Reducing Inter-Session Variability With Transitional Spectral Information

Brett R. Wildermoth and Kuldip K. Paliwal

Abstract— **In this paper, we explore the use of transitional spectral information as a means of reducing the effect of inter-sessional variation on automatic speaker recognition performance. We discuss the use of a system based feature and outline the use of an orthogonal polynomial approximation of the derivative used to represent transitional information. A verification and identification system based on the Gaussian Mixture Model is used as the classifier and experiments are carried out using the digit-spl database. Experiments showed that inter-session variability causes a significant degradation in the performance of an ASR system. The use of transitional information helps in overcoming the effect of inter-sessional variability.**

I. Introduction

AUTOMATIC Speaker Recognition (ASR) systems are useful for verifying the identity of a person; allow automated control of services by voice, such as banking transactions and also control the flow of confidential information. While retinal scans and fingerprints are considered more reliable means of identification, speech can be seen as a non-invasive biometric that can be collected with or without the speakers' knowledge and transmitted over long distances via telephone lines.

Modern day ASR systems are divided into two classes depending on their desired function: Automatic Speaker Identification (ASI) and Automatic Speaker Verification (ASV). ASI systems attempt to answer the question "who are you?", while ASV systems ask the question "are you who you claim to be?" [1]. An ASV system decides on the identity claim made by the speaker and the output of the system is in the form of a binary result, accept or deny. An ASI system returns the identity of the most likely speaker, from those enrolled in the system.

Currently available ASR systems have a number of problems. The performance of any ASR system can be affected by: background noise, cocktail party effect, channel distortion, insufficient training, inter-session variability and changes in the speaker's voice due to stress or illness. The goal of this paper is to address the problem of inter-session variability, which can be caused by model drifts and changes in the recording equipment/environment, when there is a significant time lapse between recording sessions. In this paper we explore the use of transitional information of a system based feature as a possible direction to solving this problem. The paper is organised as follows. In section II, we give a brief overview of short-time spectral representations. Section III covers the speaker verification and identification systems used in this paper, highlighting, the database, feature extraction, and speaker modeling and classification. Section IV outlines the experiments conducted and the results achieved. The paper is concluded in Section V.

II. Short-time spectral representations

Speech can be considered the combination of both a system and source component. The system component has been modeled successfully by cepstral coefficients. Cepstral coefficients are most commonly used in speaker recognition and can be easily derived through LPC analysis [2].

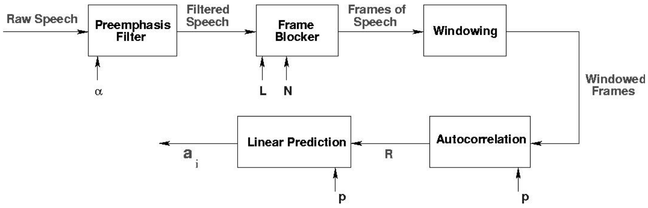


Fig. 1. LPC Analysis.

The speech utterance is segmented into short time frames, windowed, and a p^{th} order LPC analysis performed (see Fig. 1). If we represent the p linear prediction coefficients by a_i, $i = 1,\ldots,p$ then the cepstral coefficients (c_m) are derived from the LPC coefficients using:

$$c_m = a_m + \sum_{i=1}^{m-1} \left(\frac{i}{m}\right) c_i a_{m-i}, \;\; 1 < m \leq p \qquad (1)$$

The changes in cepstral coefficients with respect to time can be represented using transistional spectral information.

A. Transitional Spectral Information

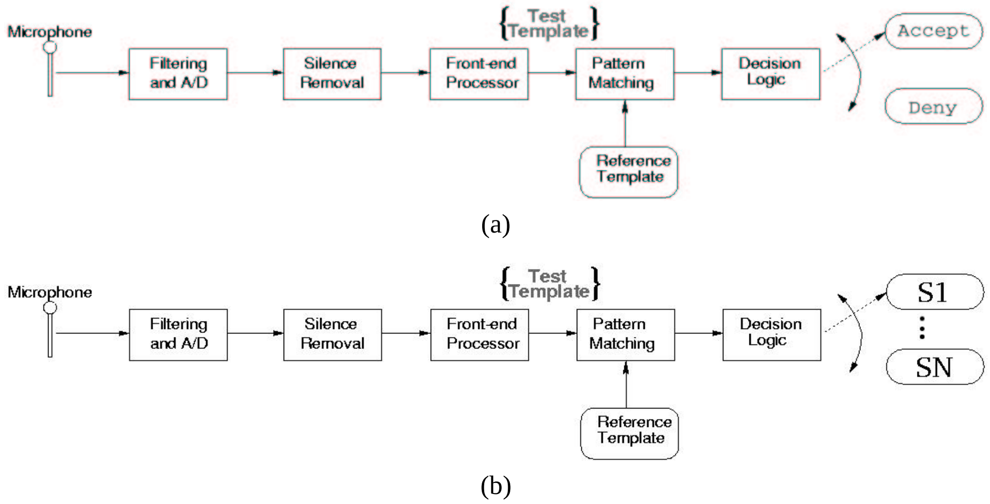
The derivatives of the short-time spectral features are used to represent the transitional spectral information [6]. Short-time spectral features very rarely have an analytical form, so finding their derivative can only be done using a finite difference. This can be successfully implemented using an orthogonal polynomial fit over a finite length window, i.e.

$$\frac{\partial c_m(t)}{\partial t} \approx \Delta c_m(t) = \frac{\sum_{k=-K}^{K} k h_k c_m(t+k)}{\sum_{k=-K}^{K} h_k k^2}, \qquad (2)$$

where $2*K+1$ defines the length of the window (h), and is usually 7 and 5 for first and second derivatives, respectively. The second order derivatives are found in the same manner. However, the first order derivatives are used instead of the cepstral coefficents.

III. Speaker Verification and Identification System

The modern day ASV/ASI system consists of six key components: filtering and A/D, silence removal, front-end processing, pattern matching, decision logic, and enrollment (see Fig. 2). The filtering and A/D section is responsible for capturing speech from the real world. The silence is then removed from the speech and converted into a series of highly representative short-time spectral features (LPCC) that highlight the speaker specific properties present in the speech. Using these features the pattern matching section relates them to stored models and calculates a distortion/probability for each model. Using the result of the pattern-matching section the system makes a decision on the



(a)

(b)

Fig. 2. The components of (a) the basic ASV system and (b) the basic ASI system.

validility of the speaker's claim, or the identity of the speaker. However, the system must first be trained to identify speakers, a process commonly referred to as enrollment. This section outlines the ASV and ASI system presented in this paper, beginning with an overview of the database used in our experiments.

A. Database

We used a multi-session database of both male and female speakers called *digit-spl*. The database was developed at Griffith University in the early part of 2001 and consists of relatively clean speech (an average SNR of 41.6dB) from 68 males and 19 females, spoken on three separate sessions. The sessions are separated by approximately 4-8 weeks. All three session contain ten utterances of two continuously spoken random sequences of five digit numbers, where each digit appeared only once per utterance. The first session contains an additional five repetitions of the isolated word set:"zero", "one", "two", "three", "four", "five", "six", "seven", "eight", "nine".

The isolated word set from session 1 was used to generate speaker models for all 19 female members of the database. The speaker models were tested using the 30 continuous utterances contained in sessions 1, 2, and 3.

B. Feature Extraction

As previously discussed the short-time spectral features employed in this paper are the cepstral coefficients derived through linear prediction analysis (LPCC) and transitional spectral information (Δ).

The speech is sampled at 8kHz with a resolution of 16 bits and preemphasized by the filter $H(z) = 1 - 0.95z^{-1}$. The speech is windowed with a 30 ms Hamming window, with a 10 ms update. A 12^{th} order cepstral coefficient feature is found via a 12^{th} order LPC analysis and the transitional spectral information is found using a polynomial approximation with a rectangular window of length 7 for the first derivative.

C. Speaker Modeling and Classification

A Gaussian Mixture Model (GMM) based text-independent speaker verification and identitification system was used to test the discriminate capabilities of the transitional information of the short-time spectral features. This system was similar to the one proposed by Reynolds [3], [4], [5]. Given a feature vector (x_t), the mixture density for speaker s is defined by

$$p(x_t|\lambda_s) = \sum_{i=1}^{M} p_i^s b_i^s(x_t), \tag{3}$$

and can be thought of as the weighted linear combination of M Gaussian densities $b_i^s(x_t)$. Each trained speaker is represented by a model, $\lambda_s = \{\mu_i,\ \Sigma_i,\ \rho_i\}$ where $i = 1,\ldots,M$, μ_i, Σ_i, and ρ_i represent the mean, variance and weighting of the i^{th} mixture respectively. Since there are generally 40 significant acoustic classes in speech, a model order of $M = 32$ was chosen. The models are trained using 15 iterations of the Expectation Maximization (EM) algorithm, with an initial model trained using the k-means algorithm.

Given the short-time feature representation of the utterance $X = \{x_1,\ldots,x_T\}$, the log likelihood of the utterance belonging to the trained speaker s ($P(X|\lambda_s)$) is found by:

$$P(X|\lambda_s) = \sum_{t=1}^{T} \log p(x_t|\lambda_s) \tag{4}$$

Hence, verification of speakers is achieved by applying an experimental threshold to the log likelihood of the trained speaker. The speaker's claim is therefore accepted only if $P(X|\lambda_s)$ exceeds the threshold, i.e.

$$\begin{aligned} if\ P(X|\lambda_s) \geq T_{experimental} \quad & ACCEPT \\ else \quad & DENY \end{aligned} \tag{5}$$

Identification of speakers is implemented using a maximum likelihood classification rule. The speaker's identity is defined by the model that produced the maximum probability, i.e.

$$i^* = arg \max_{1 \leq i \leq N_s} P(X|\lambda_i), \tag{6}$$

where N_s is the total number of trained speakers.

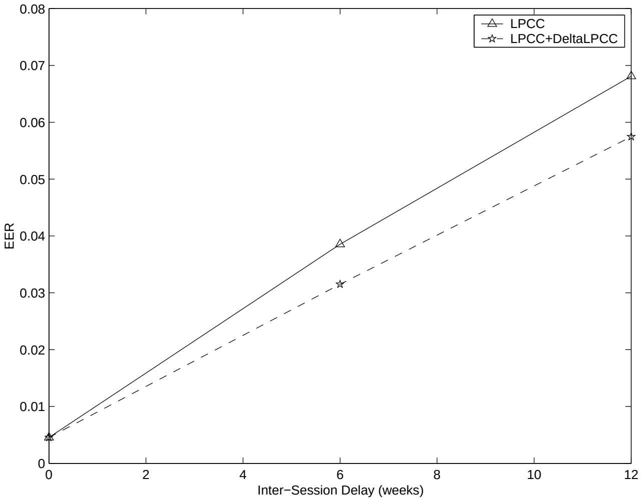


Fig. 3. The effect of intersession variation on the Equal Error Rate (EER).

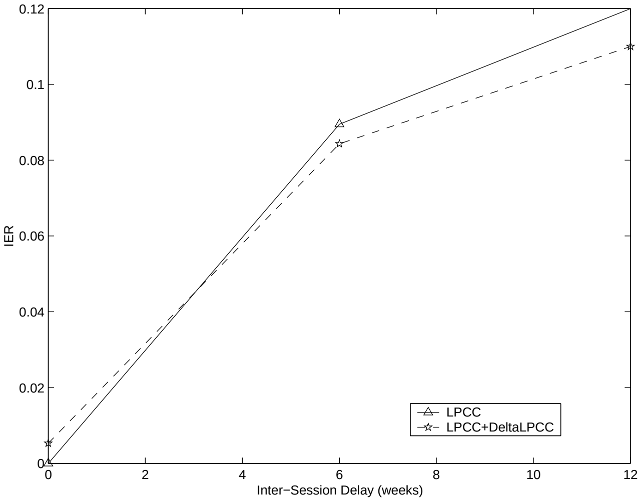


Fig. 4. The effect of intersession variation on the Identification Error Rate (IER).

IV. Experiments and Results

In this section, various combinations of transitional spectral features are used and their effect on, the equal error rate of the ASV system, and identification error rate of the ASI system is shown.

The equal error rate (EER) of an ASV system is defined as

$$ERR = P(fr) = P(fa), \tag{7}$$

where $P(fr)$ is the false rejection rate and $P(fa)$ is the false acceptance rate, defined as,

$$P(fr) = \frac{N_{fr}}{N_C}, \; P(fa) = \frac{N_{fa}}{N_I}, \tag{8}$$

where N_{fr} is the number of times a claimant is rejected by the system, N_C is the number of true claimant tests, N_{fa} is the number of times an imposter is accepted by the system, and N_I is the number of imposter tests performed.

The identification error rate (IER) of an ASI system is defined as

$$IER = \frac{N_{ii}}{N_{ti}}, \tag{9}$$

where N_{ii} is the number of incorrect identifications and N_{ti} is the total number of identifications performed.

Experiments were firstly carried out using the base LPCC feature, as a means of showing the effect of transitional spectral features. The experiment was then repeated using the LPCC feature combined with its first-order derivative, i.e.

$$\hat{C}(t) = [C_1(t), \ldots, C_p(t), \Delta C_1(t), \ldots, \Delta C_p(t)] \tag{10}$$

Figure 3 shows the effect of inter-session variability on an ASV system and Fig 4 shows the effect of inter-session variability on an ASI system. It can be seen that as the time between the training and testing sessions increase the performance of the two systems are significantly affected. However, by including transitional information in the form of first order derivatives of the LPCC feature, the overall perfomance is increased by 8% and 16% for the ASI and ASV systems respectively. It was also noticed that when no inter-session variability existed the performance of the ASI system was reduced when transitional information was used.

V. Conclusion

It has been shown that inter-session variability has a significant effect on the performance of ASR systems. Using transitional information combined with the original base feature vector such as $LPCC + \Delta LPCC$s, the effect of inter-session variability can be reduced. However, in the absence of inter-session variability, the use of transisitional information has a negative effect on the performance of an ASI system.

References

[1] J. P. Campbell, ”Speaker recognition: a tutorial,”*Proc. IEEE,* vol. 85, no. 9, pp. 1437-1462, 1997.

[2] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition,* New Jersey: Prentice Hall, pp. 14-17, pp. 52-65, pp. 112-117, pp. 183-191, 1993.

[3] D. A. Reynolds and R. C. Rose, ”Robust text-independent speaker identification using gaussian mixture speaker models,” *IEEE Trans. Speech and Audio Processing,* vol. SAP-3, no. 1, pp. 72-83, January 1995.

[4] D. A. Reynolds, ”A gaussian mixture modeling approach to text-independent speaker identification,” Ph.D. Thesis, Georgia Institute of Technology, 1992.

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

[6] F. K. Soong and A. E. Rosenberg, ”Use of instantaneous and transitional spectral information in speaker recognition,” *IEEE Trans. Acoustics, Speech, and Signal Processing,* vol. ASSP-36, no. 6, pp. 871-879, 1988.