USE OF VOICING AND PITCH INFORMATION FOR SPEAKER RECOGNITION

Brett R. Wildermoth and Kuldip K. Paliwal
School of Microelectronic Engineering, Grif£th University,
Brisbane, QLD 4111, Australia

ABSTRACT: Speech signal can be decomposed into two parts: the source part and the system part. The system part corresponds to the smooth envelope of the power spectrum and is used in the form of cepstral coef£cients in almost all the automatic speaker recognition systems reported in the literature. The source part contains information about voicing and pitch. Though this information is very important for human beings to identify a person from his/her voice, it is rarely used for automatic speaker recognition. In this paper, we propose a simple and reliable method to derive acoustic features based on voicing and pitch information and use them for automatic speaker recognition. We evaluate these features for speaker identi£cation using TIMIT, NTIMIT and IISC databases and demonstrate their effectiveness.

INTRODUCTION

A speech signal can be decomposed into two parts: the source part and the system part. The system part consists of the smooth envelope of the power spectrum and is represented in the form of cepstrum coef£cients, which can be computed by using either the linear prediction analysis or the mel £lter-bank analysis. Most of the automatic speaker recognition systems reported in the literature utilise the system information in the form of cepstral coef£cients. These systems perform reasonably well. The source information has been rarely used in the past for speaker recognition systems. The source contains information about pitch and voicing. This information is very important for humans to identify a person from his/her voice. A few studies have been reported where pitch information is used as a feature for speaker recognition. However results are not very encouraging. The main reason for this is that pitch estimation is always very much prone to errors. That is the pitch estimation methods are not very reliable, they introduce errors which affect the performance of the speaker recognition system. In this paper we propose a simple method for extracting the voicing and pitch information from the speech signal in a reliable manner. This is done by uniformly dividing the higher portion of the autocorrelation function in a number of parts and computing the maximum autocorrelation value in each of these parts. These maximum autocorrelation values (MACVs) are used as features for speaker recognition. We evaluate these MACV features on TIMIT, NTIMIT and IISC databases for speaker identi£cation task. In order to put these features in proper perspective, we compare their speaker identi£cation performance with that of pitch feature.

COMPUTATION OF PITCH FEATURE

As mentioned earlier, we compare the speaker identi£cation performance of the MACV features with the pitch feature. For determining the pitch value, we use two different methods: 1) the autocorrelation method and 2) the average magnitude difference function (AMDF) method.

Consider a speech frame $\{s(n), n = 0, 1, ..., N-1\}$. In the autocorrelation method, the autocorrelation function of the speech signal $\{s(n)\}$ is computed as follows:

$$R(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} s(n)s(n+k), \qquad k = 0, 1, ..., N-1. \tag{1}$$

Since the human pitch values normally range from 2 ms to 16 ms, this autocorrelation function is searched for a peak in this range and location of the peak de£nes the pitch value. The autocorrela-

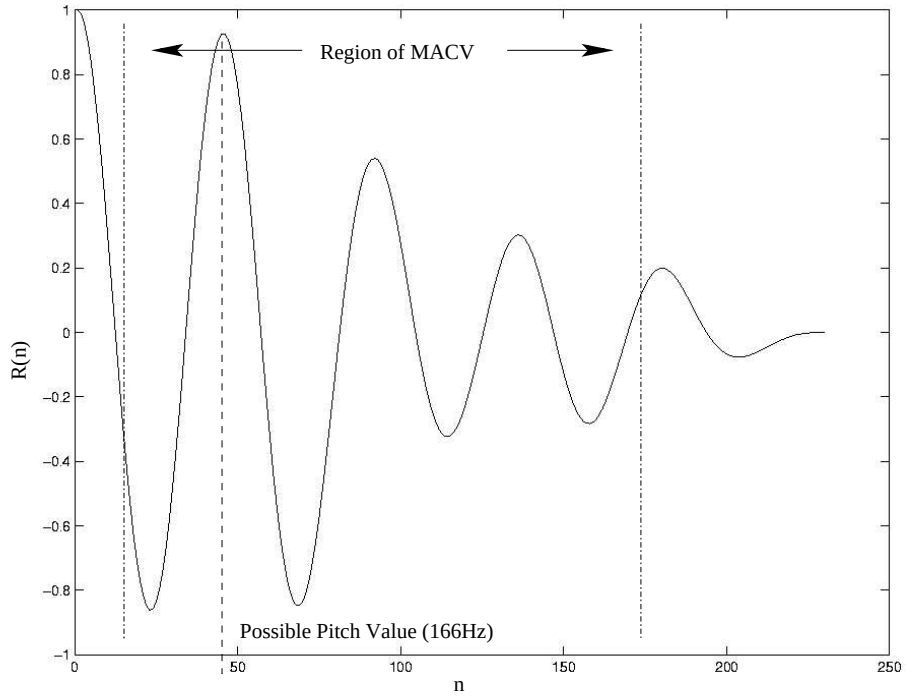

Figure 1: The autocorrelation function of 30 ms segment of vowel sound ee.

tion function of 30 ms segment of vowel /i/ sampled at 8 kHz is shown in Fig. 1. The vertical dashed line shows the peak location.

In the AMDF method, the AMDF function is computed as follows:

$$A(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} |s(n) - s(n+k)|, \qquad k = 0, 1, ..., N-1. \tag{2}$$

The AMDF function is searched for a minimum in the range of 2 ms to 16 ms. The location of minimum de£nes the pitch value.

COMPUTATION OF MACV FEATURES

Given the speech signal $\{s(n)\}$, the MACV features are computed as follows:

1. Compute the autocorrelation function $\{R(n)$ from the speech signal using Eq. (1).
2. Normalise the autocorrelation function by its value at $n = 0$, i.e.,

$$r(n) = \frac{R(n)}{R(0)}. \tag{3}$$

3. Discard the lower portion of the autocorrelation function as it contains the information about the system component of speech and is used in the speaker recognition systems in the form of cepstral coef£cients. Using only the higher portion (from 2 ms to 16 ms) of the autocorrelation function, compute the MACV features as follows:
   i. Divide the higher portion of the autocorrelation function into N equal parts (typically N=5).
   ii. Find the maximum value of the normalized autocorrelation for each of the N divisions.

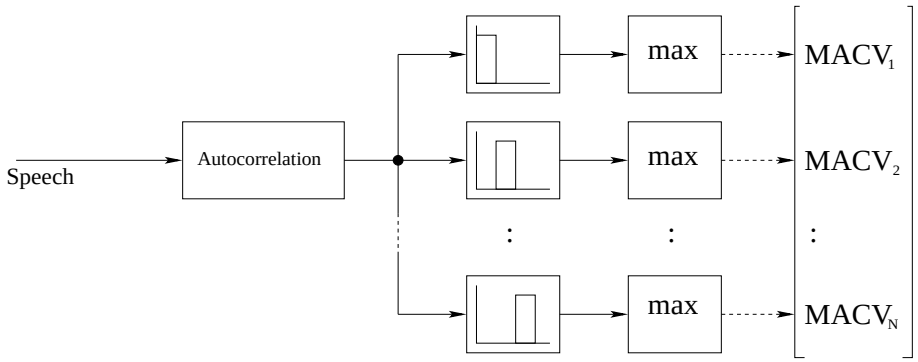

Figure 2: MACV feature extractor.

iii. These N Maximum Autocorrelation Values (MACV) correspond to N MACV features which are utilized for speaker recognition.

A block diagram of this process is shown in Figure 2.

Note that the MACV features sample the higher portion of the autocorrelation function at N points. In principle, we can use the entire higher portion of the autocorrelation function for speaker identi£cation. But, it is not possible to do because it will make the dimensionality of the feature space too high. The MACV features basically model the higher portion of the autocorrelation function in terms of N parameters.

SPEAKER IDENTIFICATION EXPERIMENTS

In our experiments, the MACV features are tested alone as well as in combination with the ceptral coef£cients derived through the LPC analysis. A text-independent Gaussian mixture model (GMM) system is used as test bed for evaluating the speaker identi£cation performance of these features.

The speaker identi£cation experiments are carried out using the TIMIT, NTIMIT and IISC data bases. TIMIT and NTIMIT databases are standard data bases available from the linguistic data consortium. They are described in references (Reynolds,1993) and (Reynolds,1995). The IISC database consists of 43 male and 37 female speakers from various regions of India. The database consists of both isolated words and continuous speech. The original database was recorded using a high quality microphone at 16 kHz with a resolution of 16 bits. The recording environment was noise free. We refer to this data base as IISC-Microphone. The recorded speech from the IISC-Microphone database was transmitted over a mobile and cordless phone to form IISC-Mobile and IISC-Cordless databases, respectively. The IISC-Mobile database was down-sampled to 8kHz due to the reduced bandwidth implied by the GSM coder of the mobile phone.

The performance of the system is measured using identi£cation error described as:

$$\%\ identification\ error = \frac{\#\ incorrectly\ identified\ segments}{total\ \#\ of\ segments} \times 100\% \tag{4}$$

In order to illustrate the positive effect of the MACV features, we also show in our results the reduction in identi£cation error which is can be calculated with respect to LPCC features (for example) as follows:

$$\%\ reduction = \frac{\%\ identification\ error_{LPCC} - \%\ identification\ error_{MACV}}{\%\ identification\ error_{LPCC}} \times 100\% \tag{5}$$

A speech segment of 24 seconds is used to train the system and utterances of 3 seconds used for testing. Using the IISC databases, 3 tests are performed per speaker while 2 tests are performed for

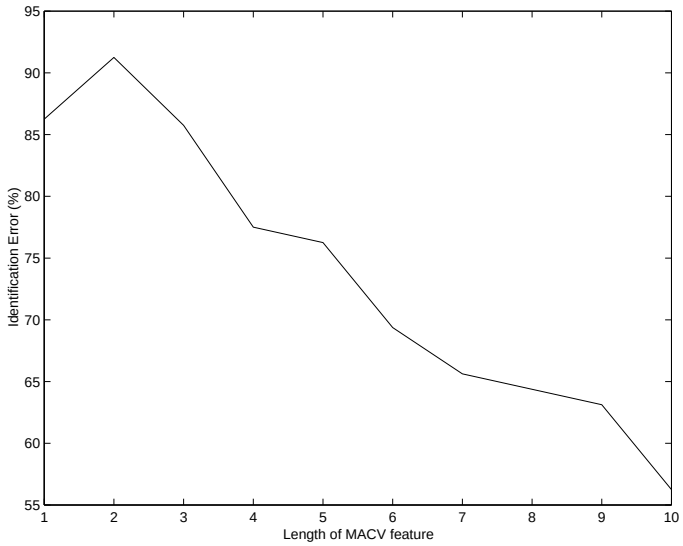

Figure 3: Speaker identi£cation error as a function of number of MACV features when used alone.

Feature Used	IISC-Microphone Id. Error (%)	IISC-Cordless Id. Error (%)	IISC-Mobile Id. Error (%)	TIMIT Id. Error (%)	NTIMIT Id. Error (%)
Pitch (1)					
Auto. method	91.3	90.0	91.9	86.9	92.1
AMDF method	90.0	98.3	93.1	82.9	79.0
$MACV(5)$	76.3	90.0	80.6	57.9	61.8

Table 1: Comparison of source based features

each TIMIT/NTIMIT speaker. For feature extraction, the speech is £ltered using a pre-emphasis £lter denoted by $H(z) = 1 - 0.95z^{-1}$ and split into 30ms frames with a 10ms update. The segments are then windowed using a hamming window.

For speaker identi£cation experiments involving the MACV and LPCC features, each speaker is modeled using a 32 mixture GMM. For the experiments with pitch, only 8 GMM mixtures per speaker are used. For training purposes, the models are initialised using a k-means algorithm and further optimised using the EM algorithm (Reynolds,1993). A variance ¤oor of 0.03 is imposed on the models (Reynolds,1995).

For 5 MACV features, the speaker identi£cation errors are listed in Table 1 for the TIMIT, NTIMIT and IISC databases. For providing comparison, the speaker identi£cation experiment is run with one pitch feature. Two different methods (the autocorrelation and AMDF method) are investigated for pitch estimation. The speaker identi£cation performance with the pitch feature is also shown in Table 1. It can be seen from this table that the use of pitch as a feature performs poorly when compared with the MACV features. The MACV features reduced the identi£cation error by 30% in comparison to pitch feature. In order to see whether the MACV features contain any additional information not represented by the LPCCs, we have conducted speaker identi£cation experiments on all the databases using the MACV features in combination with the LPCC features. Here, we use 12 LPCC features and 5 MACV features. The results are listed in Table 2. It can be seen that use of the MACV features with the LPCC features improves the speaker identi£cation performance. It reduces the speaker identi£cation error by 45.5% for the IISC-Microphone database and by 39.2% for the NTIMIT database. This reduction in identi£cation error is quite signi£cant.

In order to show the effect of the number of MACV features on speaker identi£cation performance, we carry out experiments on the IISC-Microphone database using the MACV features alone and in combination with the LPCC features. Results are shown in Figures 3 and 4, respectively. We can observe from this table that the speaker identi£cation error reduces with the increase in the number of

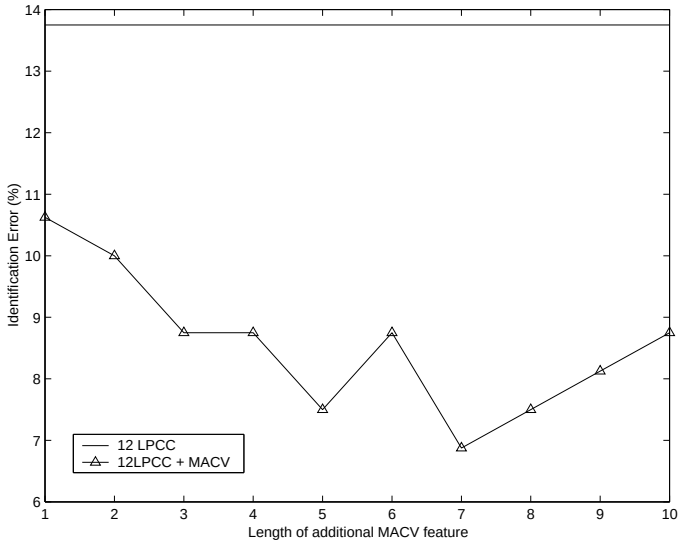

Figure 4: Speaker identi£cation error as a function of number of MACV features when used with LPCC.

Database Used	12 LPCC Id. Error (%)	12 LPCC + 5 MACV Id. Error (%)	Reduction in Id. Error (%)	5 MACV Id. Error (%)
IISC-Microphone	13.8	7.5	45.5	76.3
IISC-Mobile	6.3	5.6	10.0	80.6
IISC-Cordless	21.3	15.6	26.5	90.0
NTIMIT	21.6	13.2	39.2	61.8

Table 2: Performance of the MACV and LPCC features

MACV features when used alone and in combination with the LPCC features.

CONCLUSIONS

In this paper, we have proposed a simple and reliable method of extracting voicing and pitch information from the speech signal in the form of MACV features. We have shown that these features are more effective than the pitch feature for speaker identi£cation. When used in combination with the LPCC features, these features can reduce the speaker identi£cation errors signi£cantly.

ACKNOWLEDGMENTS

The authors would like to thank Professor T.V. Sreenivas for providing the IISC database, on which these experiments were performed.

REFERENCES

Reynolds, D. A. & Rose R. C. (1995), ”Robust Text-Independent Speaker Identi£cation Using Gaussian Mixture Speaker Models”, IEEE transaction on Speech and Audio Processing, Vol 3, No. 1 January 1995, 72 - 83.

Matsui, T. & Furui, S. (1990) ”Text-Independent Speaker Recognition using Vocal Tract and Pitch Information” , Proc. ICSLP 1990, 137-140.

Reynolds, D. A. (1993), ” A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identi£cation”, Phd Thesis, Massachusetts Institute of Technology, Lincoln Laboratory, 28-37.

Reynolds, D. A. (1995), ”Speaker identi£cation and veri£cation using Gaussian mixture speaker models”, Speech Communication 17, 91-108.