SPEAKER VERIFICATION USING FUNDAMENTAL FREQUENCY

Yoik Cheng   *Hong C. Leung*

Department of Electronic Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong

Email: ycheng@ee.cuhk.edu.hk, hcleung@ee.cuhk.edu.hk

ABSTRACT

This paper describes the use of speech fundamental frequency (F0) for speaker verification. Both Chinese and English have been included in this study, with Chinese representing a tonal language and English representing a non-tonal language. A HMM-based speaker verification system has been developed, using features based on cepstral coefficients and the F0 contour. Four different techniques have been investigated in our experiments on the YOHO database and a Chinese speech database similar to YOHO. It has been found that the pitch information results in a reduction of the equal error rates by 40.5% and 33.9% in Cantonese and English, respectively, suggesting that the pitch information is important for speaker verification and that it is more important for tonal languages. We have also found that the pitch information is even more effective when it is represented in the log domain, resulting in an ERR of 2.28% for Cantonese. This ERR corresponds to a reduction of the ERR by 54%.

1. INTRODUCTION

Many techniques have recently been applied to speaker verification [1-7]. Some of these techniques use cohort normalization [1,2] to identify groups of speakers with similar speaking characteristics. Other techniques use adaptive training methods to update the speaker's models so that the discriminability among speakers is increased [4,5]. Furthermore, techniques incorporating pitch information have also been proposed [6,7] based on the premises that the pitch contour gives information about the identity of the speaker and that pitch information is less affected by the frequency characteristics of the transmission system when compared with spectral information.

In a tonal language such as Chinese, two different words can have exactly the same base phonetic description of different tones. Since the tones are highly correlated with the pitch contour, the use of pitch information is particularly interesting for speech recognition or speaker verification in a tonal language.

In this study, we will examine the use of pitch information for speaker verification in Cantonese Chinese and English. There are several objectives. First, the importance of pitch information for speaker verification will be studied. Second, the impact of pitch information on speaker verification in these two languages will be compared. Finally, the representation of the pitch information for speaker verification will also be explored.

The organization of this paper is as follows. Section 2 summarizes the two speech databases used for our study. Section 3 describes the feature analysis techniques for pitch and spectral information. The HMM-based speaker verification system is described in Section 4. The experimental techniques are presented in Section 5, with the results reported in Section 6. Some discussions will be presented in Section 7. Finally, Section 8 summarizes our conclusion.

2. DATABASE

A Cantonese speech database was collected in home/office environments with varying telephone handsets through the public service telephone network in Hong Kong. The database was designed to be similar to the YOHO corpus [8], which was designed for speaker verification purpose with a limited vocabulary.

This Cantonese speech database is consisted of 56 two-digit numbers ranging from 21 to 97. There are 51 male speakers and 52 female speakers, resulting in a total of 103 speakers. For each speaker, there are 4 enrollment sessions of 24 phrases each, and 10 verification sessions of four phrases. Thus there are altogether 9,888 phrases in the enrollment set, and about 4,120 phrases in the verification set. Each phrase is consisted of a sequence of three two-digit numbers, such as 23-72-45.

The database recordings were digitized at an 8 kHz sampling rate using an 8-bit μ-law codec. The 8-bit μ-law digitized samples were finally converted to linear 16-bit PCM samples before further processing.

The English database used in our study was the YOHO speaker verification corpus [8]. The database is consisted of a set of randomly arranged two-digit phrases. Each phrase is also consisted of a sequence of three two-digit numbers. There are 106 male speakers and 32 female speakers, resulting in a total of 138 speakers. For each speaker, there are 4 enrollment sessions of 96 phrases each, and 10 verification sessions of 40 phrases.

3. FEATURE ANALYSIS

The speech utterances were digitized with a bandpass filter ranging from 0.3 kHz to 3.4 kHz. The digitized utterances were pre-emphasized with a first order difference digital network, and were converted to 14th order linear predictive coding (LPC) coefficients. These LPC coefficients were then converted to 12th order cepstral coefficients. When combined with the 12th order delta and 12th order delta-delta cepstral coefficients, a feature vector of 36 dimensions was obtained.

To augment the cepstral coefficients, two kinds of prosodic features were appended: log energy contour and F0 contour of the utterances. The F0 contour was obtained for every frame of speech using the normalized cross correlation function and dynamic programming [10]. Both delta and delta-delta terms of the log energy and F0 were calculated.

4. HMM-BASED SPEAKER VERIFICATION

Experiments were conducted using a continuous density Gaussian mixture, hidden Markov model (HMM) speaker verification system. Speakers were represented by a set of left-to-right HMM models. Each model contained 8 Markov states, with each state containing a maximum of 4 mixture components. For each of the two languages, a model was created for each of the following digits and "decades": "1" through "9", and "10", "20", "30", "40", "50", "60", "70", "80", "90".

The two-digit numbers in Cantonese have two possible pronunciations. For example, "thirty-six" can be pronounced as "saam-sap-luk" or "saam-ah-luk". However, only one HMM model was used for both pronunciations.

During recognition, a forced Viterbi alignment was utilized to obtain the likelihood score for each testing utterance. We assumed that speakers were giving the true text. Thus, the verification score takes the form of

$$\Lambda(O) = \log p(O \mid \lambda) ,$$

where O is the utterance and λ is the set of models comprising the digits and decades in the utterance. An acceptance or rejection decision was made accordingly if $\Lambda(O)$ was greater than an assigned threshold.

5. EXPERIMENTS

For each speaker in the database, two kinds of verifications were conducted. For each speaker, a test utterance of the true speaker was scored against the true speaker models to obtain the true speaker score. Also, the test utterances of the impostor speakers (impostors) were scored against the true speaker models to obtain a set of impostor speaker scores.

Priori thresholds were assigned and individual equal error rates were calculated. The true speaker scores and the false speaker scores were sorted to determine the threshold for which the fraction of true speaker scored less than the threshold was equal to the fraction of false speaker scored greater than the threshold. This threshold is referred to as the equal error threshold and the corresponding fraction is known as the equal error rate (ERR). ERR is a convenient predictor of verification performance.

The following experimental techniques and conditions were studied to optimize the representation for the F0 information.

5.1 F0 Value

The F0 value, measured in Hertz, was directly used as a parameter value for each frame of speech. Augmented with the delta and delta-delta terms, a feature vector of 42 dimensions was formed for each frame. The same procedures were applied to both the English and Cantonese databases.

5.2 Log F0 Value

Instead of directly using the estimated raw F0, a logarithm form of F0 was used. The log value was used to calculate the delta and delta-delta terms. Similar to the raw F0, a feature vector of 42 dimensions was formed for each frame.

5.3 Normalized F0

After estimating the F0 values for all training utterances, a speaker-dependent mean F0 was calculated from those F0 values. For each utterance within the training set of that speaker, the F0 value of each frame was normalized by this speaker-dependent mean F0. Then the normalized F0 was treated as a parameter value and was used to calculate the delta and delta-delta terms.

During verification process, the F0 value for each frame of an utterance was normalized by the speaker-dependent mean F0 belonging to the claimed speaker. In other words, the F0 contour of the test utterances from either the true speaker or the impostor was normalized with the true speaker's mean F0.

5.4 Normalized Log F0

The F0 was normalized similar to the above method, except that the normalization procedure was performed in the log F0 domain.

6. RESULTS

The results for Cantonese and English are summarized in Figures 1 and 2, respectively. The symbol "M⇒M" stands for the male true speaker model scoring against utterances spoken by the male impostors. "M⇒MF" stands for the male true speaker model scoring against utterances spoken by the male and female impostors. Similarly, "F⇒F" and "F⇒MF" stand for the female true speaker model scoring against utterances spoken by the female impostors and impostors of both genders, respectively.

Figure 1 shows that the pitch information generally improves the ERRs in Cantonese. For example, when tested on male impostors, the true male speaker ERR decreases from 6.2% to 4.5% when pitch information is available. Similarly, when tested on both male and female impostors, the true male speaker ERR decreases from 7.24% to 3.91%. However, when tested on female impostors only, the true female speaker ERR slightly increases from 2.38% to 2.52%. Nevertheless, when tested on both male and female impostors, the corresponding true female speaker ERR decreases from 2.66% to 1.97%. Overall, the ERR decreases from 4.93% to 2.93% for all male and female speakers, corresponding to a reduction by 40.5%.

A similar trend is observed in Figure 2 for English. The ERRs decreases in all cases. Similar to our Cantonese results, the only exception is for the true female speakers tested on female impostors (F⇒F) only. In this case, the ERR also increases slightly from 3.5% to 3.83%. Overall, the ERR decreases from 4.31% to 2.85% for all male and female speakers, corresponding to a reduction by 33.9%

Figure 3 compares the Cantonese results for the four F0 parameter sets and the parameter set without F0. All the four different techniques reduce the overall ERRs when compared to no F0. With the exception of F0 in Hz, the ERR is reduced for all combination of male and female true and impostor speakers. The best overall ERR of 2.28% is obtained when the pitch contour is represented in the log domain. When compared with the system with no pitch information, the ERR is reduced by 54%.

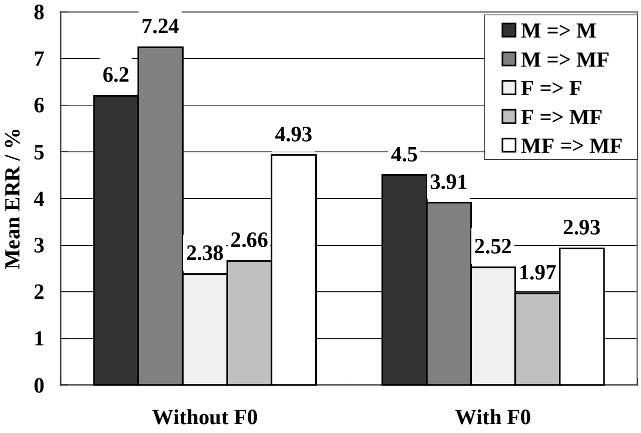

Figure 1: Mean ERRs for the Cantonese database using F0 and no F0. See text for details.

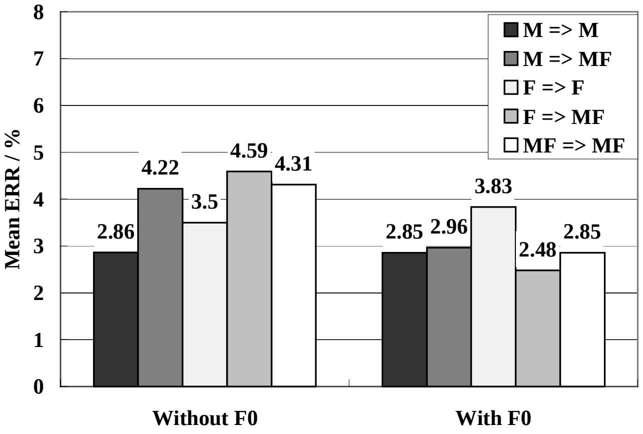

Figure 2: Mean ERRs for the YOHO database using F0 and no F0. See text for details.

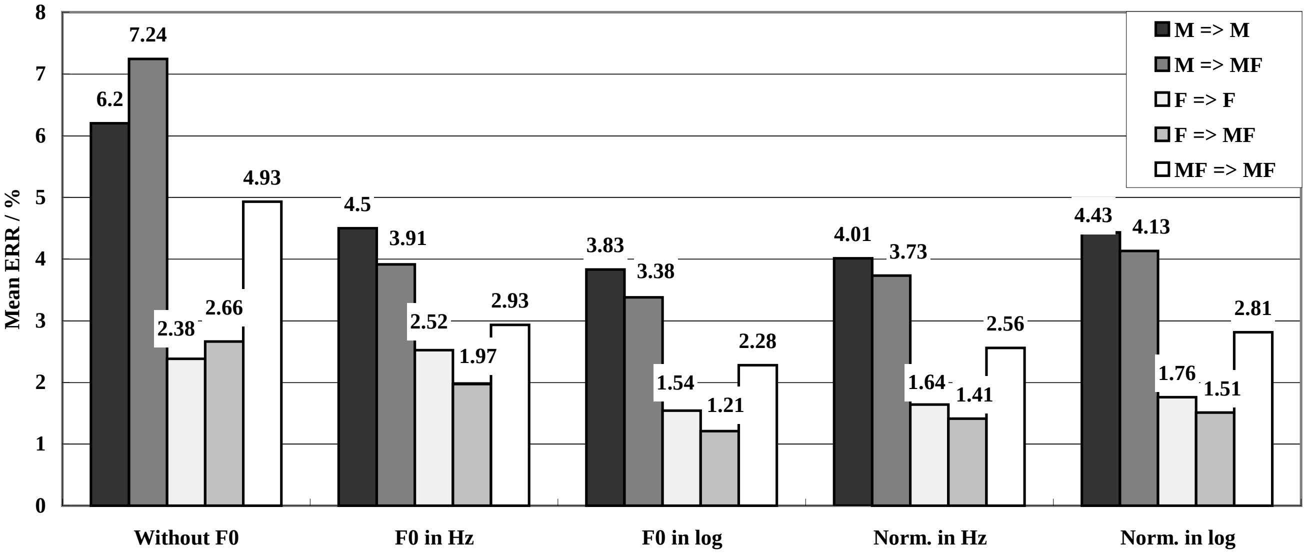

Figure 3: Mean ERRs for the five different conditions for the Cantonese database.

7. DISCUSSION

The main objective of this study is to determine the effectiveness of F0 for speaker verification. Our results show that pitch information reduces the ERR for both Cantonese and English.

It is also observed that the reduction of ERR is more significant in Cantonese than in English. The reductions in ERRs when raw F0 values are 40.5% and 33.9% in Cantonese and English, respectively. We suspect that the larger reduction of ERR in Cantonese is due to the tonal nature of the language.

With the exception of F⇒F, the ERR is consistently reduced for the male and female speakers and models. We suspect that the increase in ERR for F⇒F is due to errors in the pitch estimation.

We have found that the use of pitch information is most effective when the pitch contour is represented in the log domain. This seems to be in agreement with the cepstral representation. However, we have also found that pitch normalization may not improve the ERR, possibly due to the fact that such a normalization may actually remove specific speaker information.

8. SUMMARY

We have described our speaker verification experiments on speech databases in Cantonese and English collected over the telephone. Our results indicate that the pitch contour in raw F0 values reduces the error rate for Cantonese and English by 40.5% and 33.9%, respectively, suggesting that pitch information is improves speaker verification performance and that such information may be more important for a tonal-language. Finally, we have found that representing the pitch contour in the log domain is more effective than other techniques that we have investigated. The lowest ERRs we have obtained is 2.28%.

9. REFERENCE

1. A. E. Rosenberg, J. DeLong, C. H. Lee, B. H. Juang and F. K. Soong, “The Use of Cohort Normalized Scores for Speaker Verification,” *Proc. ICSLP-92*, pp. 599-602, 1992.
2. A. E. Rosenberg and S. Parthasarathy, “Speaker Background Models for Connected Digit Password Speaker Verification,” *Proc ICASSP-96*, pp. 81-84, 1996.
3. A. R. Setlur, R. A. Sukkar and M. B. Gandhi, “Speaker Verification Using Mixture Likelihood Profiles Extracted from Speaker Independent Hidden Markov Models,” *Proc. ICASSP-96*, pp. 109-112, 1996.
4. T. Matsui, T. Nishitani and S. Furui, “Robust Methods of Updating Model and a Priori Threshold in Speaker Verification,” *Proc. ICASSP-96*, pp. 97-101, 1996.
5. J. He, L. Liu and G. Palm, “A Discrimative Training Algorithm for Gaussian Mixture Speaker Models,” *Proc. Eurospeech-97*, pp. 959-962, 1997.
6. B. S. Atal, “Automatic Speaker Recognition Based on Pitch Contours,” *J. Acoust. Soc. Am.*, Vol. 52, pp. 1687-1697, 1972.
7. S. Furui, “Research on Individuality Features in Speech Waves and Automatic Speaker Recognition Techniques,” *Speech Communication*, Vol. 5, pp. 183-197, 1986.
8. J. Campbell, “Testing with the YOHO Speaker Verification Corpus,” *Proc. ICASSP-95*, pp. 341-344, 1995.
9. Rabiner, L. R. and Juang, B. H., “*Fundamental of Speech Recognition,*” Prentice Hall, New Jersey, 1993.
10. Talkin, D., “A Robust Algorithm for Pitch Tracking,” chapter 14, *Speech coding and Synthesis*, New York, 1995.
11. Entropic Research Laboratory, Inc., Washingtion, DC, “*HTK – Hidden Markov Model Toolkit,*” v2.1ed., 1997.