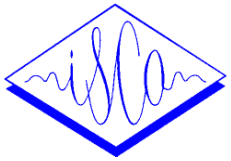

2001: A Speaker Odyssey
The Speaker Recognition Workshop
Crete, Greece
June 18–22, 2001

Feature Warping for Robust Speaker Verification

Jason Pelecanos, Sridha Sridharan

Speech Research Laboratory, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, AUSTRALIA

j.pelecanos@qut.edu.au s.sridharan@qut.edu.au

Abstract

We propose a novel feature mapping approach that is robust to channel mismatch, additive noise and to some extent, non-linear effects attributed to handset transducers. These adverse effects can distort the short-term distribution of the speech features. Some methods have addressed this issue by conditioning the variance of the distribution, but not to the extent of conforming the speech statistics to a target distribution. The proposed target mapping method warps the distribution of a cepstral feature stream to a standardised distribution over a specified time interval.

We evaluate a number of the enhancement methods for speaker verification, and compare them against a Gaussian target mapping implementation. Results indicate improvements of the warping technique over a number of methods such as Cepstral Mean Subtraction (CMS), modulation spectrum processing, and short-term windowed CMS and variance normalisation. This technique is a suitable feature post-processing method that may be combined with other techniques to enhance speaker recognition robustness under adverse conditions.

1. Introduction

In speaker verification applications, there is a need to extract information from speech that is speaker specific and robust to noise and various channel and transducer effects. Previously, a number of methods for reducing these effects was proposed. Cepstral Mean Subtraction [1] was applied to remove linear channel effects and handset mapping techniques [2] were examined to reduce mismatch between types of telephone handsets. Modulation spectrum analysis [3, 4] was also used to reduce a number of these transducer and transmission channel effects, but with limited robustness to additive noise.

For clean speech and matched conditions, there are a number of good performing features that are based on a cepstral analysis. Under mismatched conditions these basic features will become corrupted. To compensate for linear channel variations, Cepstral Mean Subtraction [1] was noted as a promising approach. However, under additive noise conditions, the feature estimates degrade significantly. An extension of this approach, for speech recognition [5] and to some extent, for speaker verification [6], was proposed and involved normalising the distribution of single cepstral features (over some specific window length) by subtracting their mean and scaling by their standard deviation. For speech recognition, it was found that this approach improved noise robustness and the effects of varied channels by forcing a consistent mean and spread of the individual cepstral features. For the speaker verification paper, normalisation was applied over the whole utterance or over a relatively small window of one second or less. This either limited the robustness to noise variations by having a long normalisation window, or reduced the resolution and response of the channel compensation portion by use of a relatively shorter window (of approximately 250-1000ms in length). A recent approach [7] successfully examined the use of a neural network structure to perform a non-linear mapping of (mean-removed) cepstral features to establish an improved parameterisation for telephone network speaker recognition. The neural network was trained to discriminate speakers by modeling speech data from speakers recorded over different handsets. The robustness of this approach across different recording environments other than what the network was trained for is currently not investigated.

An important tradeoff for speaker features exists between the quantity of unreliable information that can be removed from the speaker features versus the speaker specific information that can be preserved, to achieve optimal recognition. Thus, for clean speech using the same microphone for recordings, many of these enhancement techniques may actually reduce system performance [8].

An application of interest for using more robust features is automatic speaker verification over telephone networks. This typically requires a number of feature enhancement techniques. We propose a method that is robust to linear channel effects and slowly varying additive noise. This is achieved by warping each cepstral feature stream over a specified time interval to match a specific target distribution. Typically, the true distribution of a single feature is not of single mode. To accommodate this, the source features may be mapped to an ideal distribution of some form that may consist of multi-modal components. In this paper, we will limit the analysis to single mode mapping, although it is expected to limit performance.

The remainder of this paper discusses the cepstral feature enhancement techniques to be examined and their robustness to adverse conditions. We then propose the non-linear feature mapping technique followed by an analysis of the warping method which indicates how the features are more robust to linear channel effects and additive noise. The normalised features are then evaluated on the NIST 1999 telephone speech corpus using a state-of-the-art speaker recognition system.

2. Robust feature enhancements

In this section we identify a number of standard techniques used to improve the robustness of cepstral features to channel effects for speaker verification over telephone networks.

Cepstral Mean Subtraction (CMS) [1] was one of the ear-

lier and more effective methods of compensating cepstral features for linear channel induced effects in speech. This method removes linear channel effects by removing the mean cepstral coefficient (for speech) from each feature over the duration of the utterance. The effect of the linear channel may be reduced by this method, but the average vocal tract configuration information pertaining to the speaker is also lost. It was indicated that using CMS on clean speech with matched transducer and channel conditions degraded performance [8]. However, under different channel environments, mean subtraction can improve the core cepstral parameters significantly.

There is another class of feature processing that extracts relevant information from the modulation spectrum. Some useful feature processing techniques are (RelAtive SpecTrA) RASTA [4] and a number of other related modulation spectrum analysis methods. One approach is to filter the time-trajectories of the filterbank log-energies to remove the less useful components. For robust recognition performance, it was also noted in [3] that the DC component of the log filterbank energies was less useful for performing recognition. This relates to the success that CMS has attained which is achieved by performing mean cepstral feature removal. It was determined that the standard RASTA processing algorithm was suitable for speech recognition, but when applied to speaker verification, the specified lower cut-off frequency removed significant portions of speaker specific information. Followup investigations by these authors indicated that important speaker specific information is present at modulation frequencies down to 0.1Hz. An alternative modulation spectrum processing approach [3] was implemented with a 100 point Finite Impulse Response (FIR) Filter with a resolution of 0.5Hz. An improvement was found by using the 100 point filter to attenuate the upper modulation frequencies above 10Hz, and mean subtraction to remove the DC component of the features.

Another approach to address real-time applications of speaker verification is based on Cepstral Mean Subtraction over a relatively shortened window [9]. This window was longer than mentioned in [3] (300 points at a frame rate of 100Hz), to improve the lower modulation frequency cut-off band. There are fast implementations available for performing channel compensation over a sliding window.

An interesting extension of sliding mean removal is that of normalising the speech features according to the mean and standard deviation of the features within the current sliding window interval [5, 6]. Our experiments in Section 5 indicate significant improvements over the mean cepstral subtraction approach using a three second window. This method exploits the effect that additive noise will tend to reduce the variance of the cepstral feature parameters.

3. Feature warping

The aim of feature warping is to construct a more robust representation of the each cepstral feature distribution. This is achieved by conditioning and conforming the individual cepstral feature streams such that they follow a specific target distribution over a window of speech frames. We introduce the basic concept of the warping process, which is followed by a derivation that indicates how warping can benefit verification systems running in both additive noise and mismatched channel environments. This is followed by a proposed solution to implementing the general method with an overview of a specific form of its implementation.

The basic structure of how warping integrates into the parameterisation process is identified in Figure 1. The process begins by deriving the complete set of cepstral coefficients from the speech segment. Each cepstral coefficient is then analyzed independently as a separate feature stream over time for use in the warping process. A (typically three second) window of features is extracted from the feature stream and processed in the warping algorithm to determine a mapped feature for the initial cepstral feature in the middle of the window. The sliding window is shifted by a single frame each time and the analysis is repeated.

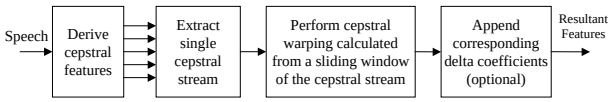

Figure 1: *Block diagram of the parameterisation process.*

For speech, the true distribution of a feature is speaker dependent and multi-modal in nature. However, various channel and additive noise influences can corrupt this distribution. We aim to perform a mapping that will condition the feature distribution. To simplify the mapping we assume that the target speaker features conform to a particular distribution type. Figure 2 indicates the mapping approach. Intuitively, this method compensates in part for the linear channel in that the short-term mean is removed, and attempts to conform the distributive shape and spread to limit additive noise effects. This warping method applied to cepstral features in speech, is similar in concept to performing histogram equalization of picture pixel intensities commonly used for image analysis.

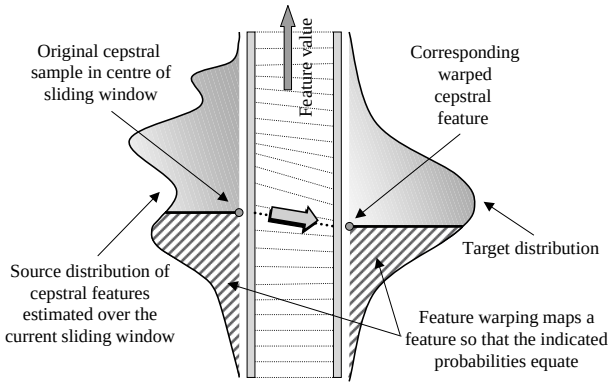

Figure 2: *Warping of features according to a target distribution shape.*

3.1. Additive noise and linear channels effects

We now present an analysis of the effects that combined additive noise and linear channel mismatch can have on cepstral based features to provide an insight into how feature warping can improve speaker recognition robustness.

The effect of the channel and additive noise on cepstral features may be observed by analysis of the filterbank energies to establish the Mel-Frequency Cepstral Coefficients (MFCCs) [10] for each speech frame. The noise corrupted log-energy $\log(E_k)$, for filterbank k, may be specified by the composition of linear channel $C(i)$ and additive noise $N(i)$ effects at each

frequency index i, determined by a complex Discrete Fourier Transform (DFT) representation. (The clean signal is symbolised by $S(i)$.) Each filterbank is assumed to be rectangular in its filtering response, with discrete frequency indexes M_{S_k} and M_{F_k} to indicate the start and finish frequency indexes for each filterbank.

$$\log(E_k) = log\left\{\sum_{i=M_{S_k}}^{M_{F_k}} |(S(i)+N(i))C(i)|^2\right\} \quad (1)$$

The channel effect may be isolated given the assumption that the channel effect is consistent over the frequency range of the filterbank. (ie. $C(M_{S_k}) \approx C(M_{S_k}+1) \approx ... \approx C(M_{F_k}) \approx C_k$) If the real and imaginary components of the speech and noise are also independent, the filterbank energy may be approximated as follows.

$$\log(E_k) \approx 2\log C_k + \log\left\{\sum_{i=M_{S_k}}^{M_{F_k}} (|S(i)|^2 + |N(i)|^2)\right\} \quad (2)$$

Let the filterbank energies be represented by $FBE_S(k)$ and $FBE_N(k)$ for the speech and noise respectively for filterbank k.

$$\log(E_k) \approx 2\log C_k + \log\{FBE_S(k) + FBE_N(k)\} \quad (3)$$

If this derivation is extended to cepstral features, and in particular MFCCs, the cepstral features become a weighted combination of the actual filterbank log-energies via the cosine transform. This may be generalized by specifying the weights as $a_1, a_2, ..., a_K$ for K filterbanks. Thus a resulting MFCC may be represented by the approximation in Equation 4.

$$\begin{aligned} &2a_1\log C_1 + a_1\log\{FBE_S(1)+FBE_N(1)\} + \\ &2a_2\log C_2 + a_2\log\{FBE_S(2)+FBE_N(2)\} + \ldots \end{aligned} \quad (4)$$

The linear channel effect may be attenuated by subtracting the mean cepstral coefficient over the current sliding window, provided the window provides sufficient frequency resolution. By observing Equation 4, a slowly changing additive noise power component generally has the effect of reducing the variance and distorting the distribution of the features. In addition, the effect of the additive noise changes the form of the distribution and can skew the shape. This is in part due to its variation in short-term signal-to-noise ratio. This is where feature warping is capable of conditioning the feature distribution. Another advantage, is that no signal to noise power estimate is required. The warping analysis in essence will map the upper percentile of the source distribution to the upper portion of the target distribution to limit the distribution skew caused by noise.

3.2. General implementation

This section describes the approach for implementing feature warping. The goal of feature warping is to map the observed cepstral feature distribution over a specified speech interval so that the accumulated distribution is similar to a target distribution. Thus, the first phase of feature warping is to select a target distribution to map the cepstral coefficients to. Since speech is multi-modal in nature, the ideal target distribution would also be multi-modal and representative of the speaker's true feature distribution. For this preliminary test, we examine only the single mode mapping.

Once a suitable target distribution is selected, the parameterisation process can begin. The speech is parameterised using cepstral coefficients. Each cepstral feature is then treated independently as its own stream of features. A window of N features in the feature stream is isolated and their values are sorted in descending order. To determine a single warped feature element given the cepstral feature that exists in the centre of the current sliding window, the ranking of the cepstral feature within the sorted list is calculated. (The most positive feature value obtains a ranking of 1 while the most negative a ranking of N.) This ranking is used as an index in a lookup table to determine the warped feature value. The lookup table is devised so as to map a rank order determined from the sorted cepstral feature elements to a warped feature using the desired warping target distribution. The process is repeated for each frame shift of the sliding window. Corresponding delta coefficients may also be calculated at this point.

The lookup table used to perform the mapping is calculated prior to the parameterisation process using Equation 5, with a generic target density function for a single warped feature stream variable z, given by $h(z)$. Thus, given an N point analysis window, and the rank R of the middle speech feature in the current window, the general mapping function to match a target distribution $h(z)$ may be calculated. The lookup table (or warped feature) may be determined by finding m.

$$\frac{N+\frac{1}{2}-R}{N} = \int_{z=-\infty}^{m} h(z)\ dz \quad (5)$$

Computationally, this may be achieved by setting the rank initially to $R = N$, solving for m by numerical integration, and repeating for each decremented value of R.

Note that the continuous form of Equation 5 that directly maps a source cepstral feature q (with measured distribution $f(y)$) to the warped component m (with distribution $h(z)$) is given by Equation 6.

$$\int_{y=-\infty}^{q} f(y)\ dy = \int_{z=-\infty}^{m} h(z)\ dz \quad (6)$$

This mapping approach may be considered as recognizing the *relative* positions of each of the features as more important rather than their *absolute* feature values.

3.3. Normal distribution warping

In this section we examine the method of mapping features to match a normal target distribution. The warping is to map a feature stream to a standard normal distribution, $h(z)$. As mentioned in Section 3.2, cepstral features are multi-modal in nature and for optimal performance should be represented in a multi-modal fashion. In this investigation we examine the simplest of mappings, that of mapping to a normal distribution. Consequently, suboptimal performance due to this simplification is expected. The distribution of a normal curve is given.

$$h(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right) \quad (7)$$

The aim is to conform the distribution of the feature elements (over a sliding window) to a particular form such that the resulting feature distributions may be made more consistent across recording environments.

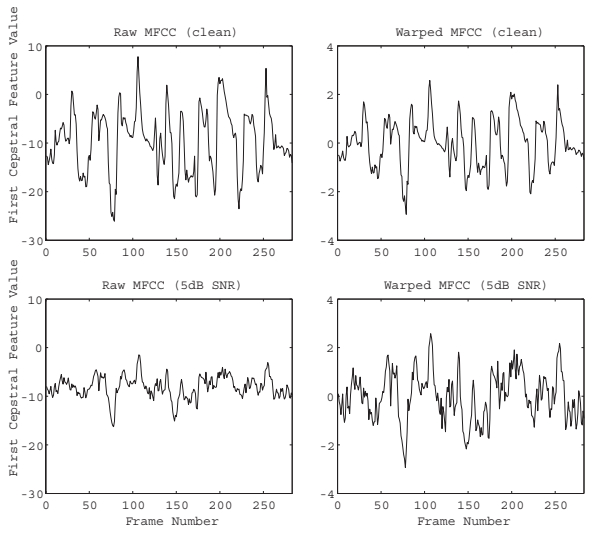

Figure 3: *Effect of additive noise on raw and warped cepstral features.*

Figure 3 indicates the effect of additive noise (simulated by adding extracts of office noise from the NOISEX speech and noise database) on both the raw cepstral features and their warped versions using a Gaussian target mapping. It is observed that the variance of the noise affected raw cepstral features generally decreases with increasing noise level. The warping remaps the distribution to improve its shape, scale and positioning (see Figure 4).

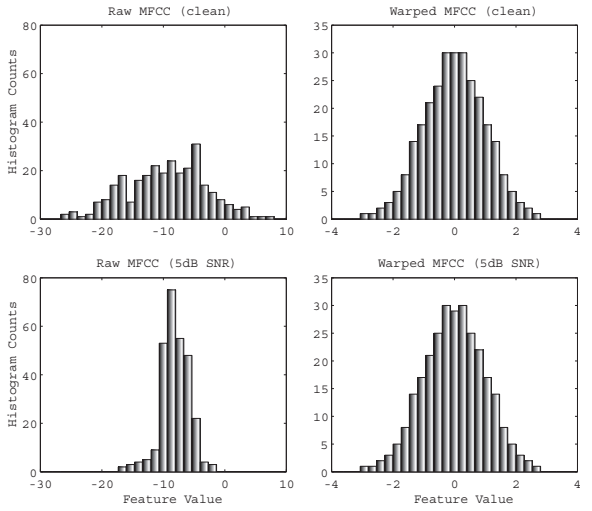

Figure 4: *Histogram of raw and warped cepstral features with and without additive noise (derived from the corresponding data used for Figure 3).*

4. Speaker modeling and classification

We discuss here the basic speaker verification system used for evaluating the different feature conditioning methods. This system consists of a Gaussian Mixture Modeling (GMM) adaptation core to establish a speaker model for which various normalisation techniques will be applied. The GMM models the distribution of the features derived from the parameterisation phase.

4.1. Parameterisation

The parameters typically used for speaker verification are cepstral based. In this system, Mel-Frequency Cepstral Coefficients [10] are derived from mel-spaced filterbank log-energies. There are 20 triangular filterbanks spanning the bandlimited region with 12 MFCCs derived from these, using 32ms frames and a 10ms frame advance. Delta coefficients were also appended. For the telephone speech evaluation, the speech was bandlimited to 300-3200Hz.

4.2. Gaussian mixture modeling

Speaker training involves a two step process; a general Universal Background Model (UBM) [11] is trained on a large quantity of exclusive speech, and a target speaker model is then adapted from the gender specific UBM. The UBM is comprised of a Gaussian Mixture Model formed from the contribution of a large number of component mixtures determined from modeling a vast quantity of speech recorded from numerous speakers.

Gaussian Mixture Modeling is used for modeling the Probability Density Function (PDF) of a multi-dimensional feature vector. A GMM forms a continuous density estimate of the PDF of a multi-variate parameter by the additive composition of multi-dimensional Gaussians. Given a single speech feature vector $\vec{x}$, of dimension D, the probability density of $\vec{x}$ given an M Gaussian mixture speaker model λ, is given by

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i g(\vec{x}, \vec{\mu}_i, \Sigma_i) \tag{8}$$

with a single Gaussian component density given as

$$g(\vec{x}, \vec{\mu}_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_i|}} \times \exp(-\frac{1}{2}(\vec{x} - \vec{\mu_i})' (\Sigma_i)^{-1} (\vec{x} - \vec{\mu_i})) \tag{9}$$

where there is the additional constraint of $\sum_{i=1}^{M} w_i = 1$ and $(\cdot)'$ represents the matrix transpose operator.

The UBM (comprised of 512 mixtures) is trained using the Expectation-Maximisation Algorithm with model seeding performed with a vector quantization pre-estimate [12].

Once the UBM is obtained, a speaker model may be established by adjusting the UBM parameters by Bayesian Adaptation [11]. Instead of maximising the likelihood of the limited quantity of training speech, the Maximum Likelihood of the information from the prior UBM with the new speech (with different weightings applying to both) is determined to form the posterior distribution model for that speaker. An adaptation weight is used to describe the proportion of information from the prior UBM distribution and the new data estimates that should be contained in the posterior distribution. The mixture

means, weights and variances were adapted in this implementation. However, adapting only the mixture means will typically improve results further.

4.3. Speaker classification

Speaker classification is achieved by testing the candidate speech segment against the adapted target speaker model and the UBM (in the form of a likelihood ratio test) and comparing this figure against a threshold. The means by which speaker scoring is performed is by calculating the expected frame-based log-likelihood ratio of the target speaker model versus the Universal Background Model. This involves the calculation of both a target speaker likelihood and the background speaker likelihood.

Thus, given a set of T independent and identically distributed feature vectors $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$, the log-likelihood of X given a speaker model λ, is determined as

$$\log p(X|\lambda) = \sum_{t=1}^{T} \log p(\vec{x}_t|\lambda) \tag{10}$$

For a time normalised score, the expected frame-based log-likelihood may be found. This normalisation alleviates some of the problems associated with the assumption of observation independence. For test trials, the set of speech feature vectors, X, is tested against both the adapted target model, λ_{tar}, and the UBM, λ_{ubm}, to determine an expected frame-based log-likelihood ratio score.

$$\Lambda = \frac{1}{T} \sum_{t=1}^{T} \left(\log p(\vec{x}_t|\lambda_{tar}) - \log p(\vec{x}_t|\lambda_{ubm}) \right) \tag{11}$$

This score for the comparison of a test segment with its target and background models is used for the basic speaker hypothesis test. Other normalisations may also be applied to the likelihood ratio statistic to compensate for various recording mismatches.

4.4. Speaker normalisation

A number of methods are used to improve the recognition performance under different recording contexts and conditions. Some of these methods focus on handset compensation and test segment normalisation. Handset compensation in this system is based on the Handset Normalisation (H-Norm) approach [11]. This compensation approach normalises a test segment score according to the derived handset class of the test utterance. Two handset classes were specified; carbon and electret. During the target speaker training process, the mean and standard deviation of the scores from a large number of standard handset test speech segments were recorded for both the carbon and electret scenarios. These two sets of statistics were used for normalising the test segment score.

An improvement on the likelihood ratio test combined with H-Norm is the inclusion of Test segment Normalisation (T-Norm) [13]. T-Norm uses the scores derived from testing the utterance against a set of standard models to adjust the target speaker score. This score is also normalised by the mean and standard deviation.

Both the H-Norm and T-Norm approaches are used in the following telephony system experiments.

5. Results

In this study, we examine the application of feature warping to speaker verification over telephone networks. We compare the methods identified earlier in Section 2 to observe their performance in adverse environments. For this evaluation, the NIST 1999 Speaker Recognition Evaluation database was used. (For further information see [14]). This database includes a collection of 230 male and 309 female target speakers, each providing approximately two minutes of training speech. There are 1448 male and 1972 female test segments of up to one minute in length.

The Detection Error Trade-off performance criteria [15] was selected to grade the systems examined for the evaluation. It represents the tradeoff error probability between false alarms and missed target speaker decisions.

In Figure 5 there are three separate classes of evaluations for each feature processing technique. The scores from the tests are partitioned according to if the target speaker used the same handset or telephone number in the training conversation as that used for the test call. The three criteria from the poorest to best performing include; same number and telephone, same number but different telephone, and different number and telephone.

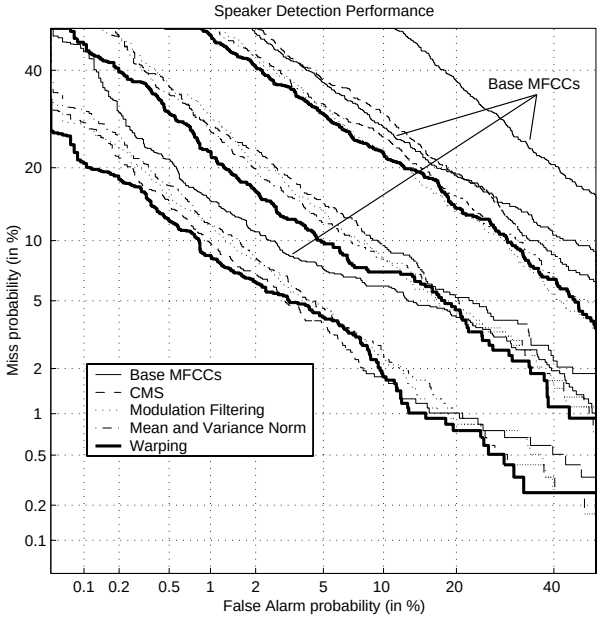

Figure 5: *The NIST 1999 DET curve performances for different feature channel compensation methods.*

Figure 5 compares a number of feature enhancements evaluated for a number of different phone number and handset conditions. For clarity, only the more contrasting techniques mentioned earlier are shown. (Sliding window mean removal was not included due to it having similar results to the segment length cepstral mean subtraction method).

As indicated, the warping approach over the majority of the error curves gives the best result, with significant improvements in error at low false alarm probabilities. The plots indicate that the warping method performs reliably all cases in contrast to the other algorithms. The warping approach is generally the better performing for each of the three classes of tests, whereas

a number of enhancement methods achieve improved results for one test and relatively degraded results for the others.

For the combined result using *all tests*, modulation spectrum processing is comparable in performance to the mean and variance normalisation on the network corpus (with the best result obtained using warping). Thus, modulation processing can be suitable for improving channel and handset mismatch, but it is also sensitive to additive noise. Feature warping and mean and variance normalisation are more robust to such effects.

Worthy of note is the degradation caused by not including linear channel compensation for telephony recordings. At 20% false alarm probability for the different number and handset condition, the basic MFCCs have almost twice the error rate of the next poorest performing method.

In addition to this experiment, we performed an evaluation of the effect of adapting only the mixture means in contrast to adapting all model parameters. The overall equal error rate was reduced further from 9.4% to 8.3%. This result indicates the importance of restricting model adaptation parameters given limited adaptation training data, and the broad suitability of feature warping to various classifier configurations.

6. Conclusion

It was found that cepstral based feature vector warping using a Gaussian target distribution is an effective method of reducing the effects of mismatch. Under adverse conditions, performance can be enhanced by conditioning the short term distribution of the individual cepstral features to a standardised distribution. The robustness of the feature processing to slowly changing additive noise characteristics and linear channel effects are attractive traits. The proposed implementation also permits its use in real time applications. It is pointed out that improved warping may be achieved by selection of a more appropriate target distribution that may also be speaker specific. The additional advantage of the approach is that it may be cascaded with other feature enhancement techniques such as some forms of modulation spectrum processing and non-linear neural network mapping approaches.

7. Acknowledgements

This work was supported by a research contract from the Australian Defence Science and Technology Organisation (DSTO).

8. References

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, pp. 254–272, 1981.

[2] T. Quatieri, D. Reynolds, and G. O'Leary, "Magnitude-only estimation of handset nonlinearity with application to speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 745–748, 1998.

[3] S. van Vuuren and H. Hermanksy, "On the importance of components of the modulation spectrum for speaker verification," in *International Conference on Spoken Language Processing*, vol. 7, pp. 3205–3208, 1998.

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

[5] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.

[6] J. Koolwaaij and L. Boves, "Local normalization and delayed decision making in speaker detection and tracking," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 113–132, 2000.

[7] L. Heck, Y. Konig, M. Sönmez, and M. Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," *Speech Communication*, vol. 31, no. 2-3, pp. 181–192, 2000.

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

[9] J. Pelecanos, S. Myers, and S. Sridharan, "Rapid channel compensation for one and two speaker detection in the NIST 2000 speaker recognition evaluation," in *International Conference on Speech Science and Technology*, pp. 306–311, 2000.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-28, pp. 357–366, 1980.

[11] D. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Eurospeech*, vol. 2, pp. 963–966, 1997.

[12] J. Pelecanos, S. Myers, S. Sridharan, and V. Chandran, "Vector quantization based Gaussian modelling for speaker verification," in *International Conference on Pattern Recognition*, vol. 3, pp. 298–301, 2000.

[13] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 42–54, 2000.

[14] National Institute of Standards and Technology, "NIST speech group website." http://www.nist.gov/speech, 2001.

[15] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Eurospeech*, vol. 4, pp. 1895–1898, 1997.