SIGNAL MODELING FOR SPEAKER IDENTIFICATION
Li Liu, Jialong He, and Günther Palm
Abteilung Neuroinformatik
University of Ulm
89069 Ulm, Germany
li@neuro.informatik.uni-ulm.de

ABSTRACT
A large number of parameters, including pitch, LPCC, ΔLPCC, PARCOR, MFCC, AMFCC, and residual cepstrum (RCEP) were extracted from speech signals and their effectiveness for text-independent speaker identification was evaluated. In addition, the usefulness of two signal processing techniques, preemphasis and cepstral weighting, was also studied. The VQ-based speaker recognition method with codebooks fine-tuned by LVQ algorithm was used. It was shown that both LPCC and MFCC are effective representations, for smaller number of parameters, LPCC representation performs better but is surpassed by MFCC if the analysis order is larger. Pitch is an independent parameter so that it can be used jointly with other spectral features. In an evaluation experiment, the correct identification rate for 112 male speakers with test utterances of less than one second reached 98.2%.

1. INTRODUCTION
Parameterization of a speech signal is the first step in any automatic speaker recognition systems. Most of these techniques are borrowed directly from speech recognition. Both speech and speaker recognition rely primarily on spectral features, but speaker recognition makes use of some other information. A dominant spectral feature representation is the LPC based cepstrum (LPCC). Translational spectral information (ΔLPCC) has been suggested to be complementary to the LPCC and can be used jointly with the LPCC to improve the performance of text-independent speaker recognition systems [1]. Other spectral feature representations such as partial correlation (PARCOR) and mel-scaled FFT based cepstrum (MFCC) have also been used. Pitch period as measured in a contour over time has been found to be useful for text-dependent speaker recognition [2]. More recently, we have extracted a new set of spectral features from pre-diction residual signals (named as residual cepstrum or RCEP for abbreviation) and found them useful for speaker identification [3]. Besides, some commonly used signal processing techniques developed for speech recognition, such as preemphasis and cepstral weighting, are also applied when modeling a signal for speaker recognition.

In this research, we studied the effectiveness of pitch and various spectral feature sets for text-independent speaker identification. The learning vector quantization (LVQ) network served as the classifier. This classifier is inherently text independent and essentially the same as the well-known VQ based method [4], the only difference is that the codebook is further fine-tuned by the discriminative LVQ algorithm after generated by a VQ algorithm.

2. FEATURES
A. LPCC and ΔLPCC
From each frame of speech signals, LPCC and ΔLPCC are derived from the LPC coefficients with the following equations [5]:
\[ c_m = a_m + \sum_{n=1}^{m-1} (n/m)c_n a_{m-n} \quad 1 \leq m \leq p \] (1)

\[ \Delta c_m(t) = \sum_{i=-k}^{k} c_m(t+i) \cdot i \] (2)

where \( \{a_m\} \) are the LPC coefficients, \( p \) is the LPC analysis order. The index \( k \) specifies the number of frames over which the ΔLPCC were calculated (\( k=1 \) in our experiments).

In many speech applications, the speech signal is often pre-emphasized with a first-order FIR filter prior to analysis
\[ H(z) = 1 - \alpha z^{-1} \quad 0.9 \leq \alpha \leq 1.0 \] (3)
The idea is to flatten the spectrum and make it less susceptible to finite precision effects later in the signal processing. For comparison, we calculated spectral features from signals either with ($\alpha=0.97$) or without preemphasis.

The LPC-derived cepstral coefficients decay at least as fast as $1/m$, the cepstral coefficients tend to be concentrated around $m=0$. Since the higher order cepstral coefficients have been found to be as important as the lower order cepstral coefficients in their ability to separate one speaker from the others [1], a number of researchers use the spectral slope distance in order to equalize the contribution from individual cepstral coefficients.

$$d_{\text{slope}} = 2 \sum_{m=1}^{p} m^2 (c_m - c'_m)^2$$

This distance measurement is equivalent to a linearly ramped weighting function.

B. PARCOR

The partial correlation coefficients (PARCOR, also known as reflection coefficients) can be obtained either directly from the LPC equations when they are solved with the Durbin method [5], or from the LPC coefficients using a recursive relation.

C. MFCC and AMFCC

We calculated MFCC coefficients from each frame of signals with 40 mel-scaled filter banks. The details of the procedure can be found in [6]. AMFCC coefficients were calculated in a similar way to the calculation of $\Delta$LPCC coefficients described above.

D. Pitch

The parallel processing algorithm (PPROC) proposed in [7] was used to estimate the average pitch of each frame. The pitch value was then divided by a normalization factor so that it can be used together with other spectral features. The coincidence number, a by-product from this algorithm, was used in conjunction with energy measurement for voiced/unvoiced detection.

E. RCEP

The residual cepstrum (RCEP) derived from prediction residual signals have been found to be useful for both speaker identification [3] and speaker verification [8]. This is not surprising if we think of the fact that the LPC analysis is based on the all-pole assumption which is not perfect for some sounds, thus the prediction residual signals still contain useful information not captured by the LPC based parameters. RCEP were calculated by FFT based method. It should be noted that since RCEP and LPCC are calculated by different methods, when they are used jointly, a proper normalization is necessary. We studied the optimal order of RCEP parameters when used together with the LPCC.

3. DATABASE

The evaluation speech data were selected from the TIMIT [9]. Since the voice difference between male and female is fairly large, it is more reasonable to evaluate a speaker recognition system with data from the same gender. Thus we chose the speech data from 112 male speakers. Two "sa" and five "sx" sentences were used as the training data and the rest three "si" sentences as the test ones. To evaluate the effects of test utterance length, we first concatenated the three test sentences to form a long one and then cut it into several fixed length pieces as separate test utterances. The final sentence level performance is the percent of correctly identified utterances over all test utterances. The analysis window size was 32 ms and the overlap between successive windows was 16 ms. Silence segments were detected and discarded based on an adaptive energy threshold.

4. CLASSIFIER

The main objective of the present research is to evaluate the effectiveness of various signal modeling techniques rather than to achieve the highest recognition rate, so we decided to use the learning vector quantization (LVQ) network as the classifier. Although its performance is not so good as that of a MLP network, it needs much less training time [10]. The initial placement of the code vectors was determined by the LBG vector quantization algorithm. The codebook size in all our experiments was fixed at 16 code vectors per speaker. In the identification phase, each test vector was labeled through the nearest comparison of this vector with all code vectors. The frame identification rate is the percentage of vectors that have been correctly classified, which gives the speaker identification performance with short segments of signals (32 ms in our case). The classification decision for a test sentence is made by finding the minimum accumulated distance.

5. EXPERIMENT RESULTS

The speaker identification performances (frame & sentence level) as a function of feature vector dimension are shown in Figure 1. It is seen that identification rates at frame-level increase monotonically with the vector di-
mension. This tendency is more clear when the vector dimension is less than 24. On the other hand, the sentence level performance rises quickly at the beginning until it reaches the level of about 95%, after that, no significant improvement can be made by adding more parameters. Hence simply increasing the analysis order is not an efficient way to squeeze more information from signals. Another interesting result evidenced in Fig. 1 is that the LPCC provides a better performance for lower vector dimensions, but are surpassed slightly by the MFCC when the vector dimension is large than 26, suggesting that speaker related information concentrates more on the lower order LPCC parameters, but distributes somewhat evenly over the MFCC parameters. Certain kinds of combinations of the LPC based method and the FFT based method might be more effective to improve the recognition performance. This leads to our idea of calculating some RCEP coefficients from the prediction residual signals [3]. At last, we can see from Fig. 1 that the simple cepstral weighting technique of multiplying each cepstral coefficient by its index to compensate the decay of LPC-derived cepstral coefficients degrades the frame level performance, even though the sentence level performance is not seriously affected. Normally, the weighting function should be estimated from the training data. A linear ramp is here clearly not a proper approximation to the weighting function.

Figure 1: speaker identification performances at both frame level (solid lines) and sentence level (dashed lines) as a function of feature vector dimensions. The sentence level identification rates were obtained from 60 frames (960 ms).

It has been reported by several researchers that speaker identification performance increases with the length of test sentences. Our experimental results support this claim. Figure 2 shows the identification performance as a function of test utterance length. As expected, the performance improvement at sentence level was dramatic as the test sentence length increasing. Furthermore, even though the performance with vector dimension 32 is better than that with 16, this difference becomes smaller for longer test utterances. This demonstrated that high sentence level performance can be reached if sufficient long test utterances are available, regardless what type and how many spectral parameters are used. Thus when comparing the performances of different systems, the length of test sentences should be an important criterion.

Figure 2: sentence level performances as a function of test utterance length. Solid lines are the performances with feature vector dimension 32 and dashed lines with dimension 16.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Frame Training (%)</th>
<th>Frame Test (%)</th>
<th>Sentence Test (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPCC</td>
<td>58.0</td>
<td>26.9</td>
<td>92.0</td>
</tr>
<tr>
<td>MFCC</td>
<td>54.6</td>
<td>24.1</td>
<td>90.2</td>
</tr>
<tr>
<td>ΔLPCC</td>
<td>21.4</td>
<td>3.4</td>
<td>12.5</td>
</tr>
<tr>
<td>ΔMFCC</td>
<td>21.3</td>
<td>3.1</td>
<td>8.8</td>
</tr>
<tr>
<td>LPCC+</td>
<td>60.1</td>
<td>28.5</td>
<td>92.9</td>
</tr>
<tr>
<td>Pitch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC+</td>
<td>56.8</td>
<td>25.5</td>
<td>91.1</td>
</tr>
</tbody>
</table>

Table 1: Performance of different spectral features (16 parameters) and a pitch period.

Table 1 compares the identification results using different feature sets. To our surprise, the performances with ΔLPCC and ΔMFCC are very low. Soong and Rosenberg [1] have demonstrated that ΔLPCC do carry speaker relevant information for fix vocabulary (digits) even though they are less effective than LPCC. The extremely low performance with the dynamic feature sets (ΔLPCC & ΔMFCC) in our experiments might be due to the different vocabulary used in training and test data, that is, our system was run in real text-independent mode. Pitch is usually used in text-dependent speaker recognition systems to form contours [2], it was shown in our experi-
ments that pitch is an independent parameter and can be quantized together with other spectral features to improve the performance of text-independent speaker recognition.

The effect of preemphasis technique was also studied (Table 2). For comparison, results from a set of composite features with 30 LPCC coefficients plus 2 RCEP coefficients is also included in the table. It can be seen that the performance is not always improved at frame level but is higher at sentence level if the preemphasis technique is applied. Since the computation load of preemphasizing a signal is very light, it is worth to do before extract spectral features. We also found that the performance were significantly improved if the LPCC coefficients are used jointly with two RCEP coefficients.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>With preemphasis</th>
<th>Without preemphasis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frame</td>
<td>Sentence</td>
</tr>
<tr>
<td>LPCC</td>
<td>35.2</td>
<td>97.3</td>
</tr>
<tr>
<td>MFCC</td>
<td>36.5</td>
<td>97.3</td>
</tr>
<tr>
<td>PARCOR</td>
<td>28.7</td>
<td>97.3</td>
</tr>
<tr>
<td>LPCC+RCEP</td>
<td>39.1</td>
<td>98.2</td>
</tr>
</tbody>
</table>

Table 2: effect of preemphasis on the performance. The feature vector dimension is 32, the composite feature set consisting 30 LPCC and 2 RCEP.

6. CONCLUSIONS

In this study, we investigated several issues of modeling speech signals for speaker recognition, including the effectiveness of various spectral features, pitch, preemphasis, and cepstral weighting. The following is a summary of the conclusion remarks:

- Both LPCC and MFCC are effective spectral feature representations. LPCC performs better for smaller number of parameters but can be surpassed slightly by MFCC when the analysis order is large than some value (25 in our case).
- ΔLPCC and ΔMFCC are not suitable for text-independent (unconfined vocabulary) speaker identification.
- Pitch is an independent feature, i.e., it can be used jointly with other spectral features in text-independent speaker recognition.
- Preemphasizing a signal by the first-order FIR filter enhances the effectiveness of spectral features.
- Cepstral weighting by multiplying each LPCC coefficient with its index degrades the frame-level performance.
- Performance distinction of different static spectral representations at sentence level is large when the feature vector dimension is lower and/or test utterances are short, but becomes negligible with long test sentences and relative higher analysis order.
- Residual cepstrum (RCEP) contain extra information not captured by the LPC based parameters, they can be used jointly with LPCC to further improve the performance. A proper number of RCEP parameters, when used with the LPCC, is two.

7. REFERENCES