The results of these experiments are shown in Table 3.

From these results, we can see that while the MFDP-based JFA speaker verification system is outperformed in all conditions by the MFCC-based approach, the fusion of the two systems can provide better performance than the MFCC-based approach in both utterance lengths under evaluation. This effect is particularly the case for the shorter *10sec-10sec* condition, suggesting that making use of phase information may allow for the extraction of complementary information in short-utterance speaker verification.

6. Conclusion

In this paper, we investigated the use of MFDP coefficients in comparison to, and in fusion with, traditional MFCC features within the JFA speaker verification framework. We found that, while MFDP features could be shown to work reasonably well for the task of JFA-based speaker verification, they were not competitive with traditional MFCC features, which provided a third the EER of MFDP in best conditions. However, we did find that in score fusion between MFDP- and MFCC-based JFA speaker verification systems, the MFDP features could provide complementary information to the MFCC, resulting in improved performance in both the *short2-short3* and *10sec-10sec* conditions on the common set of the NIST SRE 2008 evaluation data. We found this to be particularly the case for the *10sec-10sec* condition, suggesting that there may be fruitful use for phase-based features in providing complementary information to improve the performance of short utterance speaker verification.

7. Acknowledgements

This project was supported by the Cooperative Research Centre for Advanced Automotive Technologies (AutoCRC).

8. References

[1] T. Thiruvaran, E. Ambikairajah, and M. Epps, "Group delay features for speaker recognition," in *Information, Communications & Signal Processing, 2007*, pp. 1–5, IEEE, 2007.

[2] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum with application to voice activity detection and speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, no. 99, pp. 1–1, 2011.

[3] K. Murthy and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *Signal Processing Letters, IEEE*, vol. 13, no. 1, pp. 52–55, 2006.

[4] Y. Wang, J. Hansen, G. Allu, and R. Kumaresan, "Average instantaneous frequency (AIF) and average log-envelopes (ALE) for ASR with the AURORA 2 database," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[5] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech 2008*, (Brisbane, Australia), pp. 853–856, September 2008.

[6] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "Experiments in SVM-based speaker verification using short utterances," in *Proc. Odyssey Workshop*, pp. 83–90, 2010.

[7] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector based speaker recognition on short utterances," in *Interspeech 2011*, pp. 2341–2344, 2011.

[8] L. Alsteris and K. Paliwal, "Importance of window shape for phase-only reconstruction of speech," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04)*, vol. 1, pp. I–573, IEEE, 2004.

[9] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[10] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*, pp. 4057 –4060, April 2009.

[11] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, pp. 213–218, ISCA, 2001.