TEXT-INDEPENDENT SPEAKER RECOGNITION USING MULTIPLE INFORMATION SOURCES

Konstantin P. Markov *Seiichi Nakagawa*
markov@slp.tutics.tut.ac.jp *nakagawa@tutics.tut.ac.jp*

Dept. of Information and Computer Sciences, Toyohashi Univ. of Technology,
1-1 Hibarigaoka, Tempaku-chou, Toyohashi-shi, Aichi-ken, 441 JAPAN

ABSTRACT

In the speaker recognition, when the cepstral coefficients are calculated from the LPC analysis parameters, the LPC residual and pitch are usually ignored. This paper describes an approach to integrate the pitch and LPC-residual with the LPC-cepstrum in a Gaussian Mixture Model based speaker recognition system. The pitch and LPC-residual are represented as a logarithm of the $F0$ and as a MFCC vector respectively. The second task of this research is to verify whether the correlation between the different information sources is useful for the speaker recognition task. 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 using the correlation with the cepstral coefficients does not have big effect. The best achieved results 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.

1. INTRODUCTION

In the last decade, the research has been focused on using the spectral information, especially the cepstral coefficients, for speaker recognition. There have been several studies, for example [1, 2, 3], 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 in speech. The approach taken in [1], 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. In [3] 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 scores obtained from conventional Gaussian Mixture Model (GMM) cepstral system.

In our speaker recognition system, which is based on GMM, we combine the the cepstral feature vector with the pitch parameter at the frame level. This prompted as to use two models per speaker (as in [1]) for voiced and unvoiced speech segments respectively. Another issue of interest which to our knowledge has not been addressed yet, is whether there is a correlation between the pitch and cepstral coefficients and whether it is useful for the speaker recognition task.

A by-product of the LPC analysis is the prediction error 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. Thus, the prediction error essentially carries all information that has not been captured by the LPC coefficients. In [4, 5] the LPC residual is transformed into cepstral coefficients using FFT - much like MFCC for the speech signal. In [6] the LPC residual is represented in terms of power difference spectrum in subband (PDSS) which is derived also from the FFT spectrum. In [4, 6] the LPC cepstral coefficients and the representation of the LPC residual are treated as a separate feature streams and the scores of the respective models are linearly combined. In contrast, in [5] they are combined at the feature vector level, furthermore, only voiced segments of the speech signal are used for feature extraction.

In our speaker recognition system, the LPC residual is transformed into cepstral coefficients obtained using mel frequency filter bank analysis. We have tried both approaches to combine the conventional LPC cepstral coefficients with 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. Finally, we have experimented with the 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 normalization technique [7, 8], which had a significant effect on our baseline system. Here, we also applied this technique and achieved further improvements of the system performance.

2. FEATURE PARAMETERS

2.1. LPC Residual Cepstrum

The prediction residual signal, according to the LPC model, is found from:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \qquad (1)$$

where α_k are the LPC prediction coefficients, p is the prediction order and $s(n)$ are the samples of the speech signal. It is evident that $e(n)$ might contain information which has

not been captured by the LPC coefficients and which can be useful for the speaker recognition task.

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 DCT.

2.2. Pitch Parameter

The pitch frequency is estimated using an algorithm based on the normalized short-time autocorrelation function which does not require the selection of the frame length [9]. For the minimization of the pitch extraction errors, such as pitch doubling or pitch halving, a post-processing is applied as proposed in [10].

Pitch frequency values are extracted from the digitized speech signal 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 set to 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.

2.3. 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. 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 pitch

In our system, each speaker is represented by two GMMs trained on the corresponding collections of the unvoiced and voiced frames.

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). 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. To overcome this problem, we have chosen to take a linear combination of the likelihoods as follows:

$$L(X) = \alpha L(X_{uv}|\lambda_{uv}) + (1-\alpha)L(X_v|\lambda_v) \qquad (2)$$

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.

3.2. Using LPC residual

As mentioned in Section **2.3.**, 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 modeled by two GMMs - one for CEP and one for R-CEP features. The utterance score in this case is obtained by a linear combination of the two models scores in the same way as Eq.(**2**).

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, it can be captured and used when the model's probability density functions 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 **3.1.**.

4. EXPERIMENTS

4.1. 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. For training the models, 10 sentences for each speaker from one session were used. Five other sentences/session from the other four sessions uttered at normal, fast and slow speeds were used as test data. 10 mel-cepstrum coefficients (CEP) were calculated by the 14th order LPC analysis at every 8 ms with a window of 21.33 ms. Each session's cepstral data were also mean normalized (CMN). Regressive (ΔCEP) coef-

ficients were calculated separately for each of the voiced and unvoiced data streams giving in the same time Δpitch parameters.

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, the ΔCEP and ΔR-CEP are also combined.

4.2. Results using pitch

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. Table **1** compares the

Table 1: Speaker recognition rates using pitch

Model type	Using Δ's	CEP	CEP + pitch			
			ML test			
		ML test	W/o Cor.	With Cor.	Cohort test	WMR test
Identification rate (%)						
4 mix.	no	92.3	93.9	95.3	95.1	96.0
	yes	94.1	93.9	95.3	94.4	96.6
8 mix.	no	96.1	96.3	97.1	96.9	97.7
	yes	97.0	96.8	97.4	97.0	97.6
Verification equal error rate (%)						
4 mix.	no	2.50	2.46	1.66	1.33	0.84
	yes	1.64	2.28	1.45	1.11	0.64
8 mix.	no	1.66	1.48	1.21	0.96	0.50
	yes	1.18	0.98	0.89	0.80	0.41

recognition rates among the baseline ("CEP"), the independent pitch modeling case ("W/o Cor.") and the case when the correlation between the pitch and the cepstral coefficients is used ("With Cor."). In the columns, "ML test" stands for the Maximum Likelihood test. These results show, that the pitch/cepstral correlation is effective and that the gain in the performance is bigger than the case when this correlation is not used.

The columns "Cohort test" and "WMR test" of the Table **1** show the recognition rates when the frame level likelihood normalization technique is applied to the system using pitch/cepstral correlation [8]. 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. 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 was improved to 97.4% with the WMR test. The corresponding rates for the slow speed test are 93.0%, and 96.5%. The verification 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.

4.3. Results using LPC residual

In the first evaluation experiments with LPC residual, it was modeled as a separate feature stream. Each speaker was modeled by a pair of GMMs 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. In the next experiments, the CEP

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

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

and R-CEP vectors were combined into one 20 dimensional feature vector. The results of these experiments are summarized in Table **2** in the column "20 dim.". The poor performance of the 8 mixture, full covariance matrix GMM suggests that probably the training data became insufficient when the model dimension became doubled. Thus, we decided to reduce the R-CEP vectors dimension to 4 using Karuhnen-Loewe (K-L) transformation.

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 diagonalises the covariance matrices. 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 mixture GMM) and 20 dimension feature vector. Obtained results 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 small which confirms the fact that the CEP and R-CEP coefficients hold different information and are almost uncorrelated.

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 baseline (CEP) and CEP + R-CEP (CEP+R) 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 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.

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

Mod. type	Using Δ's	Cohort test		WMR test	
		CEP+R	CEP	CEP+R	CEP
Identification rate (%)					
4 mix. full	no	95.3	92.4	96.1	92.4
	yes	96.7	94.8	95.7	95.2
8 mix. full	no	96.3	96.2	97.3	96.6
	yes	97.4	97.0	97.7	97.3
32 mix. diag.	no	96.0	95.2	97.0	95.0
	yes	97.4	96.3	97.6	95.3
64 mix. diag.	no	97.3	94.9	97.9	96.2
	yes	97.9	95.9	97.7	95.8
Verification equal error rate (%)					
4 mix. full	no	1.48	2.14	1.04	1.31
	yes	0.90	1.33	0.90	0.84
8 mix. full	no	0.66	1.38	0.42	0.66
	yes	0.58	0.96	0.45	0.52
32 mix. diag.	no	0.81	1.29	0.69	0.91
	yes	0.52	1.00	0.48	0.95
64 mix. diag.	no	0.57	1.20	0.39	0.72
	yes	0.29	0.86	0.21	0.60

4.4. Results using both pitch and LPC residual

In these experiments, we added the pith parameter to the best performing CEP + R-CEP 14 dimension vector, thus increasing the dimension of the voiced vectors to 15. The experimental set up was the same as explained in Section **4.2.**. 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 pith parameter further improves the identification rate of all the tests. The best results is 98.5% of the WMR test. However, no improvement was observed in the speaker verification experiments.

5. CONCLUSIONS

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 among the combination approaches. Significant improvement was also obtained for the fast and slow utterances. Including additionally the pitch parameter gives further improvements, however, at the cost of increased system complexity. When the frame level likelihood normalization technique was applied, in average, further performance improvements were achieved.

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

REFERENCES

1. T. Matsui and S. Furui, "Text-independent speaker recognition using vocal tract and pitch information," in *Proc. ICSLP*, pp. 137–140, 1990.
2. M. J. Carey *et al.*, "Robust prosodic features for speaker identification," in *Proc. ICSLP*, pp. 1800–1803, 1996.
3. 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'97*, pp. 1391–1394, 1997.
4. P. Thevenaz and H. Hugli, "Usefulness of the LPC-residue in text-independent speaker verification," *Speech Communication*, vol. 17, pp. 145–157, Aug. 1995.
5. J. He, L. Liu, and G. Palm, "On the use of features from prediction residual signals in speaker identification," in *Proc. EUROSPEECH'95*, pp. 313–316, 1995.
6. 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).
7. K. P. Markov and S. Nakagawa, "Text-independent speaker identification utilizing likelihood normalization technique," *IEICE Transactions on Information and Systems*, vol. E80-D, pp. 585–593, May 1997.
8. 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.
9. K. H.Fujisaki and S.Seto, "Proposal and evaluation of a new scheme for reliable pitch extraction of speech," in *Proc. ICSLP*, pp. 473–476, 1990.
10. A.Ogihara and S.Yoneda, "A method for selecting the most suitable pitch from some candidates utilising its time continuation," *Trans. IEICE*, vol. J74-A, no. 7, pp. 948–956, 1991. (in Japanese).