RAPID CHANNEL COMPENSATION FOR SPEAKER VERIFICATION IN THE NIST 2000 SPEAKER RECOGNITION EVALUATION

J. Pelecanos and S. Sridharan
Speech Research Lab, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology

ABSTRACT: A technique is proposed for rapidly compensating for channel effects of telephone speech for speaker verification. The method is generic and can be applied to both the one and two speaker detection tasks without re-training the separate systems. The technique has the advantages that it can be performed in real time (except for the small initial buffering), it does not suffer from a relatively long settling time such as certain RASTA processing techniques, and in addition, it is computationally efficient to apply. Results of the application of this technique to the NIST 2000 Speaker Recognition Evaluation are reported.

1. INTRODUCTION

Speaker Verification is the process of accepting or rejecting the claimed identity of a speaker based on a sample of their voice. Applications of speaker verification include secure building access, credit card verification and over-the-phone security access. High performance speaker recognition has been achieved under controlled laboratory and office recording conditions (Liou and Mammone, 1995) and is suitable for practical implementation under these circumstances. Unfortunately, performance of these systems severely degrades under adverse environmental and mis-matched conditions. High performance speaker verification performed over the telephone network is consequently a challenging task. In the recent NIST Speaker Recognition evaluation (NIST, 2000), the recognition performance reported for matched recording conditions is significantly better than mis-matched tests and the latter remains a formidable challenge. The NIST evaluation is an annual international event aimed at advancing the state-of-the-art technology in speaker recognition. A large portion of research has been directed at minimising the effects of varied channels and handsets. Of interest in this research, is the compensation of multiple channel sources with the aim of enhancing recognition performance. In addition, there is a goal of not retraining a speaker recognition system for different speaker detection scenarios. A constraint in this experiment requires the channel compensation technique to perform well under the one and two speaker detection tasks. In this way, once a speaker model is obtained, there is no need to re-evaluate it given a different testing scenario.

The two scenarios of interest are the one and two speaker detection tasks. The one speaker detection task is the most basic. It is the process of accepting or rejecting the claimed identity of a speaker from their voice signal when the voice signal contains the content of a single speaker. In contrast, with two speaker detection, the speech signal contains up to two speakers, one of which may be the target speaker. In the NIST 2000 evaluation (NIST, 2000), the two-speaker test utterance is formed by the addition of the two channels of the speaker conversation into a single channel. Compensating for channel effects is now more difficult. This is due to there being two separate sources of speech, with each source being affected by a different channel.

We propose a computationally efficient method of performing channel compensation on the speech with one or more speakers present in the voice segment content. In addition, we compare the performance of this method across both the one and two speaker detection tasks with varied window lengths. These experiments utilised the speaker recognition system submitted by the authors for the NIST 2000 evaluation.

2. CHANNEL COMPENSATION AS APPLIED TO PARAMETERISATION

The traditional and effective method of channel compensation for a single channel source has been to subtract the mean of the corresponding cepstral coefficients determined over the entire speech segment. The problem with this approach when the inclusion of multiple speech sources through different channels is the case, is that this approach would average the channel effects rather than remove them. Ignoring this effect may be somewhat damaging to recognition performance.

Given linear channels (and ignoring handset transducer effects), the sampled output signal, $Y(t)$, can be considered as the summation of the two speech signals $S_1(t)$ and $S_2(t)$, convolved with their corresponding channel impulse responses $H_1(t)$ and $H_2(t)$.

$$Y(t) = H_1(t) * S_1(t) + H_2(t) * S_2(t) \quad (1)$$

Diagrammatically, the recording configuration is indicated in Figure 1.

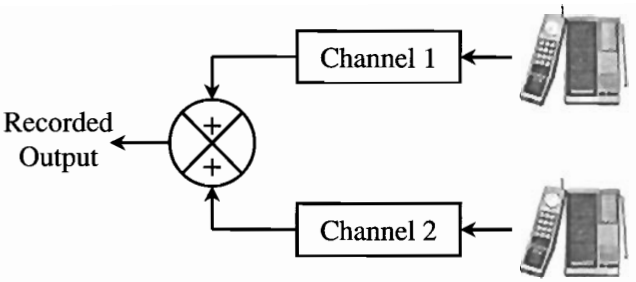

Figure 1. Configuration of the two speaker detection recordings.

In the one speaker detection configuration, the second speech signal source $S_2(t)$, and channel $H_2(t)$, are disregarded. Adding further signal source contributions can accommodate for the N-Speaker detection case.

To curtail the problem of multiple channel effects, there have been several methods proposed. These include sub-segment length feature adjustment techniques such as general IIR (Infinite Impulse Response) RASTA processing (Hermansky and Morgan, 1994) and LDA-FIR (Linear Discriminate Analysis - Finite Impulse Response) Modulation Spectrum analysis (Van Vuuren, 1999). RASTA processing was introduced for speech enhancement purposes and improving speech recognition performance. This technique has since been applied to speaker verification, particularly for over-the-phone applications. This method has been found to have only a limited improvement over the standard segment length mean cepstral coefficient subtraction for one speaker speech segments. The other issue is the settling time of the IIR filter at the start of a speech segment. For short test speech segments (~3 seconds), such as some speech examples trialled in the NIST evaluation, the performance can significantly degrade in comparison with other channel estimate techniques. One of the issues with the IIR filter is how to initialise the output feedback component of the filter. An alternative to this approach is the use of an FIR filter. Here, the N coefficients (~300) of the filter are determined from an LDA of speakers examined across different conditions. This system performs comparably to the standard RASTA method. The drawback of this approach is the effort required for determining the filter using a data-driven approach on an external set of phonetically transcribed speech segments that are consistent with speech in the target application. Recent work (Van Vuuren, 1999) has determined the filter properties based on minimising the signal variation across handsets exclusive of the channel and not the handset deviations with varied telephone channels.

To avoid many of the inherent difficulties with that of the LDA-FIR and IIR RASTA techniques we propose another method to handle one and two speaker speech segments. This method lends itself to compensating for the presence of several speakers also. The rapid channel compensation technique in part, applies a box-car filter to the corresponding cepstral features running in time. Here, the output of the filter is subtracted from the corresponding raw cepstral features. The formal representation of this approach is indicated in z-transform notation in equations (2) and (3), where X(z) is the set of input feature observations and $Y(z)$ is the corresponding output.

$$Y(z) = X(z) - \frac{1}{2N+1}\sum_{i=-N}^{N} X(z)z^{-i} \qquad (2)$$

for a window length of $2N+1$.

$$Y(z) = X(z) - \frac{1}{2N}\sum_{i=-N}^{N-1} X(z)z^{-i} \qquad (3)$$

for a window length of $2N$.

Initially we selected a window size of (say) 300 speech frames at intervals of 10ms. Thus, once a score for the first 300 frames was accumulated to estimate the summation term, proceeding summation estimates could be determined quickly by a simple addition of the next feature coefficient and removal of the last frame of the window. A mean estimate is subtracted from the feature present in the middle of the current window.

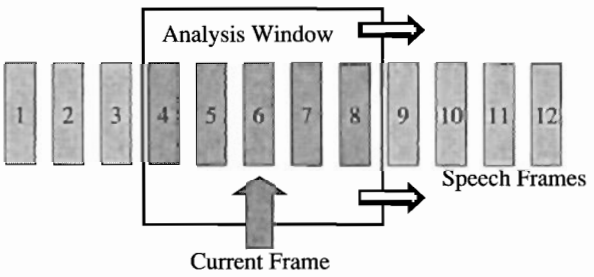

Figure 2. Diagram of the Filtering Approach for Channel Compensation.

The simplicity of this system also allows for the simple compensation of the features at the ends of the speech segment. This is useful for applications of limited length recording segments. To estimate the channel compensated features for the beginning of the speech segment, the nearest available windowing mean estimate is subtracted from the initial set of features. A similar approach is applied to determine the features at the end of the file.

This method indicated by Figure 2 and equations 2 and 3, is significantly faster to calculate than the FIR-LDA Filtering approach. In the LDA-FIR scheme, each compensated coefficient requires the weighted addition of the features spanning the window to be calculated. An alternative being the IIR RASTA method, has the difficulty of seeding the RASTA equation with an initial estimate of the output variable. Hence, a certain number of initial speech frames would have to be ignored to allow the filter to produce a stabilised estimate. As identified earlier, the box-car filter method is not limited to such an extent by this problem.

This style of compensation is suitable for varied channel sources over time such as for two and N speaker detection and speaker tracking. But for these instances, there remains the issue of selecting a suitable window length. A window size that is too short will not capture the channel specific

information, while a longer window length will increase the probability of having two speakers present within the window estimate period. Under this circumstance, the channel estimates of the two speaker signal source would become somewhat averaged. Thus, a suitable window length to balance these effects must be selected. The method of channel compensation proposed and the effects of window size on performance will be examined in our speaker verification system.

3. SPEAKER VERIFICATION SYSTEM OVERVIEW

Introduction

The general structure of the speaker verification module applied to the one speaker detection task is given in Figure 3. One of the differences with this system and the two speaker detection system is that there is no speaker score normalisation in the testing phase of the two speaker detection process. In addition, the distribution of the raw frame based log-likelihood ratio scores was analysed to determine the two speaker detection scores.

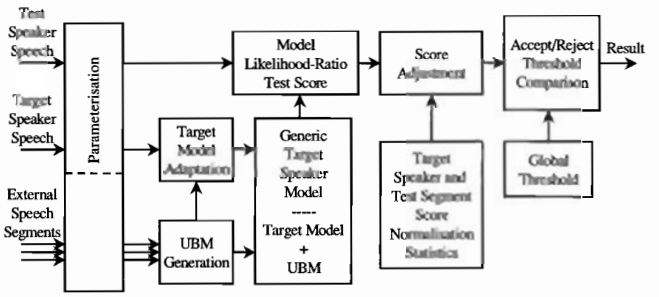

Figure 3. Block diagram of the Adapted-UBM One-Speaker Detection System.

Parameterisation

We used 24 parameters comprised of 12 MFCCs (using 20 filterbanks) with their corresponding delta coefficients. The speech frames were generated using 32ms of speech, offset at 10ms intervals. The signal was bandlimited from 300 to 3200 Hz. Channel compensation was applied to the baseline MFCCs before the delta coefficients were calculated. Silence removal was performed using an energy based histogram approach.

Target Speaker Modeling

We performed speaker modeling by use of the adapted Universal Background Model (UBM) method (Reynolds, 1997). This procedure adjusts the mixtures of a standard speech UBM model toward the distribution of the target speech. The model adaptation process requires the training of a high order GMM on a large quantity of speech. A GMM is a combination of $k = 1, 2, ..., N$, single Gaussian components with dimensionality D, mixture weights p_k, means μ_k, and diagonal covariance matrices Σ_k. For a single speech feature vector observation, X, the probability density function for a speaker model λ, is described.

$$p(\vec{X} \mid \lambda) = \sum_{k=1}^{N} p_k \; g(\vec{X}, \vec{\mu}_k, \Sigma_k) \tag{4}$$

with

$$g(\vec{X}, \vec{\mu}_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_k|}} \exp\left[-\frac{1}{2}(\vec{X} - \vec{\mu}_k)^T \Sigma_k^{-1} (\vec{X} - \vec{\mu}_k)\right] \tag{5}$$

For the verification system, there are two gender dependent models (male and female UBMs) using *orthogonal mixture* GMMs with 512 mixtures. Each UBM was trained on electret handset data from a large portion of the NIST 1999 Evaluation Target Speaker Set. After silence removal, only one in three parameterised frames were kept as training data. This was performed because adjacent frames are typically highly correlated, and keeping the extra data contributes little to the accuracy of the UBM but adds significantly to the training time. Target models were generated by adapting the corresponding gender-specific UBM to the target speaker using MAP adaptation. Both the UBM and the adapted model are stored for the testing phase.

In addition, validation speech was incorporated for performing Handset/Target Speaker Score Normalisation for each target speaker. The NIST 1999 data was partitioned such that validation speakers were not members of the speakers used to train the UBM. This speech data was trialled against the target models to derive the distributional statistics of the impostor speaker set for different handset types. This process called H-Norm, is performed for the carbon and electret handset types to improve performance across multiple handsets (Reynolds, 1997).

Testing Phase

Testing is performed for each frame of a test file, by finding the log-likelihood ratio (LLR) of a given target speaker model with its UBM (male or female depending on the target speaker). Given a speech feature vector X_t, a target speaker model λ_{TARGET}, and a UBM λ_{UBM}, the log-likelihood ratio may be determined.

$$\Lambda_t = \log p(\vec{X}_t \mid \lambda_{TARGET}) - \log p(\vec{X}_t \mid \lambda_{UBM}) \tag{6}$$

Only the top 5 scoring mixtures from the UBM were used for each frame, and the corresponding adapted 5 mixtures (McLaughlin et al, 1999) were used for all hypothesized target speaker tests. By taking advantage of the correspondence between the UBM mixtures and the adapted model mixtures, testing times can be dramatically improved.

The one speaker detection result was determined by averaging these LLR scores over the speech based segments and then performing H-Norm. The two speaker detection result was located by use of a bi-modal Gaussian mixture analysis of the log-likelihood-ratio scores and using the score of the highest scoring Gaussian mean (Myers, 2000). These systems have had proven performance in the NIST evaluations.

4. EVALUATION

Experiment Database

Of interest in this experiment is the performance of the fast channel compensation method and the effect of window size on the performance of one and two speaker detection. We aim to locate a suitable window size to suit both detection tasks. This experiment was examined according to the NIST 2000 speaker recognition specification (NIST, 2000). The database contained 457 male and 546 female target speakers, each with approximately two minutes of telephone speech. The one and two speaker detection tasks used these same target speakers to perform the test. Thus, by modeling each speaker in a universal fashion, the speaker models would not have to be retrained for each task.

One and Two Speaker Detection Results

Presented in Figures 4 and 5 are the one and two speaker detection results. Results are indicated in the form of a Detection Error Trade-off curve (DET). The better performing system has the lower Miss and False Alarm probabilities. For details concerning the DET representation see (Martin et al, 1997).

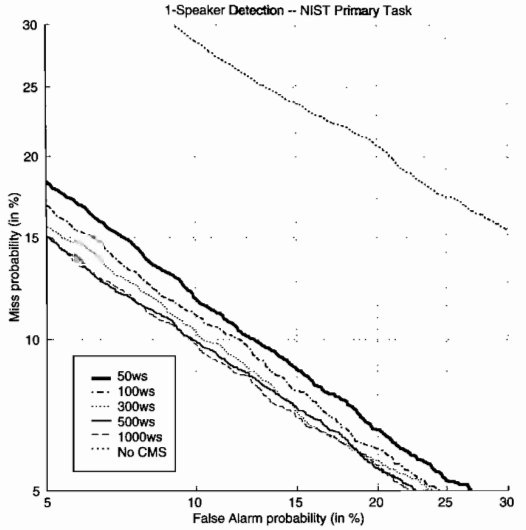

Figure 4. One Speaker Detection DET curve results.

The plot in Figure 4 indicates a generally improving trend of speaker recognition performance with increasing window length for channel compensation. As expected, the 1000ws (1000 frame window length) performed marginally better than the 500 frame compensation. This indicates that the longer the window length (to a certain limit) the better the channel/average vocal tract estimate. This demonstrates that whole utterance length cepstral mean subtraction is quite effective for one speaker detection. Figure 4 also contrasts the difference in performance between cepstral mean removed and the uncompensated speech features. It indicates that ignoring linear telephone network channel effects is detrimental to speaker verification performance.

For the two speaker detection task (Figure 5), an optimal performance was achieved for the 500 frame window length configuration and not the 1000 frame approach (as in the one speaker task). This shows that applying cepstral mean

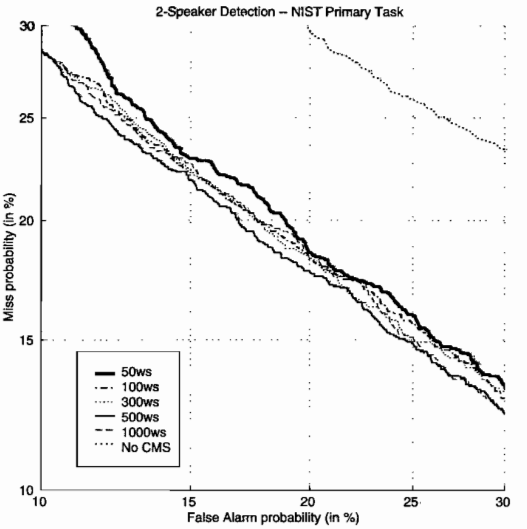

Figure 5. Two Speaker Detection DET curve results.

subtraction over long periods (or whole utterances) with multiple speakers and channels present will degrade multi-speaker detection performance.

5. CONCLUSIONS

It was determined that the running mean box-car filter cepstral removal approach for channel compensation was a successful approach. The optimal window length for both the one and two speaker detection tasks was 500 frames. This particular method of channel compensation is orders of magnitude faster to execute than FIR RASTA alternatives and more stable at the beginning of speech files than IIR based RASTA filter approaches. This method can also be adapted for a fast real-time implementation of speaker recognition applications.

ACKNOWLEDGEMENTS

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

REFERENCES

Hermansky H and Morgan N. (1994) "RASTA Processing of Speech," *IEEE Trans. Speech and Audio Processing*, 2(4), 587-589.

Liou H and Mammone R. (1995) "A Subword Neural Tree Network Approach to Text-Dependent Speaker Verification," *Proc. ICASSP*, pp 357-360.

Martin A, Doddington G, Kamm T, Ordowksi M and Przybocki M. (1997) "The DET Curve in Assessment of Detection Task Performance," *Proc. Eurospeech*, 4, 1895-1898.

McLaughlin J, Reynolds D and Gleason T. (1999) "A Study of Computation Speed-Ups of the GMM-UBM Speaker Recognition System," *Proc. Eurospeech*, 3, 1215-1218.

Myers S, Pelecanos J and Sridharan S. (2000) "Two Speaker Detection by Dual Gaussian Mixture Modelling," *Proc. Australian Internat. Conf. Speech Sci. and Tech.*, pp 300-305.

NIST. (2000) *NIST's Speech and Speaker Recognition Website*. http://www.nist.gov/speech

Reynolds D, (1997) "Comparison of Background Normalization Methods for Text-Independent Speaker Verification," *Proc. Eurospeech*, 2, 963-966.

Van Vuuren S. (1999) *Speaker Verification in a Time-Feature Space*, PhD Thesis.