ROBUST SPEAKER VERIFICATION WITH PRINCIPAL PITCH COMPONENTS

R. M. Nickel[1]*, S. P. Oswal*[2]*, and A. N. Iyer*[3]

Department of Electrical Engineering
The Pennsylvania State University
University Park, PA 16802
{rmn10[1], szo107[2], ani103[3]}@psu.edu

ABSTRACT

We are presenting a new method that improves the accuracy of text dependent speaker identification systems. The new method exploits a set of novel speech features that is derived from a principal component analysis (PC) of voiced speech segments. The new PC features are only weakly correlated with the corresponding cepstral features. A distance measure that combines both, cepstral and PC pitch features provides a discriminative power that cannot be achieved with cepstral features alone. It is well known that the discriminative power of cepstral features declines if the dimensionality of the feature space is increased beyond its optimal value. By augmenting the feature space of a cepstral baseline system with PC pitch features we are able to reduce the equal error probability of incorrect customer rejection versus incorrect impostor acceptance by 12.5% beyond the discriminative limit of the cepstral analysis.

1. INTRODUCTION

The majority of text dependent automatic speaker verification systems in use today employ cepstral features in combination with either dynamic time warping (DTW) or hidden Markov modeling (HMM) to account for possible non linear alignments in the feature space [1]. For text dependent speaker verification, cepstral features exhibit a discriminative power that is, as of now, unsurpassed by any other feature representation for speech [1]. It is, thus, not surprising that, in recent years, the research community has focused more on the pattern matching end and noise reduction part of the problem. The great success in these areas, however, warrants a revisit of the feature extraction problem since the performance of any speaker verification system is inherently limited by the discriminative power of the underlying speech feature. If we want to improve speaker verification systems beyond the discriminative limit of cepstral features we must incorporate additional features that provide independent[1] information.

The features that we are considering in this paper are derived from the local structure of voiced sections of the speech signal [2]. Since every speaker is bound to use the same vocal apparatus for each utterance, it is expected that the generated waveforms will also bear striking similarities *if* the vocal apparatus is operated at the same *operating point*. By comparing suitably chosen waveforms from different utterances we should be able to obtain insight into the identity of the given speaker.

This work has been supported in parts with a grant from the Pittsburgh Digital Greenhouse initiative, Pittsburgh, Pennsylvania.

[1] It is statistically sufficient to provide uncorrelated information.

The caveat of this approach is that: (i) we must restrict ourselves to waveforms that are not chaotic in nature (i.e. utterly unpredictable), and (ii) the operating point variability of the vocal tract must be within reasonable bounds between two utterances of the same word from the same speaker.

Condition (i) is easily satisfied by excluding waveforms from unvoiced (and silent) sections of the utterance. Condition (ii) warrants an *averaging procedure* that focuses on the principal components of the observed waveforms.

The details of the employed feature construction method are outlined in the following section. Section 3 describes the experiment with which we evaluate the discriminative power of the proposed features. In section 4 we present our results.

2. METHODS

The proposed method uses different schemes for training and testing. The principal pitch components are obtained by performing a singular value decomposition on pitch synchronous segments of voiced speech sections from the same phonetic unit. Pitch features derived from different phonetic units are placed into different classes. The block diagram in figure 1 illustrates the training scheme of the proposed method. The testing scheme is illustrated in figure 3.

We will use the notation $\mathbf{s}[n]$ to represent a segment of sampled speech $s[n]$:

$$\mathbf{s}[n] = [\, s[n-L] \;\ldots\; s[n] \;\ldots\; s[n+L] \,]^{\mathrm{T}} \qquad (1)$$

where n is the sample index and L denotes the number of samples that correspond to a segment half-length of 10 msec.

2.1 Utterance Training

The training scheme involves the computation of the principal pitch components, the optimal pitch basis (OPB) feature matrices and the LPC cepstrum. The various steps of the training scheme are discussed below.

2.1.1 Silence/Voiced/Unvoiced Classification

In an initial step we extract the voiced portions of a given training utterance. Indicators such as the short-time energy contour, short-time zero-crossing rate, normalized autocorrelation coefficients and short-time entropy are used in a statistical decision method to identify the voiced sections of the incoming speech signal. A detailed description on the employed methods can be found in [3].

2.1.2 Pitch Aligned Segmentation

In a second step a pitch synchronous segmentation is performed on each voiced section of the given training utter-

ance. The segmentation uses a robust pitch estimator based on the super resolution algorithm proposed by Medan et al. described in [4].

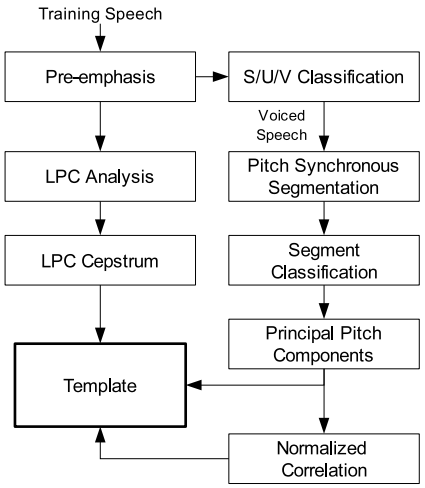

Fig. 1. Block diagram of the training scheme for the proposed speaker verification system.

The pitch contour is then used in a peak picking algorithm to identify the most predominant peak within each pitch frame. The identified peaks (at time indices n_k) serve as a means to align similar glottal events within each pitch frame and across neighboring frames. A pitch frame $\mathbf{s}[n_k]$ is constructed by extracting a 20 msec long segment around each predominant peak as described in equation 1.

2.1.3 Identification of Pitch Classes

The resulting pitch frames are then grouped into classes of similar phonetic units. A weighted spectral slope (WSS) distance [5] between adjacent pitch frames serves as a characterization for the spectral variation between frames. The WSS measure is a weighted difference of thirty six overlapping filters with increasing bandwidths. We use the notation $\mathbf{Q}(\mathbf{s}[k])$ to indicate the vector of weighted spectral slopes of segment $\mathbf{s}[k]$. If j denotes the index of the pitch class and k_j denotes the index of the first segment in the class then an average WSS measure $\bar{\mathbf{Q}}_j[k]$ for class j is computed as:

$$\bar{\mathbf{Q}}_j[k] = \frac{1}{k-k_j+1}\sum_{p=k_j}^{k}\mathbf{Q}(\mathbf{s}[p]) \quad \text{for } k \geq k_j. \tag{2}$$

A new pitch class is formed when the next pitch frame $\mathbf{s}[k+1]$ deviates from the average WSS measure of the current class by more than a fixed threshold ρ:

$$\begin{aligned} \text{if} \quad & \|\bar{\mathbf{Q}}_j[k] - \mathbf{Q}(\mathbf{s}[k+1])\| > \rho \\ \text{then} \quad & k_{j+1} = k+1. \end{aligned} \tag{3}$$

A threshold value of $\rho = 15$ was found to work best based on the training data. All the segments that belong to the same class j are collected to form a class data matrix $\mathbf{C}_j$:

$$\mathbf{C}_j = [\,\mathbf{s}[k_i]\;\mathbf{s}[k_i+1]\;\ldots\;\mathbf{s}[k_{i+1}-1]\,]^{\mathrm{T}} \tag{4}$$

Class segments with less than 4 segments are purged. Such classes are considered to be invalid and formed due to peak picking errors.

2.1.4 Optimal Pitch Bases Expansions (OPB)

An optimal pitch bases expansion [6] is used to derive a principal pitch component as a representation of each pitch class j. Each class data matrix is subjected to a singular value decomposition (SVD):

$$\mathbf{C}_j = \mathbf{U}_j\mathbf{D}_j\mathbf{V}_j \quad \text{with} \quad V_j = [\,\mathbf{v}_j^1\;\mathbf{v}_j^2\;\mathbf{v}_j^3\;\ldots\,]^{\mathrm{T}} \tag{5}$$

Above, we are assuming that $\mathbf{v}_j^1$ is the eigenvector associated with the largest singular value of $\mathbf{C}_j$. $\mathbf{U}_j$ is a unitary matrix and $\mathbf{D}_j$ is a diagonal matrix.

2.1.5 OPB Feature Matrices

In a next step we generate a principal pitch component similarity measure by computing the normalized correlation of the principal pitch component $\mathbf{v}_j^1$ with the training utterance itself. For each pitch class j we obtain a separate representation. The computation of the normalized correlation $\vartheta_j[n]$ of the segments $\mathbf{s}[n]$ with the principal pitch component $\mathbf{v}_j^1$ for each class j at each time instant n can be mathematically expressed as:

$$\vartheta_j[n] = \frac{1}{\|\mathbf{s}[n]\|}\mathbf{s}[n]^{\mathrm{T}}\mathbf{v}_j^1 \tag{6}$$

For the testing part we need features that are aligned with the non pitch synchronous segmentation that is used for the associated cepstral features (see section 2.1.6). We are generating such frame aligned features by determining the maximum value of the normalized correlation $\vartheta_j[n]$ within each non synchronous segment.

$$\varphi_j[m] = \max\{\,|\vartheta_j[Mm-L]|\ldots|\vartheta_j[Mm+L]|\,\}. \tag{7}$$

where M is the frame shift parameter, which makes the segmentation synchronous with the LPC-Cepstral analysis described in section 2.1.6. The match measures $\varphi_j[m]$ for each class j at frame number m are then read into a feature vector $\phi[m]$:

$$\phi[m] = [\,\varphi^1[m]\;\varphi^2[m]\;\varphi^3[m]\ldots\,]^{\mathrm{T}}. \tag{8}$$

The collection of all the feature vectors $\phi[m]$ for all frame numbers m forms the OPB feature matrix $\mathbf{P}$:

$$\mathbf{P} = [\,\phi[0]\;\phi[1]\;\phi[2]\;\ldots\,] \tag{9}$$

Figure 2 shows a typical example for $\mathbf{P}$ derived from 8 pitch classes. The top panel shows the OPB feature template from the training utterance. The second panel shows an OPB feature matrix of an utterance from the same person saying the same word (customer) in a different instance. Note the good match between the customer OPB feature matrix and the OPB template. The third panel shows an OPB feature matrix of a different speaker saying the same word (impostor). It is obvious that the match between the template feature and the impostor feature is poor.

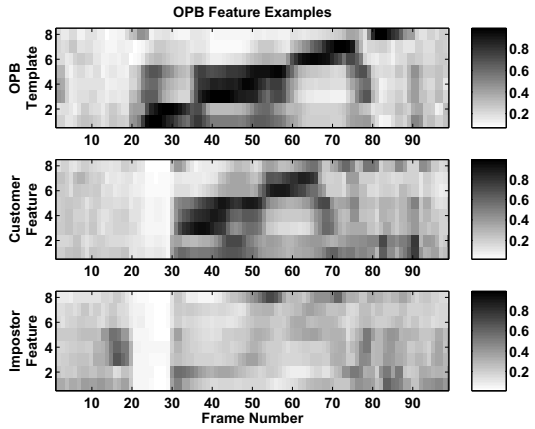

Fig. 2. A typical example for an OPB feature: the OPB template from the training utterance (top panel), a typical customer feature (middle panel), and a typical impostor feature (bottom panel).

2.1.6 LPC Cepstral Analysis

The LPC Cepstral coefficients (LPCC) are computed on segments of speech which are non-synchronous with the pitch aligned segments. The segments $\mathbf{s}[mM]$ are constructed as described in equation 1, where m is the segment index and M is the segment shift parameter. An LPC analysis of order 22 was performed on each of the segments $\mathbf{s}[mM]$ and 24 cepstral coefficients were computed [7]. The LPC order and number of cepstral coefficients were chosen experimentally to obtain the lowest equal error rates.

In summary, a template representaion consists of three parts: (i) the principal pitch components $\mathbf{v}_j^1$ for each class j, (ii) a LPCC representation, and (iii) the OPB feature matrix computed from the training utterance.

2.2 Utterance Testing

The testing scheme of the proposed method is much simpler compared to the training scheme. The testing phase of the proposed method is summarized in a the block diagram shown in figure 3. In the testing scheme, the principal pitch components $\mathbf{v}_j^1$ of each class j are used as matched filters on the testing utterances. The OPB feature matrix is computed on the testing utterance as described in equations 6 through 9, in which $\mathbf{s}[n]$ is the sampled speech from the testing utterance and $\mathbf{v}_j^1$ is the principal pitch templates computed from the training data.

2.2.1 Dynamic Time Warping

A dynamic time warping procedure is employed to perform the pattern matching step. The optimal time alignment between two utterances is determined from the cepstral features and then used to also align the OPB features of the training and testing utterances. The optimal time alignments were determined with a type I local path constraints (as described in [7]), a relaxed endpoint constraint with maximum offset of 7 frames, uniform slope weighting and path normalization. The final combined distance is obtained from an appropriately weighted linear combination of the cepstral distance d_c and the OPB distance d_o:

$$d = d_c + w_f \cdot d_o \tag{10}$$

Based on experiments a value of $w_f = 5$ for the weighing factor was determined best.

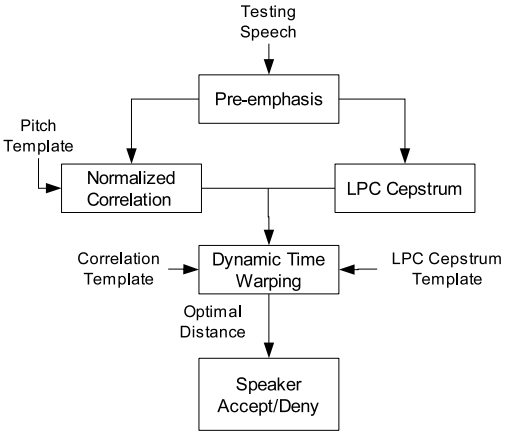

Fig. 3. A block diagram of the proposed utterance testing procedure.

3. EXPERIMENTS

The performance of the proposed method was evaluated with verification experiments over a data set of 3200 utterances from the TI46 speech corpus. All verification results were based on utterance-by-utterance comparisons.

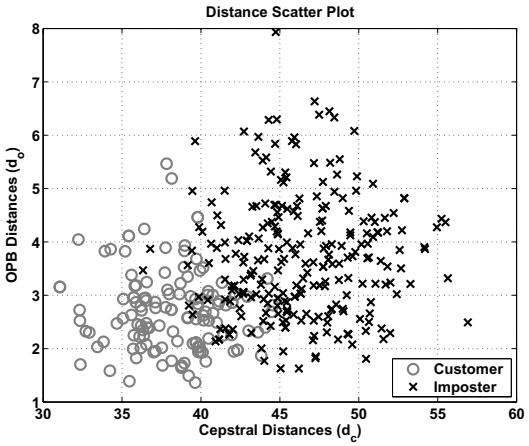

Fig. 4. Scatter plot of a random subset of the inter-speaker distances (○) and the intra-speaker distances (×).

Two verification experiments are compared: 1) a verification system based on 24 LPCC[2]coefficients alone, and 2) a combined system with 24 LPCC[2] coefficients and the new OPB feature matrices. Each system was successively trained with a single utterance from the TRAIN set of the database and then tested against 8 customer utterances (same speaker, same word) and 8 impostor utterances (different speaker, same word). The testing was performed for all 20 words of

[2]Based on an AR model of order 22.

the TI20 sub-corpus and all 16 subjects leading to a total of 5120 comparisons.

A scatter plot of a random subset of distances from the LPCC system (d_c) and from the OPB feature matrices (d_o) is presented in figure 4. It is readily visible that the distances obtained from the cepstral features are only weakly correlated with the distances obtained from the OPB features.

4. RESULTS

The results of the experiments are summarized in the receiver operating characteristics (ROC) shown in figure 5. The speaker verification performance of the 24 LPCC system is indicated by a gray solid line and the performance of the combined system with 24 LPCC and the OPB feature matrices is indicated by the black solid line. The inclusion of the OPB feature matrices significantly improves the performance. The equal error rate (EER) of the proposed scheme is 7% compared to 8%, a 12.5% reduction in EER over the LPCC coefficients. The gray dotted line is the ROC of a system built with 26 LPCC coefficients, which clearly shows a degradation in the performance with a small increase in the number of LPCC coefficients.

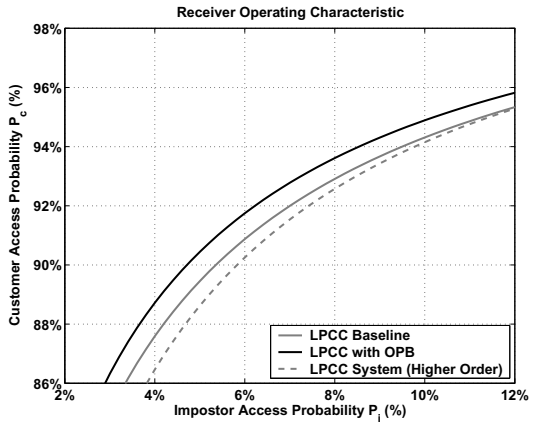

Fig. 5. Receiver operating characteristics for different speaker verification schemes.

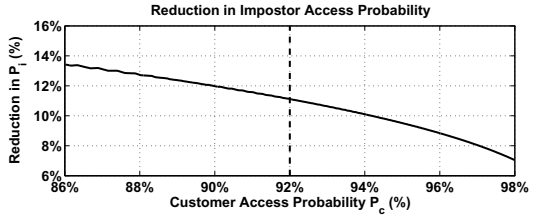

Fig. 6. Reduction in impostor access probability from the ROC curves in the region of interest.

The reduction in the impostor access rate against the customer access rate is shown in figure 6. The curve is generated by computing the relative[3] difference in the impostor access rate of the proposed system and the LPCC system. At a 92% customer access rate, we are able to decrease the wrongful impostor access by 11%.

[3] Relative to the LPCC verification scheme.

5. CONCLUSIONS

A new set of features derived from the principal pitch components was investigated for speaker verification. The new technique has significantly improved the performance when incorporated into a traditional system based on cepstral features. Experimental results have shown a reduction in the probability of imposter access by one third in comparision to a sole use of cepstral features. Furthermore, we have shown that the proposed method can improve the accuracy of a verification system beyond that provided by a system that is based on cepstral features alone. We are only using the main principal component of each pitch class. We are anticipating further improvements when several components are employed.

REFERENCES

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. of the IEEE*, vol. 85, no. 9, September 1997.

[2] R. M. Nickel and W. J. Williams, "On local time-frequency features of speech and their employment in speaker verification," *Journal of the Franklin Institute, Special Issue on Time-Frequency Analysis and Applications*, vol. 377, pp. 469–481, 2000.

[3] L. R. Rabiner, C. E. Schmidt, and B. S. Atal, "Evaluation of a statistical approach to voiced-unvoiced-silence analysis for telephone-quality speech," *The Bell System Technical Journal*, vol. 56, no. 3, pp. 455–482, March 1977.

[4] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Transactions on Signal Processing*, vol. 39, no. 1, January 1991.

[5] D. H. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," *Proc. of the 1982 IEEE Conference of Acoustics, Speech, and Signal Processing ICASSP, Paris*, pp. 1278–1281, May 1982.

[6] R. M. Nickel and S. P. Oswal, "Optimal pitch bases expansions in speech signal processing," *in Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers*, November 2003.

[7] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Inc., Englewood Cliffs, New Jersey 07632, 1993.

[8] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 24, no. 3, June 1976.

[9] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing, A Guide to Theory, Algorithm, and System Development*, Prentice Hall, Upper Saddle River, New Jersey, 2001.

[10] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, New York, 1993.