2016 12th World Congress on Intelligent Control and Automation (WCICA)
June 12-15, 2016, Guilin, China

The Influence of Fundamental Frequency on Speaker Recognition System

Yi Zhang, Yanyi Xie, and Kejia Wang

Abstract—The impact of fundamental frequency of MFCC usually is neglected in the application and research of speech signal based on feature MFCC. Analysis shows that the fundamental frequency affects the accurate description of channel characteristics, thus affect the performance of speaker recognition system, so a new speech signal feature named SMFCC (MFCC smoothing) is proposed based on the smoothed amplitude spectrum. Experiments on the YOHO speaker recognition database show that the performance of SMFCC is much more precise than MFCC. Especially in the female speaker data set, the performance of SMFCC is obvious, and SMFCC has better robustness.

[1] Introduction

MEL frequency spectrum coefficient (MFCC) is one of the short term acoustic characteristic parameters[1,2] widely used in speaker recognition system. A set of triangular filters is used to filter the short-time amplitude spectrum of the speech signal. The center frequency and bandwidth of each filter simulate the auditory perception of the human ear. That is, the frequency resolution is high in the low frequency band, and low in the high frequency band. A large number of experiments show that MFCC can achieve good recognition performance in speech recognition and speaker recognition. Since MFCC was put forward, there were many studies from different angles to analyze and improve it. For example, in order to enhance the robustness to additive noise, AMFCC (MFCC autocorrelation) was extracted from autocorrelation sequence[3]. HFCC (factor cepstral coefficents human) was proposed from the angle of auditory perception to reset the bandwidth of each filter[4]. In order to improve the system recognition rate of speaker recognition system, information of speech voicing and pitch frequency DMCEP (dynamic Mel cepstrum) was proposed[5]. Literature[6] proposed WDFT-WFBA-MFCC (DFT weighted, filter bank wrapped MFCC analysis) to improve the spectral resolution of the low frequency part of the speech signal and improve the robustness of the MFCC. In literature, MFCC was improved from the point of view of Chinese speech recognition[7]. The influence of different filter bank on speech recognition performance was compared[8].

MFCC is the short-time spectral magnitude based on the extraction of the voiced speech signal features. For voiced speech signal, the amplitude spectrum can be regarded as the sampling of the spectral envelope of the harmonics, which contains the spectral envelope and pitch frequency information[9]. For the male speaker with low frequency pitch, Mel filter can smooth voiced amplitude spectrum, to eliminate under 1kHz is about 200 Hz. the filter output is heavily affected by harmonic[10]. In addition to the characterization of the main channel characteristics, MFCC is also affected by pitch frequency.

For the speaker recognition system, the extracted feature parameters can not be changed by the speaker's pronunciation in different time, so the system should be time robust. Because of the difference of environment, mood and physical condition, the voice is different in different time. One of the most important random varieties is the fundamental frequency[11]. The pitch frequency lead to the increase of inconsistent of the same speaker MFCC in training and testing, so it can affect the performance of the recognition system.

In order to eliminate the influence of fundamental frequency, a smooth amplitude spectrum based on SMFCC (smoothing MFCC) is proposed. Unlike MFCC, SMFCC does not directly extracted MFCC from the short-time amplitude spectrum, but first smoothing voiced signal amplitude spectrum, then use the Mel filter bank, take the logarithm, and DCT transform. Smoothing is aimed at the voiced, but not voiceless consonant. Comparative experiment of SMFCC and MFCC in YOHO database showed that, SMFCC in the speaker recognition system was better than MFCC, and the performance is especially obvious in the female speaker data set, and SMFCC also has better robustness.

[2] Effects of Pitch Frequency on Mfcc

MFCC is extracted from the short-time spectrum of speech signal. The voiced speech signal can be regarded as the excitation source signal through the channel modulation[12], and then through the radiation of lips (Fig. 1). For short time analysis, it is necessary to take a window function $w[n]$ on $s[n]$ at length of 20 ~ 30 ms, then get short frame $x[n]$.

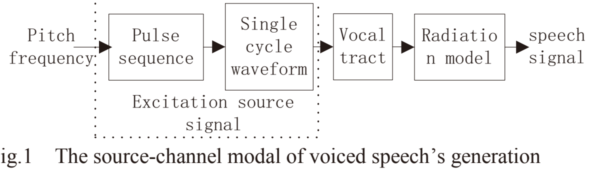


Fig.1 The source-channel modal of voiced speech's generation

The Fourier transform of $g[n]$, $h[n]$, $r[n]$, $w[n]$

Yi Zhang is with the School of Advance Manufacture Engineering, Chongqing University of Posts and Telecommunications. His research interests include signal processing, speech recognition and HCI. (e-mail: zhangyi@cqupt.edu.cn).

Yanyi Xie is with the School of Automation, Chongqing University of Posts and Telecommunications.His research interests include speech recognition and voiceprint recognition. (corresponding author: 15320504050; e-mail: 811719530@qq.com).

Kejia Wang is with the School of Automation, Chongqing University of Posts and Telecommunications. Her research interests include speech recognition.(e-mail: 937469214@qq.com).

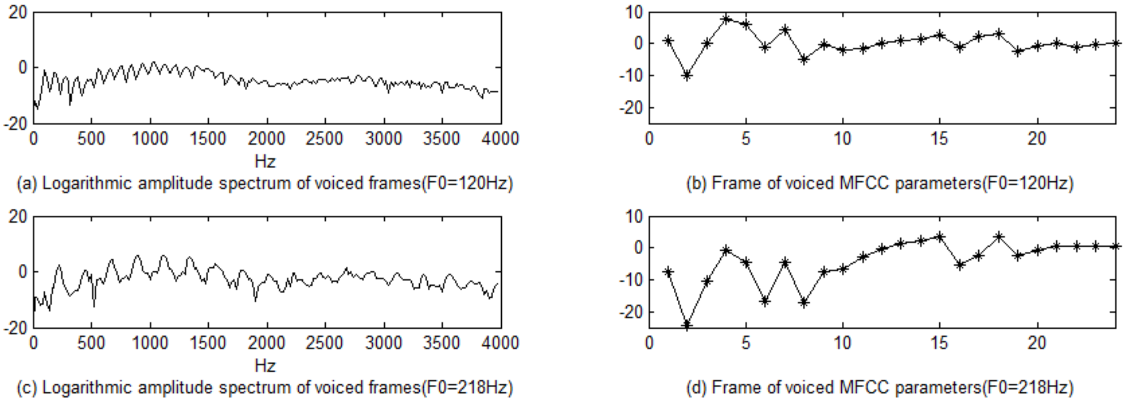
Fund Project: Projects of Chongqing Science and Technology Committee(cstc2015jcyjBX0066)

978-1-4673-8414-8/16/$31.00 ©2016 IEEE

respectively is set to be $G(w)$, $H(w)$, $R(w)$,and $W(w)$. So the spectrum of the x[n] is

$$X(\mathrm{w}) = F_0 W(w)*[G(w)H(w)R(w)\sum_{k=-\infty}^{\infty}\delta(w-kw_0) \quad (1)$$
$$= F_0\sum_{k=-\infty}^{\infty}H_{SE}(kw_0)W(w-kw_0)$$

$H_{SE}(w)=G(w)H(w)R(w)$ shows the envelope shape of the voiced frames signal amplitude spectrum, usually called spectral envelope (SE). $H_{SE}(w)$ is determined by the shape of the vocal tract pulse and the shape of the vocal tract. In addition to the semantic information, it also describes the physiological characteristics of the speaker's vocal organs, pronunciation habits, and so on. It also has a strong distinction between different speakers[13]. Voiced speech signal amplitude spectrum of $X(w)$ can be seen as the sampling of the harmonics of the fundamental frequency based on the spectral envelope, which contains information of spectrum envelope $H_{SE}(w)$ and the pitch frequency w_0. In order to eliminate the harmonic effect, the Mel filter smooths the signal amplitude spectrum in the extraction of MFCC, so that it mainly reflects the information of the spectral envelope $H_{SE}(w)$. But at frequency below 1k Hz, triangular Mel filter nonzero area is in the range of about 200 Hz, and the pitch frequency range is at around 60 to 400 Hz. When the fundamental frequency is large, Mel filter cannot eliminate harmonics very well, thus cause error of MFCC described spectral envelope $H_{SE}(w)$. Figure 2 shows the log spectrum and Mel filter log energy of voiced frames(pitch frequency respectively 120 Hz and 218 Hz) of a male and a female speaker. In Fig. 2, due to the low pitch in men, Mel filter can do well on the amplitude spectrum smoothing to eliminate the influence of pitch frequency. And for high frequency speaker (as with most women), low-frequency output of the Mel filter has obvious harmonic structure, so that the MFCC is affected by pitch frequency. To a speaker in different time, the pitch frequency may be differed with different environment, emotional and physical condition. Hypothesis two frames with the same spectral envelope of the voice, pitch frequency respectively are F_0, so the kth harmonic difference of two frames speech spectral magnitude is kF_0. The peak position of the voiced sound amplitude spectrum corresponds to the position of each harmonic, so with the increase of the frequency, the difference of the two frame speech amplitude spectrum peak position becomes larger.



(a) Logarithmic amplitude spectrum of voiced frames(F0=120Hz)

(b) Frame of voiced MFCC parameters(F0=120Hz)

(c) Logarithmic amplitude spectrum of voiced frames(F0=218Hz)

(d) Frame of voiced MFCC parameters(F0=218Hz)

Fig.2. Logarithmic amplitude spectrum and the output of Mel filterbank of a voiced speech

For each filter in the Mel filters, in the range of its bandwidth, the difference of the two frame speech amplitude spectrum peak position makes the Mel filter outputs of the two speeches different. Fig. 3 shows the harmonic position of the two speech frames with the pitch frequency respectively being 201Hz and 220 Hz, and the corresponding Mel filters. On the tenth filter, the output of speech with higher pitch frequency is much smaller than the output of speech with lower pitch frequency.

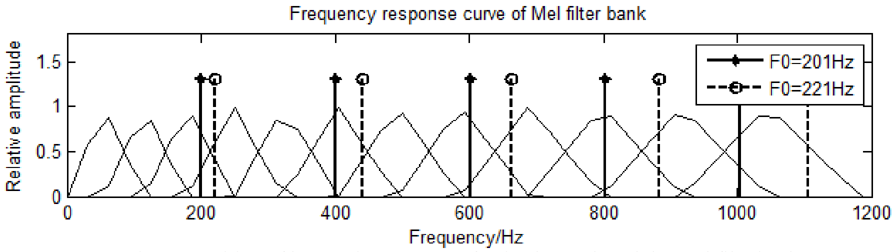


Fig.3 Position of harmonic spectrum of voiced speech and the Mel filterbank

According to the above examples, for the voice with high pitch frequency, the output of the Mel filter bank is not the same even if the difference of pitch frequency is small, so MFCC is also different. Fundamental frequency in different time will randomly change, which will cause different MFCC in training and testing, thus bring adverse effects to the

III. SMFCC

In order to eliminate the adverse effects of the vocal tract characteristics on MFCC caused by the fundamental frequency, a method based on spectral amplitude smoothing(SMFCC) was proposed in this paper. Before Mel

performance of the recognition system. even if the difference of pitch frequency is small, so MFCC is also different. filtering on the amplitude spectrum, the amplitude spectrum should be smoothed to eliminate harmonic. The impact of fundamental frequency of MFCC usually is neglected in the application and research of speech signal based on feature MFCC.

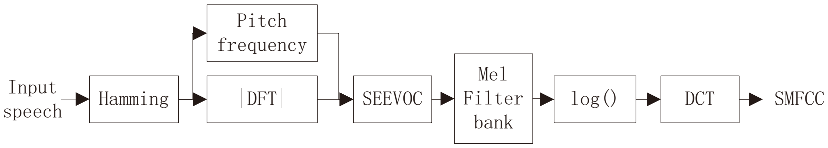


Fig.4. The process of extracting SMFCC

The extraction of SMFCC is shown in Fig. 4, unlike MFCC, there is no Mel filter bank directly acting on voiced signal amplitude spectrum, but adding a module of the smoothed amplitude spectrum before the filter. Through the interpolation of each harmonic peak amplitude spectrum of voiced signal points, the new amplitude spectrum of the frequency points between different harmonic points is obtained. In order to get the approximate spectral envelope $A(w)$ which has nothing to do with pitch frequency, the output of the $A(w)$ through Mel filter bank is obtained after the log transform and DCT transform to get SMFCC.

The interpolation of amplitude spectrum is the improvement of spectral envelope estimation algorithm proposed by Paul (envelope estimation vocoder SEEVOC) [14]. For a frame of voiced sounds speech signal, firstly the basic audio rate w_0 should be estimated. Then based on the frame of the short range of speech spectrum $X[w]$, next step is to find the peak A_1 and its location w_{11} in the interval $[w_0/2, 3w_0/2]$, and then find the peak A_2 and its location w_{12} in the interval $[w_0/2+w_{11}, 3w_0/2+w_{11}]$, cycle to find the peak $\{[A_k,\ w_{1k}]\}, k=1,2,...$, until at the voice bandwidth of the boundary. The second step is to find the bottom B_1 and its location w_{21} in the interval $[w_0/2, 3w_0/2]$, and then find the peak B_2 and its location w_{22} in the interval $[w_0/2+w_{21}, 3w_0/2+w_{21}]$, cycle to find the peak $\{[B_k,\ w_{2k}]\}, k=1,2,...$, until at the voice bandwidth of the boundary. The third step is to get the middle value of A_k and B_k named C_k, and the corresponding frequency of w_k .The spectral amplitude value of the frequency point w is obtained by linear interpolation between the amplitude value of the adjacent two peaks. The interpolation formula is

$$C(\mathrm{w}) = C_{k-1} + \frac{C_k - C_{k-1}}{w_k - w_{k-1}}(w - w_{k-1}) \qquad (2)$$

Fig. 5 shows the amplitude spectrum of a frame of voiced signal and the amplitude spectrum envelope smoothed by SEEVOC algorithm, which also shows that pitch frequency is basically eliminated in the smoothed spectral envelope. For comparison, Fig. 6 shows two frames of speeches with the same envelope spectrum, but pitch frequency respectively are 201 Hz and 220 Hz, also shows the magnitude of spectrum, Mel filter bank outputs and the corresponding characteristic parameters in pre and post smoothing. The F0 relative change is about 10%, the degree of SMFCC change is far less than MFCC in Fig.6, which shows that the use of SMFCC as the feature of the speaker recognition system has better time robustness.

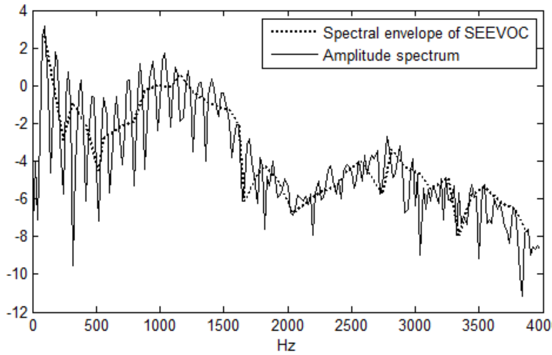


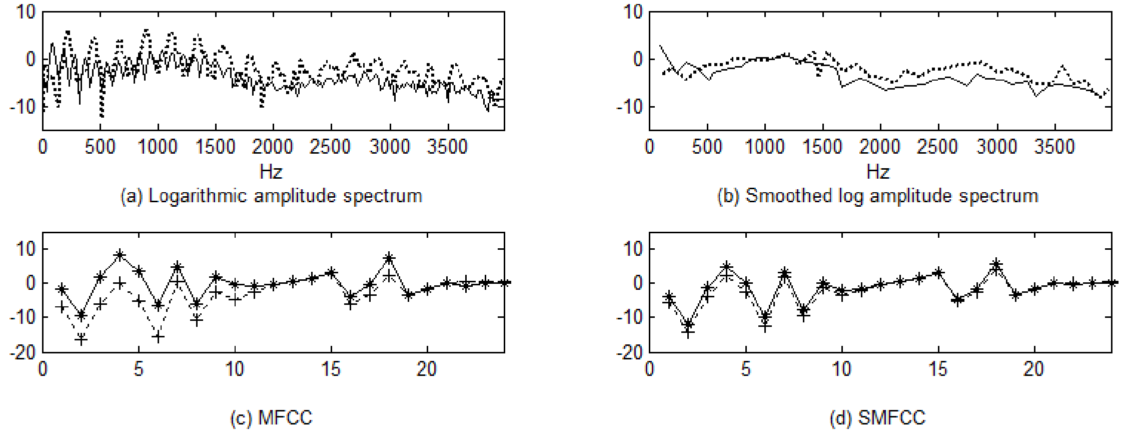
Fig.5. Log amplitude spectrum of a frame of voiced speech and the spectral envelope smoothed by SEEVOC algorithm

IV. Experimental Results and Analysis

In order to compare the performance of MFCC and SMFCC, experiments on the YOHO speaker recognition database [15] was carried out. YOHO speaker database is a large (a total of 138 people, including 106 male, 32 female) multi-session voice database, including recorded speeches in different ages, occupations and educational backgrounds at different times (within three months). Each speaker's training speech included 96 recorded speeches (each time section 24 records) in 4 different time periods. Test speech consisted of 40 recorded voices in 10 different periods. The content of each voice pronunciation was the English string with the length of 6, pronunciation length was about 3s.

A. Comparison of recognition performance of SMFCC and MFCC

YOHO database was divided into female and male speaker data sets by gender, and there were 4240 sets of male test data, and the number of female data sets was 1280. The speaker recognition performance of MFCC and SMFCC was also compared. When the feature parameters were extracted, the Hamming window of 25 ms was used for speech frame. Extraction process of MFCC and SMFCC parameters were as shown above, SMFCC and MFCC contained 16 dimensional static parameters and 16 dimensional first-order dynamic parameters as the feature vector. In the training stage, 96 training speech at 4 different time periods were used to establish the Gauss mixture model of 64 degree for each speaker. and the covariance matrix of the model was diagonal matrix. In the test, the likelihood ratio score of each extracted test speech feature parameters and each speaker model were calculated to select the maximum score model corresponding to the speaker, which is shown in TAB. 1 as one of the test results.



(a) Logarithmic amplitude spectrum

(b) Smoothed log amplitude spectrum

(c) MFCC

(d) SMFCC

Fig.6. Comparison of two voiced speech before and after smoothed

TAB.1 THE RECOGNITION RATE OF MFCC,SMFCC IN MALE,FEMALE DATABASE TRAINING WITH 4 SESSION/ 1 SESSION

data set	Error rate(4 session/ 1 session)	
	MFCC	SMFCC
Male	2.06%/7.73%	1.85%/7.25%
Female	3.23%/12.67%	1.54%/6.32%

The performance of the recognition system based on SMFCC was better than the system based on MFCC by TAB. 1, but it was different for men and women. The reason was that the male speaker pitch frequency was low for the majority. Mel filtering is a good method to eliminate the impact of harmonics, thus pitch frequency has little effect on MFCC, so the adoption of SMFCC can only be a small improvement. While the female speaker's fundamental frequency is high, the Mel filter output of low frequency periods is affected by the harmonic, thus the MFCC cannot accurately characterize the vocal tract characteristics. Through the smoothing of SEEVOC algorithm on the amplitude spectrum, SMFCC effectively reduced the influence of pitch frequency on Mel filter output, which made the system performance significantly improved, the error recognition rate fell from 3.23% to 1.54%.

B. Time robustness comparison between SMFCC and MFCC

People differ in pronunciation at different times. This kind of pronunciation change may cause inconsistent of the speaker recognition parameter of training and test, so as to affect the performance of the recognition system. The time robustness of the recognition system means, the performance of the system identification does not decline over time. For a practical speaker recognition system, it is very important to have good time robustness.

A method to improve the time robustness of the system is to collect enough different time speeches to train the speaker model. In the 3.1 section of the experiment, the training data for each speaker included 96 pronunciation training speaker models in 4 different time periods. The actual speaker recognition system in the training phase can only get the pronunciation in the training of speech. The recognition process is usually carried out at other times. For a practical speaker recognition system, it is very important to have good time robustness. At this point, to reduce the degree of the characteristic parameters variation of the same speaker at different times is important.

In order to test the system with the characteristics of MFCC and the time robustness of the system based on SMFCC, only 24 speeches of one time periods was used in training. Test speeches were still the voice of all-time period, the other aspects of the system are the same as the 3.1 section. The results of the experiment were shown in TAB. 1. As can be seen from TAB. 1, the recognition rates of the 2 parameters of male and female both declined. In particular, due to the influence of pitch frequency on MFCC of female speaker, women MFCC's error rate rose from 2.82% to 16.14%. The fundamental frequency of different times of pronunciation was instable, which showed that the SMFCC can effectively reduce the influence of fundamental frequency, thus its error rate was far less than the system based on MFCC. At the same time, for the male speakers, effects on MFCC were very small because of its low audio frequency, so SMFCC and MFCC performances were nearly similar.

V. CONCLUTION

The influence of pitch frequency on MFCC and the impact of this phenomenon on the performance of speaker identification system were analyzed in this paper. Experiments showed that for women with higher pitch frequency, the difference of MFCC caused by pitch frequency changes had a negative impact on the performance of speaker recognition. This effect was particularly evident in the case of insufficient training data, which is detrimental to speaker recognition. The SMFCC proposed in this paper can effectively reduce this effect, and it is better than MFCC in the female speaker recognition performance.

REFERENCES

[3] S. CHATTERJEE, W. B. KLEIJN, "Auditory Model-Based Design and Optimization of Feature Vectors for Automatic Speech Recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, pp. 1813-1825, Jun. 2011.

[4] C. KONIARIS, S. CHATTERJEE, W. B. KLEIJN, "Selecting static and dynamic features using an advanced auditory model for speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 4, pp. 115-126, Jun. 2011.

[5] J. A. Morales-Cordovilla, A. M. Peinado and V. Sanchez, "Feature Extraction Based on Pitch-Synchronous Averaging for Robust Speech Recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, pp. 640-651, Jun. 2011.

[6] G. D. O'Clock, W. L. Yong and L. Jongwon, "A Simulation Tool to Study High-Frequency Chest Compression Energy Transfer Mechanisms and Waveforms for Pulmonary Disease Applications," *IEEE Trans. Biomedical Engineering*, vol. 57, pp. 1539-1546, Feb. 2010.

[7] S. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Audio, Speech and Language Processing,* vol. 28, pp. 357-366, Nov. 1980.

[8] S. Mortia, M. Unoki and L. xugang, "Robust voice activity detection based on concept of modulation transfer function in noisy reverberant environments," in *international Symposium on Chinese Spoken Language Processing*, Singapore, 2014, pp. 108–112.

[9] L. Wei and Y. Lili and X. B. ling, "Based on the modified MFCC parameters of Chinese whispered speech recognition," *Nanjing University Journal (NATURAL SCIENCE)*, vol. 42, pp. 54-62, Jan. 2006.

[10] K. D. Suk, L. S. Young and M. K. Rhee, "Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 55-69, Nov. 1999.

[11] Z. Guofu, C. Jinbao and L. Chao, "Pattern recognition approach to identify loose particle material based on modified MFCC and HMMs," *Neurocomputing*, vol. 155, pp. 135-145, Mar. 2015.

[12] Y. H. Goh, P. Raveendran and S. S. Jamuar, "Robust speech recognition using harmonic features," *Iet Signal Processing*, vol. 8, pp. 167-175, Mar. 2014.

[13] K. Yanagisawa, K. Tanaka and I. Yamaura, "Detection of the fundamental frequency in noisy environment for speech enhancement of a hearing aid," *in Conf. Rec. 1999 IEEE Int. Control Applications*, pp. 1330–1335.

[14] S. M. Caballero, F. R. Trujillo, "Evolutionary approach for integration of multiple pronunciation patterns for enhancement of dysarthric speech recognition," *Expert Systems with Applications Iet Signal Processing*, vol. 41, pp. 841-852, November 2014.

[15] L. H. Shih, B. Chen and Y. Ming, "Exploring the Use of Speech Features and Their Corresponding Distribution Characteristics for Robust Speech Recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, pp. 84-94, July 2009.

[16] D. Paul, "The spectral envelope estimation vocoder," *IEEE Trans. on Acoustics,Speech and Signal Processing*, vol. 29, pp. 786-794, Jan. 1981.

[17] J. P. Campbell, "Testing with the YOHO CD-ROM voice verification corpus, " in *Conf. Rec. 1999 IEEE Int. Acoustics,Speech and Signal Processing*, pp.341-344.