Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition

Konstantin P. Markov and Seiichi Nakagawa

Department of Information and Computer Sciences, Toyohashi University of Technology, 1-1 Hibarigaoka, Tenpaku-cho, Toyohashi, 441-8522, Japan

(Received 17 July 1998)

In the speaker recognition, when the cepstral coefficients are calculated from the LPC analysis parameters, the prediction error, or LPC residual signal, is usually ignored. However, there is an evidence that it contains a speaker specific information. The fundamental frequency of the speech signal or the pitch, which is usually extracted from the LPC residual, has been used for speaker recognition purposes, but because of the high intra-speaker variability of the pitch it is also often ignored. This paper describes our approach to integrating the pitch and LPC-residual with the LPC-cepstrum in a Gaussian Mixture Model (GMM) based speaker recognition system. The pitch and/or LPC-residual are considered as an additional features to the main LPC derived cepstral coefficients and are represented as a logarithm of the $F0$ and as a filter bank mel frequency cepstral (MFCC) vector respectively. The second task of this research was to verify whether the correlation between the different information sources is useful for the speaker recognition task. For the experiments we used the NTT database consisting of high quality speech samples. The speaker recognition system was evaluated in three modes - integrating only pitch or only LPC-residual and integrating both of them. The results showed that adding the pitch gives significant improvement only when the correlation between the pitch and cepstral coefficients is used. Adding only LPC-residual also gives significant improvement, but in contrast to the pitch, using the correlation with the cepstral coefficients does not have big effect. The best results we achieved using both the pitch and LPC-residual and are 98.5% speaker identification rate and 0.21% speaker verification equal error rate compared to 97.0% and 1.07% of the baseline system respectively.

Keywords: speaker recognition, LPC cepstral, LPC residual, pitch

PACS number: 43.72.Pf

1. INTRODUCTION

At the early stage of the automatic speaker recognition research, the fundamental frequency of the speech signal, or the pitch, had attracted many researchers' attention. The pitch contour, as well as the long time pitch average, have been extensively investigated and it has been found that they carry much speaker specific information [1]. However, as those results show, using the pitch alone is not enough to build high performance long term speaker recognition system. This is, probably, the reason that in recent years interest in the use of the pitch seems to have diminished. Another problem which was encountered is that it was difficult to integrate the pitch in a text-independent system. The pitch extraction was not also much reliable and computationally expensive.

In the last decade, research has been focused on using the spectral information, especially the cepstral coefficients, for speaker recognition [2–4]. There have been several studies, for example [5–7], trying to use both the pitch and the cepstral coefficients. The main problem in such combination, in the case of text-independent speaker recognition, is that there are voiced and unvoiced parts of speech, i.e. the pitch is not present in all the frames and this makes the modeling complicated. The approach taken in [5], where VQ codebook is used as a model, is to train two separate models for each speaker from the voiced and unvoiced parts of the training data respectively. For dealing with different kinds of feature parameters, an appropriate distribution normalization is applied. In [7] the pitch is modeled separately using mixture model which takes into account the probability of pitch extraction errors - pitch halving and doubling. The relative entropy between pitch distributions of the model and the test utterance is used as a pitch score which is further combined with the score obtained from the conventional GMM cepstral system.

In our speaker recognition system, which is based on GMM, we combine the cepstral and pitch information at the frame level by augmenting the cepstral feature vector with the pitch parameter. Since for the unvoiced speech segments no pitch can be extracted, in this case, the cepstral vectors are used as they are. This prompted as to use two models per speaker (as in [5]) for voiced and unvoiced speech segments respectively. Another issue of interest which to our knowledge has not been addressed yet, is whether the correlation between the pitch and cepstral coefficients is useful for the speaker recognition task. The study [8] shows that the change of the pitch results in the change of the cepstrum and, therefore, the pitch/cepstral correlation may carry speaker specific information. Using models with a full covariance matrix gives us very simple way of utilizing such correlation.

The LPC technique is a very powerful and popular method for speech analysis, because it provides extremely accurate estimates of speech spectrum and is computationally inexpensive. A by-product of the LPC analysis is the prediction error signal, also called LPC residual signal. If the speech could be perfectly modeled by the all-pole model, the residual signal would be very small. However, this model is not suitable for nasal and fricative sounds. For example, nasal sounds having anti-formant frequencies have useful acoustic characteristics for speaker recognition [9]. Thus, the prediction error essentially carries all information that has not been captured by the LPC coefficients. On the other hand, the LPC residual signal is generally considered as an approximation of the glottal flow which obviously differs among speakers. The information lost in the LPC analysis contains the fundamental frequency (pitch), the shape of the glottal flow signal and those spectral elements which cannot be modeled by the all-pole LPC model. Therefore, we have enough reasons to believe that the LPC residual contains additional speaker specific information. However, only recently it has attracted researchers interest and the published works where LPC residual is used for speaker recognition are very few, for example [10–12]. Since the LPC residual is a time domain signal as the speech itself, in order to extract information from it some kind of spectral analysis is necessary. The approach taken in [10, 11] is to transform the LPC residual into a cepstral coefficients using FFT - much like Mel Frequency Cepstral Coefficients (MFCC) for the speech signal. In [12] the LPC residual is represented in terms of power difference spectrum in subband (PDSS) which is derived also from the FFT spectrum. The next issue is how to combine the two types of information sources. In [10, 12] the LPC cepstral coefficients and the representation of the LPC residual are treated as a separate feature streams and are combined at the model level, i.e. the scores of the respective models are linearly combined. In contrast, in [11] they are combined at the feature vector level, i.e. by augmenting the LPC cepstral vector. Furthermore, only voiced segments of the speech signal are used for feature extraction. We have to note that in all mentioned works both features are modeled using vector quantization technique (LVQ in the case of [11]).

In our speaker recognition system, the LPC residual is transformed into cepstral coefficients obtained using mel frequency filter bank analysis - MFCC. We have considered this analysis method because it also gives very good spectral representation, but does not require the source signal to be modeled by an all-pole filter. We have tried both approaches to combining the conventional LPC cepstral coefficients and LPC residual MFCC, i.e. by treating them as separate feature streams and by forming one feature vector from both types of cepstral coefficients. In all the cases we use a GMM for the modelization.

Although the pitch is contained in the LPC residual signal and, generally, the LPC residual representation in the form of cepstral coefficients should include the pitch information as well, in practice, in order to keep a reasonable number of model parameters, only the low order cepstral coefficients are used. The pitch, however, as a frequency of the signal is included in the higher order coefficients. Therefore, it is reasonable to assume that the MFCC of the LPC residual may not contain all of the pitch information and that adding the pitch parameter explicitly would have some effect. That is why, we have experimented with combination of both the pitch and LPC residual by adding the pitch parameter to the augmented cepstral vector and again using two models (voiced and unvoiced) per speaker.

As a baseline system for comparisons we used a conventionally trained GMM using only LPC derived cepstral coefficients. Previously, we have developed and experimented with the frame level likelihood transformation technique [13, 14]. There was a significant effect of applying this technique to our baseline system. In this research, we also experimented with this technique and we achieved further improvements of the system performance.

2. FEATURE PARAMETERS

2.1 LPC Mel-Cepstrum and ΔCepstrum

The basic idea of the LPC analysis is that a given speech sample can be approximated (or predicted) by a linear combination of the past p samples:

$$s(n) \approx \sum_{k=1}^{p} a_k s(n-k) \tag{1}$$

Thus, we can define a linear predictor as an all-pole system whose output is:

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{2}$$

where a_k are the LPC prediction coefficients, p is the prediction order and $s(n)$ are the samples of the speech signal. There are many methods to calculate LPC coefficients. The most popular is the autocorrelation method which allows a_k to be efficiently calculated by the Durbin's recursive algorithm [15]. The LPC coefficients are then transformed into cepstral coefficients using:

$$c_k = a_k + \sum_{n=1}^{k-1} \frac{n}{k} c_n a_{k-n} \tag{3}$$

Since it has been found that mel-wrapped features perform better, the LPC cepstral coefficient can be further transformed into a mel-frequency scale. This is usually done by bi-linear frequency warping using an all-pass filter [16].

The delta spectral coefficients which provide a transitional spectral information can be found using:

$$\Delta c_k(t) = \frac{\sum_{i=-L}^{L} i h_i c_k(t+i)}{\sum_{i=-L}^{L} h_i i^2} \tag{4}$$

where h_i is a symmetric window of length $2L+1$ frames.

2.2 LPC Residual Cepstrum

The prediction residual signal, or the prediction error, is found directly from Eq.(1) and Eq.(2):

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{5}$$

As described in Introduction, the LPC residual signal is interpreted as the excitation of the LPC model of speech and approximates the glottal flow which obviously differs among the speakers and, thus, provides a speaker specific information.

In practice, the LPC residual is obtained by inverse filtering of the speech signal using its autoregressive parameters computed by the standard LPC analysis as filter coefficients.

Obtained LPC residual signal is then transformed into cepstral coefficients using the standard mel frequency filter-bank analysis technique. In more detail, this method consists of the following steps:

a) Framing the LPC residual with the same rate and length as the original speech signal.

b) Applying a Hamming window.

c) Obtaining the magnitude spectrum with FFT.

d) Forming M filter banks in the mel scale.

e) Computing the log filter-bank amplitudes.

f) Calculating d cepstral coefficients from the filter-bank amplitudes using Discrete Cosine Transform (DCT).

2.3 Pitch Parameter

Besides the LPC-residual spectrum, the fundamental speech frequency (pitch) is widely used as a representation of the glottal flow information. The pitch frequency is extracted from the LPC-residual signal and estimated using an algorithm based on the normalized short-time autocorrelation function which does not require the selection of the frame length [17]. For minimization of the pitch extraction errors, such as pitch doubling or pitch halving, a post-processing is applied as proposed in [18].

Pitch frequency values are extracted at intervals, corresponding to the cepstral frames time rate. In other words, the extraction of the pitch and cepstral coefficients is synchronized such that for each cepstral vector there exists a pitch value. The pitch value is zero for the unvoiced parts of the speech signal. This scheme is particularly useful when deciding whether the current cepstral vector represents a voiced or unvoiced speech interval.

2.4 Combined Feature Vectors

In our speaker recognition system, when using the pitch information, the LPC derived cepstral vector, denoted by CEP, is augmented with the logarithm of the pitch frequency. For the unvoiced parts of speech where the pitch value is zero, cepstral vectors are kept unchanged. Thus, a given speech utterance is represented by two types of feature vectors - voiced and unvoiced which have the following structure:

$$x_t^{voiced}{=}(c_{1t}, c_{2t}, \ldots, c_{dt}, \log F0_t)$$
$$x_t^{unvoiced}{=}(c_{1t}, c_{2t}, \ldots, c_{dt})$$

where c_{it} is the ith cepstral coefficient at time t and $\log F0_t$ is the logarithm of the pitch frequency. We used $\log F0$ instead of $F0$ because as shown in [7] the distribution of the $\log F0$ is closer to the normal distribution. Note that the two types of feature vectors have different dimension: $d+1$ for voiced and d for unvoiced vectors.

When using the LPC residual cepstral coefficients, denoted by R-CEP, we investigated two approaches. The first treats the R-CEP features as a separate stream and, thus, they are modeled by a separate GMM. The second approach is to form one long feature vector consisting of both CEP and R-CEP coefficients. Adding the pitch parameter, in the latter case, again leads to a split of the feature vectors into voiced and unvoiced sets.

3. DECISION PROCEDURE

3.1 Using LPC-cepstrum

A GMM is a weighted sum of M component densities and is given by:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i N(x; \mu_i, \Sigma_i) \qquad (6)$$

where x is a d-dimensional vector, $N(x;\mu_i,\Sigma_i)$ is the i^{th} Gaussian density, w_i is the mixture weight and λ represents all GMM parameters. The log-likelihood of an observation sequence $X = x_1, x_2, \ldots, x_T$ is:

$$L(X|\lambda) = \log p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda) \qquad (7)$$

In the speaker identification task, it has to be decided to whom of a group of N known speakers a given speech sample belongs. The conventional maximum likelihood approach is to decide in favor of that speaker whose model i^* is:

$$i^* = \arg\max_i L(X|\lambda_i) \qquad (8)$$

The speaker verification task is a binary decision problem, where it has to be decided whether the speech sample belongs to the claimant speaker or not. The general approach is to apply likelihood normalization:

$$l(x) = \frac{p(X|\lambda_c)}{p(X|\lambda_{\overline{c}})} \qquad (9)$$

where λ_c is the claimant speaker model and $\lambda_{\overline{c}}$ is a model representing all other possible speakers. The $l(x)$ is then compared with a threshold and if it is bigger, the speech sample is accepted as being uttered by the claimant speaker. Otherwise it is rejected. In our speaker recognition system, $p(X|\lambda_{\overline{c}})$ is approximated by a collection of B *background* speakers:

$$p(X|\lambda_{\overline{c}}) \approx \frac{1}{B} \sum_{b=1}^{B} p(X|\lambda_b) \qquad (10)$$

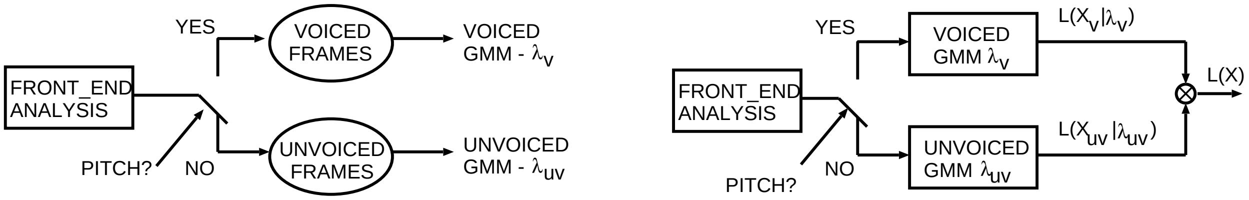
This kind of likelihood normalization we call *sentence level* likelihood normalization in contrast to the *frame level* likelihood normalization which is briefly discussed in Section 4.

In our baseline system, each speaker is represented by one GMM and the vector x consists of only LPC-cepstral coefficients.

3.2 Using pitch

In our system, each speaker is represented by two Gaussian mixture models (GMM) trained on the corresponding collections of the unvoiced and voiced frames. Fig.1a shows the block diagram of the training algorithm.

After the front-end analysis, the training feature vectors are divided into two subsets, voiced X_v and unvoiced X_{uv}, by checking their dimension. Then from each subset a GMM is trained using the conventional Maximum Likelihood Estimation (MLE). We



a) Training

b) Testing

Fig. 1 Block diagram of the training and testing algorithms.

have to note that since our pitch extraction algorithm is not perfect, there may be kind of errors where pitch was not extracted for a voiced frame or, conversely, unvoiced frame was given a pitch value. Such errors occur at the beginning or at the end of some voiced speech segments. Therefore, the subsets of voiced and unvoiced frames contain small part of falsely assigned frame. However, we believe that this does not have any significant effect on the system performance because both the training and test data are subject to such kind of errors. In fact, we found that roughly 1% of the voiced frames were falsely judged as unvoiced and that this caused no misrecognition errors.

When the Gaussian densities of the models have full covariance matrix, for the voiced GMM, it has the following structure:

$$\Sigma = \begin{bmatrix} \sigma_{11}\sigma_{12}\cdots\sigma_{1d}\,\rho_1 \\ \sigma_{21}\sigma_{22}\cdots\sigma_{2d}\,\rho_2 \\ \cdots\cdots\cdots\cdots\cdots \\ \sigma_{d1}\sigma_{d2}\cdots\sigma_{dd}\,\rho_d \\ \rho_1\ \ \rho_2\ \cdots\ \rho_d\ \ \ \rho \end{bmatrix} \tag{11}$$

where σ represents the cepstral coefficients co/variances, ρ is the pitch variance and $\rho_i, i = 1, \ldots d$ are the pitch/cepstral covariances. Therefore, using a full covariance matrix, we can model not only the pitch itself, but its correlation with the cepstral coefficients as well.

A given test utterance is first divided into voiced and unvoiced parts in the same manner as the training data. Then, the log-likelihood of each part with respect to the corresponding GMM is calculated. However, the whole test utterance score cannot be obtained by a simple addition of the two log-likelihoods. This is because the voiced and unvoiced vectors have different dimension and, therefore, their likelihoods will have different dynamic range. Also, when dealing with two different information sources (voices and unvoiced frames can be viewed as different information sources), the general approach is to take a weighted sum of the two likelihoods since there can be a difference in effectiveness for the recognition task. That is why, we have chosen to take a linear combination of the likelihoods as follows:

$$L(X) = \alpha L(X_{uv}|\lambda_{uv}) + (1-\alpha)L(X_v|\lambda_v) \tag{12}$$

where X_{uv} and X_v denote the unvoiced and voiced subsets of the feature vectors respectively and then the $L(X)$ is used for identification or verification decision. Fig.1b shows the block diagram of the test procedure.

3.3 Using LPC residual

As mentioned in Section 2.4, the LPC cepstral and LPC residual features are combined in two ways. When the R-CEP coefficients are treated as a separate stream, each speaker is represented by two GMMs - one for CEP and one for R-CEP features. The utterance score in this case is obtained by linear combination of the two models scores in the same way as Eq.(12).

When CEP and R-CEP are combined in one feature vector, one GMM per speaker is used and the speaker recognition system structure does not differ from the conventional one. If there is any correlation between CEP and R-CEP coefficients, in this case, it can be captured and used when the model's pdfs are with full covariance matrices in the same manner as the pitch/CEP correlation.

Adding the pitch parameter to the combined CEP/R-CEP vector allows to use both the LPC residual and pitch in the same time. The speaker recognition system in this case is similar to that explained in Section 3.2.

4. FRAME LEVEL LIKELIHOOD NORMALIZATION

4.1 Using background speakers

We apply a likelihood normalization on the frame likelihoods [13, 14]. Given the speaker model λ_i and vector x_t, the following formula is used for the normalization:

$$p_{norm}(x_t|\lambda_i) = \frac{p(x_t|\lambda_i)}{\frac{1}{B}\sum_{b=1}^{B} p(x_t|\lambda_b)} \tag{13}$$

where λ_b, $b = 1, \ldots, B$ are the background speaker models. For a sequence $X = \{x_t\}$, the log normalized likelihood is:

$$L^{norm}(x|\lambda_i) = \frac{1}{T}\sum_{t=1}^{T} \log p_{norm}(x_t|\lambda_i) \tag{14}$$

4.2 Weighting Models Rank (WMR)

We have proposed the WMR frame level likelihood normalization technique in [13, 14]. The essence of this approach is to compute likelihoods of all speaker models and then to sort them in order, corresponding to the value $p(x_t|\lambda_i)$. Further, each model is given a weight corresponding to its position or rank in the sorted list. Accumulated weights over all sequence of frames form the model's WMR score.

5. EXPERIMENTS

5.1 Database

For the evaluation experiments we used the NTT database for speaker recognition which consists of recordings of 35 speakers (22 males and 13 females) collected in 5 sessions over 10 months in a sound proof room [2]. Each session data contain 10 equal sentences uttered at normal, slow and fast speed and 5 different sentences uttered only at normal speed. The average utterance duration is about 4 sec. For training the models, 5 equal and 5 different sentences uttered at normal speed for each speaker from one session - August 1990 (90.8) were used. Five other sentences/session from the other four sessions (90.9, 90.12, 91.3, 91.6) uttered at normal, fast and slow speeds were used as test data. The input speech was sampled at 12 kHz. 14 cepstrum coefficients were calculated by the 14th order LPC analysis at every 8 ms with a window of 21.33 ms. Then these coefficients were further transformed into 10 dimensional mel-cepstrum (cep) vector which serves as our baseline feature (CEP). Each session's cepstral data were also mean normalized (CMN). Pitch parameter was added to the voiced vectors, thus their dimension is 11. Regressive (ΔCEP) coefficients were calculated separately for each of the voiced and unvoiced data streams using a frame window of 9 cepstral frames (L=4). Note that the ΔCEP for the voiced frames contains the Δpitch parameter as well.

LPC analysis parameters of each frame of the speech signal were stored and then used to obtain the residual signal by inverse filtering of the same speech frame. Then the LPC residual was transformed into 10 MFCC (R-CEP) using 24 mel-scaled filter banks. When the R-CEP coefficients were used separately, ΔR-CEP coefficients were calculated in the same manner as ΔCEP coefficients. When combined with the CEP coefficients in one vector, obtaining Δ's of this vector gives both ΔCEP and ΔR-CEP simultaneously.

5.2 Results using pitch

In the evaluation experiments, the voiced and unvoiced GMMs were set with the same number of mixtures. This was possible because the amounts of voiced and unvoiced training data were roughly the same. The total number of mixtures per speaker was 4 or 8.

In order to assess the effect of using the correlation between the pitch and the cepstral coefficients, we made additional experiments, where the pitch was modeled as an independent feature stream and this correlation was not used. This was done by making the voiced GMM's covariance matrices block-diagonal, i.e. setting their last column and last row elements ρ_i to zero in Eq.(11) (except the diagonal element).

In the columns "ML test" which stands for the Maximum Likelihood test, Table 1 compares the recognition rates among the baseline ("CEP"), the independent pitch modeling case ("Without Correl.") and the case when the correlation between the pitch and the cepstral coefficients is used ("With Correl."). The column "Using Δ's" shows whether the ΔCEP and Δpitch parameters are used. They are modeled as a separate (independent) feature stream with its own voiced and unvoiced GMM. The overall score is a simple summation of the CEP + pitch and ΔCEP + Δpitch scores. This approach to combining the delta features is the same for all cases of this work. These results show, that including the pitch parameter is effective when the pitch/cepstral correlation is used. When this correlation is not used, the system performance is similar to that of the baseline.

The columns "Cohort test" and "WMR test" of the Table 1 show the recognition rates when the

Table 1 Speaker recognition rates using pitch.

		Baseline (CEP)			CEP + pitch			
Model	Using				ML test			
type	Δ's	ML	Cohort	WMR	Without	With	Cohort	WMR
		test	test	test	Correl.	Correl.	test	test
				Identification rate (%)				
4	no	92.3	92.4	92.4	93.9	95.3	95.1	96.0
mixture	yes	94.1	94.8	95.2	93.9	95.3	94.4	96.6
8	no	96.4	96.2	96.6	96.3	97.1	96.9	97.7
mixture	yes	97.0	97.0	97.3	96.8	97.4	97.0	97.6
				Verification equal error rate (%)				
4	no	2.50	2.14	1.31	2.46	1.66	1.33	0.84
mixture	yes	1.64	1.33	0.84	2.28	1.45	1.11	0.64
8	no	1.66	1.38	0.66	1.48	1.21	0.96	0.50
mixture	yes	1.18	0.96	0.52	0.98	0.89	0.80	0.41

frame level likelihood normalization technique is applied to both the baseline system and the system using pitch/cepstral correlation (full covariance matrices). The term "Cohort" means that the background speakers for the frame level likelihood normalization are chosen to be the most acoustically close speakers to the target speaker (see Eq.(13)). The number of the background speakers is $B = 5$. For the speaker verification task, we have also applied sentence level likelihood normalization using the top 10 speakers as background speakers (see Eq.(10)) to both systems with (Cohort test, WMR test) and without (ML test) frame level normalization. It can be seen that this technique works well improving further the performance.

For the fast and slow speed test utterances, even bigger improvement was achieved. The baseline fast speed test best result of 94.0% identification rate increased to 95.9% and to 97.4% with the WMR test. The corresponding rates for the slow speed test are 93.0%, 95.6% and 96.5% respectively. The verification Equal Error Rate (EER) also decreased from 1.43% to 0.64% (with WMR) and from 2.06% to 0.87% (with WMR) for the fast and slow speed tests respectively. For details of calculating the EER see [14].

As stated in Section 3.2, the overall likelihood of the test utterance was calculated using a linear combination of the likelihoods from voiced and unvoiced GMMs. In such a case, an important issue is how to set the combination parameter α. Fig.5.2 shows the speaker identification rate as a function of this parameter.

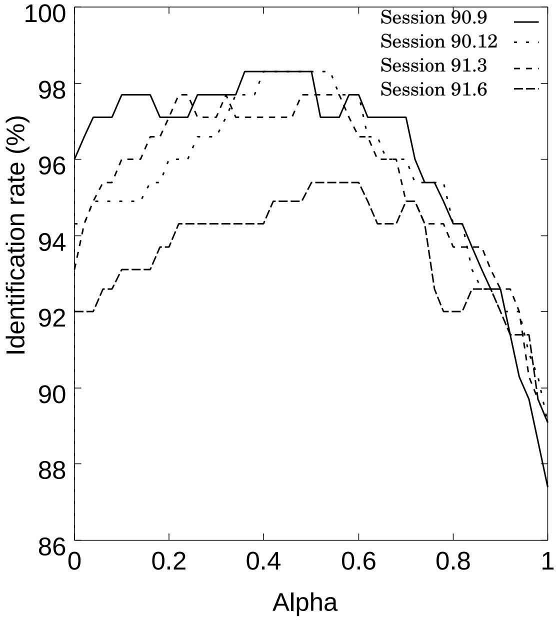


Fig. 2 Speaker identification rate vs. parameter α.

As this figure shows, in average, the maximum identification rate occurs in a wide range of α values - from 0.4 to 0.6, which means that the results are not much sensitive with respect to α. It also can be observed that at $\alpha = 0.5$, i.e. when a simple summation of the voiced and unvoiced log-likelihoods is used, the results are very close to those for the optimum α. This suggests that the effectiveness for speaker recognition of the voiced and unvoiced parts is similar.

Table 2 Speaker identification rates (%) using CEP and R-CEP features. Maximum Likelihood (ML) test.

Mod.	Using	Combined CEP and R-CEP			Baseline
type	Δ's	Lin.Comb.	20 dim.	14 dim.	CEP 10 dim.
4 mix.	no	96.0	96.9	96.0	92.3
full	yes	96.6	96.4	96.9	94.1
8 mix.	no	97.0	96.3	96.4	96.4
full	yes	97.0	96.0	97.4	97.0
32 mix.	no	95.9	95.6	96.6	94.4
diag.	yes	97.7	96.0	97.7	95.9
64 mix.	no	96.4	96.1	98.0	94.1
diag.	yes	96.1	97.3	98.1	95.9

5.3 Results using LPC residual

In the first evaluation experiments with LPC residual, it was modeled as a separate feature stream. We ran several tests using different types of GMM, with full or diagonal covariance matrices and different number of mixtures. Each speaker was modeled by a pair of GMMs of the same type corresponding to CEP and R-CEP features. The overall utterance score was obtained by a linear combination of non-normalized scores from the two models. The optimal combination parameter for all cases was between 0.3 and 0.4. In Table 2, the column "Lin.Comb." shows the speaker identification rates using the standard Maximum Likelihood (ML) test when this combination parameter was set to 0.36.

In the next experiments, the CEP and R-CEP vectors were combined into one 20 dimensional feature vector. Since they were obtained using different analysis techniques and it is not guaranteed that their components have at least similar variances, R-CEP coefficients were scaled appropriately. The results of these experiments are summarized in Table 2 in the column "20 dim.". There is no big difference between these and previous results. However, the poor performance of the 8 mixture, full covariance matrix GMM suggests that probably the training data became insufficient when the model dimension was doubled. Thus, we decided to reduce the R-CEP vectors dimension to 4 using Karuhnen-Loewe (K-L) transformation, since it preserves most of the information from the original vectors.

The transformed R-CEP vectors were combined with the 10 dimension CEP vectors resulting in a 14 dimension feature vectors. The identification results using this new vector are shown in the "14 dim" column of the Table 2. The biggest improvement in this case is seen for the models with diagonal covariances. It is not surprising, because the K-L transformation also diagonalizes the covariance matrices, and thus, almost all information is presented in the diagonal elements of the new covariances. Comparing the performance of the all CEP + R-CEP cases with the baseline, it is clear that using the R-CEP features gives significant improvement up to 4%, which shows that the LPC-residual signal carries speaker specific information not presented in the standard CEP vectors.

Investigating the correlation between CEP and R-CEP coefficients, we ran experiments using models with block-diagonal covariance matrix (4 and 8 mixtures per GMM) and 20 dimension feature vector. Unfortunately, the test with 8 mixture GMM failed because of the underestimated model parameters. Obtained results with 4 mixture GMM were 96.3% without the Δ's and 96.1% when they were used. The difference from the case of full covariance matrix (Table 2, column "20 dim.") is minimal which confirms the fact that the CEP and R-CEP coefficients hold different information and are almost uncorrelated.

Since the 14 dimension CEP + R-CEP vectors performed the best among other cases, we used only them for the next experiments involving the frame level likelihood normalization technique (see Section 4). Table 3 shows the speaker identification rates as well as speaker verification equal error rates when the Cohort and WMR tests were applied to both the CEP (baseline) and CEP + R-CEP cases. Using the Cohort test did not improve the identification performance of the CEP + R-CEP system and the WMR test was better only in the half of the cases. However, the verification error rates were improved in both the Cohort

Table 3 Speaker recognition rates using 14 dimensional CEP + R-CEP feature vector.

Mod.	Using	ML test		Cohort test		WMR test	
type	Δ's	CEP+R-CEP	CEP	CEP+R-CEP	CEP	CEP+R-CEP	CEP
Identification rate (%)							
4 mix.	no	96.0	92.3	95.3	92.4	96.1	92.4
full	yes	96.9	94.1	96.7	94.8	95.7	95.2
8 mix.	no	96.4	96.4	96.3	96.2	97.3	96.6
full	yes	97.4	97.0	97.4	97.0	97.7	97.3
32 mix.	no	96.6	94.4	96.0	95.2	97.0	95.0
diag.	yes	97.7	95.9	97.4	96.3	97.6	95.3
64 mix.	no	98.0	94.1	97.3	94.9	97.9	96.2
diag.	yes	98.1	95.9	97.9	95.9	97.7	95.8
Verification equal error rate (%)							
4 mix.	no	1.58	2.50	1.48	2.14	1.04	1.31
full	yes	1.04	1.64	0.90	1.33	0.90	0.84
8 mix.	no	0.85	1.66	0.66	1.38	0.42	0.66
full	yes	0.77	1.18	0.58	0.96	0.45	0.52
32 mix.	no	1.01	1.65	0.81	1.29	0.69	0.91
diag.	yes	0.62	1.29	0.52	1.00	0.48	0.95
64 mix.	no	0.60	1.60	0.57	1.20	0.39	0.72
diag.	yes	0.29	1.07	0.29	0.86	0.21	0.60

and WMR test giving the smallest EER of 0.21%.

Significant improvement was obtained for the fast and slow speed test. Thus, the best ML test result for the fast speed is 97.4% compared to the 94.0% of the baseline. The WMR test further improved the result to 98.1% which is very close to the normal speed test results. For the slow speed test we achieved 96.4% (with WMR) from the baseline's 93.0%. The best verification EERs (with WMR) are 0.39% and 0.69% for the fast and slow speeds respectively. The fast and slow speed tests introduce a bigger mismatch between the test data and model distributions, and the significant improvements achieved with these tests show that integrating the LPC-residual information makes the speaker recognition system more robust against variations of the speaking rate.

5.4 Results using both pitch and LPC residual

In these experiments, we added to the best performing CEP + R-CEP 14 dimension vector the pitch parameter, thus increasing the dimension of the voiced vectors to 15. The experimental set up was the same as explained in Section 5.2. Table 4 presents the speaker recognition results using ML, Cohort and WMR tests.

Comparing the results from Table 4 with those from Table 3, we can see that including the pitch parameter further improves the identification rate in most of the cases. The best result is 98.5% of the WMR test. The improvement achieved by including the pitch in addition to the LPC-residual is due to the fact that the cepstral representation of the LPC-residual (4 K-L transformed coefficients) is unable to represent all the pitch information. Therefore, including explicitly the pitch parameter would have effect. However, in the speaker verification experiments, an improvement was observed only for the GMM with 32 mixtures and diagonal covariance matrix.

6. Conclusion

We have introduced a GMM based text independent speaker recognition system, where the pitch and LPC residual were integrated with the standard LPC derived cepstral coefficients.

The experimental results showed that using the pitch information is most effective when the correlation between the pitch and the cepstral coefficients is used.

The combination of the cepstral and LPC residual features is also effective without big difference

Table 4 Speaker recognition rates using CEP and both pitch and R-CEP features.

Mod. type	Using Δ's	ML Test	Cohort Test	WMR Test
Identification rate (%)				
4 mix.	no	96.3	96.2	96.4
full	yes	95.3	96.2	96.3
8 mix.	no	97.5	97.6	97.6
full	yes	97.3	97.3	97.9
32 mix.	no	98.0	97.9	98.3
diag.	yes	96.8	96.9	98.3
64 mix.	no	97.9	98.0	98.5
diag.	yes	96.7	97.9	98.0
Verification equal error rate (%)				
4 mix.	no	2.41	2.20	1.65
full	yes	1.45	1.19	1.29
8 mix.	no	0.90	0.83	0.46
full	yes	0.38	0.39	0.50
32 mix.	no	0.98	0.78	0.38
diag.	yes	0.48	0.47	0.29
64 mix.	no	0.74	0.62	0.44
diag.	yes	0.38	0.34	0.28

among the combination approaches. However, using the LPC residual increases the number of the free system parameters, which sometime cannot be reliably estimated due to limited training data. Including the pitch parameter gives further improvements; these improvements come at the cost of increased system complexity, however.

We have applied our frame level likelihood normalization technique to all cases and, in average, further performance improvements were achieved. Thus, the baseline best identification rate of 97.0% is improved to 97.3% by the WMR technique [13,14] and further to 98.5% by using the LPC residual signal. The corresponding verification equal error rates are 1.07%, 0.52% and 0.21% respectively. Using the same experimental setup and only LPC-cepstral coefficients, Matsui and Furui [2] have reported 95.6% identification rate and 2% verification error rate.

REFERENCES

[1] B. Atal, "Automatic recognition of speakers from their voices," in *Proc. IEEE*, Vol. 64, pp. 460–475, 1976.

[2] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and Discrete/Continuous HMMs," in *Proc. ICASSP*, Vol. II, pp. 157–160, 1992.

[3] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, pp. 18–32, Oct. 1994.

[4] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on SAP*, Vol. 3, No. 1, pp. 72–83, 1995.

[5] T. Matsui and S. Furui, "Text-independent speaker recognition using vocal tract and pitch information," in *Proc. ICSLP*, pp. 137–140, 1990.

[6] M. J. Carey *et al.*, "Robust prosodic features for speaker identification," in *Proc. ICSLP*, pp. 1800–1803, 1996.

[7] M. K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition," in *Proc. EUROSPEECH*, pp. 1391–1394, 1997.

[8] N. Minematsu and S. Nakagawa, "Modeling of variations in cepstral coefficients caused by F0 changes and its application to speech processing," in *Proc. ISCLP*, Vol. 3, pp. 1063–1066, 1998.

[9] S. Nakagawa and T. Sakai, "Feature analyses of Japanese phonetic spectra and considerations on speech recognition and speaker identification," *J. Acoust. Soc. Japan*, Vol. 35, No. 3, pp. 111–117, 1979, (in Japanese).

[10] P. Thevenaz and H. Hugli, "Usefulness of the LPC-residue in text-independent speaker verification," *Speech Communication*, Vol. 17, pp. 145–157, Aug. 1995.

[11] J. He, L. Liu, and G. Palm, "On the use of features from prediction residual signals in speaker identification," in *Proc. EUROSPEECH*, pp. 313–316, 1995.

[12] S. Hayakawa, K. Takeda, and F. Itakura, "Speaker recognition using the harmonic structure of linear prediction residual spectrum," *Trans. IEICE*, Vol. J80-A, pp. 1360–1367, Sept. 1997, (in Japanese).

[13] K. P. Markov and S. Nakagawa, "Text-independent speaker identification utilizing likelihood normalization technique," *IEICE Trans. on Inf. and Sys.*, Vol. E80-D, pp. 585–593, May 1997.

[14] K. P. Markov and S. Nakagawa, "Text-independent speaker recognition using non-linear frame likelihood transformation," *Speech Communication*, Vol. 24, pp. 193–209, June 1998.

[15] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[16] S. Nakagawa, *Speech Recognition Based on Stochastic Model.* IEICE, 1988, (in Japanese).

[17] H. Fujisaki, K. Hirose, and S. Seto, "Proposal and evaluation of a new scheme for reliable pitch extraction of speech," in *Proc. ICSLP*, pp. 473–476, 1990.

[18] A.Ogihara and S.Yoneda, "A method for selecting the most suitable pitch from some candidates utilizing its time continuation," *Trans. IEICE*, Vol. J74-A, No. 7, pp. 948–956, 1991, (in Japanese).

Konstantin P.Markov was born in Bulgaria. He received his B.E. degree in Electrical Engineering from the Department of Cybernetics, Leningrad Polytechnical Institute, Russia in 1984. He joined the Institute of Communication Industry, Sofia, Bulgaria in 1986 as a Research Associate. He received his M.E. and D.E. degrees in Electrical Engineering from Toyohashi University of Technology, Dep. of Information and Computer Sciences, Japan in 1996 and 1999 respectively. He is currently a visiting research engineer in ATR, Japan. His research interests include speech and speaker recognition, language identification and pattern recognition.

Seiichi Nakagawa received his B.E. and M.E. degrees in Electrical Engineering from Kyoto Institute of Technology in 1971 and 1973, respectively, and his Dr. of Eng. degree from Kyoto University in 1977. He joined the Faculty of Kyoto University in 1976 as a Research Associate in the Department of Information Sciences. From 1980 to 1983, he was an Assistant Professor; from 1983 to 1990, he was an Associate Professor; and, since 1990, he has been a Professor in the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi. From 1985 to 1986, he was a Visiting Scientist in the Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, USA. He is the author of *Speech Recognition Based on Stochastic Model* (Inst. Elect. Inform. Comm. Engrs., Japan, 1988). Dr. Nakagawa was a co-recipient of the 1977 Paper Award from the IEICE and the 1988 J.C. Bose Memorial Award from the Institute of Electro. Telecomm. Engrs. His major interesting research areas are automatic speech recognition/speech processing, natural language processing, and artificial intelligence.