Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/070445

Robust FHPD Features from Speech Harmonic Analysis for Speaker Identification

Shuiping Wang[1,2,3,*], *Zhenmin Tang*[1] ,*Ye Jiang*[1] *and Ying Chen*[1]

[1]School of Computer Science & Technology, Nanjing University of Science & Technology, Nanjing 210094, P. R. China
[2]Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing 210044, P. R. China
[3]School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, P. R. China

Received: 25 Nov. 2012, Revised: 3 Jan. 2013, Accepted: 29 Jan. 2013
Published online: 1 Jul. 2013

Abstract: Speaker identification accuracy decreases significantly in the presence of additive noise. In this paper, we propose a robust speech feature extraction method, which is based on the harmonic structure of voiced segments. The robust features are composed of fundamental and harmonic peak data from short-time spectrum. These features are evaluated by thirty speaker data from TIMIT database and additive noise signals from NOISEX-92 database with clean training and noisy testing samples. Results reflect that under low SNR (signal-to-noise ratio) environments new features achieve better performance than conventional MFCC (Mel-Frequency Cepstral Coefficients) parameters.

Keywords: Harmonic Analysis, Speaker Identification, Spectrogram, Gaussian Mixture Model

1. Introduction

Automatic speaker recognition (ASR) refers to recognizing persons from their voice. It can be classified into speaker identification and speaker verification[1]. Automatic speaker identification (ASI) involves determining if a speaker is a specific person or is among a group of persons. It is an N-way classification process. Automatic speaker verification (ASV) decides a speaker is whom he/she claims to be.

Speaker recognition systems have been studied actively for several decades. It can also be used for some mobile applications[2]. The performance of them under clinical and controlled conditions is good but degrades significantly in real-world noisy environments in case of mismatched channel or additive noise. Recently, several linear and nonlinear compensation methods, such as front-end enhancement, feature domain processing and model domain compensation, have been proposed to reduce the effect of channel mismatch and additive noise. Front-end enhancement methods such as spectral subtraction[3], Wiener filtering[4] and Kalman filtering[5] were proposed to improve the signal-to-noise ratio (SNR), meanwhile, they suppress background noises[6]. Front-end enhancement methods are all based on building a statistical estimation for noise and removing it from the noisy speech. However, one drawback of them is that imperfect noise estimates may result in removing both the noise and speaker-dependent information of the original speech[7]. In the feature processing section, researchers are dedicated to looking for robust acoustic features or reprocessing the features which are extracted from noisy speech. Some feature normalization methods, such as cepstral mean normalization[8], Relative Spectra (RASTA) filtering[9] and feature warping[10], are often stacked with each other to reduce the impact of environmental noise. Examples of model domain compensation methods, include parallel model combination (PMC)[11] and model-domain spectral subtraction[12], are based on the assuming that the statistical model of the noise is available.

Short-time and low-level acoustic information, such as cepstrum characters, is commonly used in the current speaker recognition systems. Linear prediction cepstral coefficients (LPCC) and Mel-frequency cepstral coefficients (MFCC) are the mostly used features. The performance of these features under controlled conditions is good but degrades significantly in real-world noisy

* Corresponding author e-mail: shuipingw@126.com

environments. Although some researchers engaged in adding △MFCC and △△MFCC into the basic MFCC to improve the performance of systems in noisy conditions, the increase in the number of dimensions of feature vector will improve the amount of computation.

This letter describes a novel feature extraction technique based on short-time spectrum analysis of clear and noisy speech. The new feature parameters, called fundamental and harmonic peak data (FHPD), are generated by applying the excitation source harmonic characteristics. To evaluate the results, TIMIT databases are used. Experimental results show FHPD parameters that can achieve higher recognition rate in speaker identification application than MFCC parameters under lower SNR noisy conditions.

This paper is organized as follow: Section 2 analyzes voiced speech spectrogram and FHPD feature, Section 3 describes the FHPD parameters extraction process and their performance evaluation, Section 4 shows the experiments and results, and Section 5 reflects the conclusions.

2. FHPD Feature Analysis

The innate difference of speakers' vocal organs is mainly expressed in the frequency of the structure of voices. Excitation source and channel characteristics are included in the short-time spectrum of speech. They can reflect the speakers' physiological differences. In conclusion, the short-time spectrum of speech signals may show the speakers' personality traits and can be used as the characteristic parameters of the speaker recognition applications. In accordance with the human pronunciation manner, speech can be divided into unvoiced, voiced and plosive. Among them, voiced speech can be considered as a quasi-periodic signal, and it contains most of speaker-dependent information.

2.1. Short-term Spectrum Analysis

Firstly, we add a hamming window to every frame of speech signal, and then calculate the Fourier amplitude spectrum with normalization step. At last we use 20 times of logarithm normalized spectrum to plot the speech spectrogram. The flow chart of the short-time spectrum calculation is shown as Figure 1.

The short-time spectrum results without noise and with a 10dB SNR noise are shown in Figure 2. Both the waveform and the spectrum are also shown in this figure. The speech is selected from TIMIT database. The voice sampling rate is 16kHz, and the window is with 256 samples.

Figure 3 is plotted by using the logarithm amplitude of three typical frames, which are (a) voiced frame, (b) unvoiced frame, and (c) noisy frame. As shown in Figure 3, there is an obvious difference among them. The spectrum of voiced segment has a harmonic structure apparently, and the energy is mainly concentrated at fundamental and harmonic waveform peaks. However, the energy of unvoiced or noisy spectrum is not regular, and both of them do not have the harmonic structure.

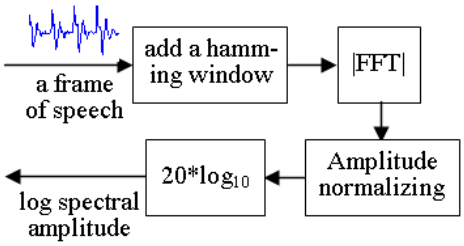


Figure 1: short-time spectrum analysis flowchart

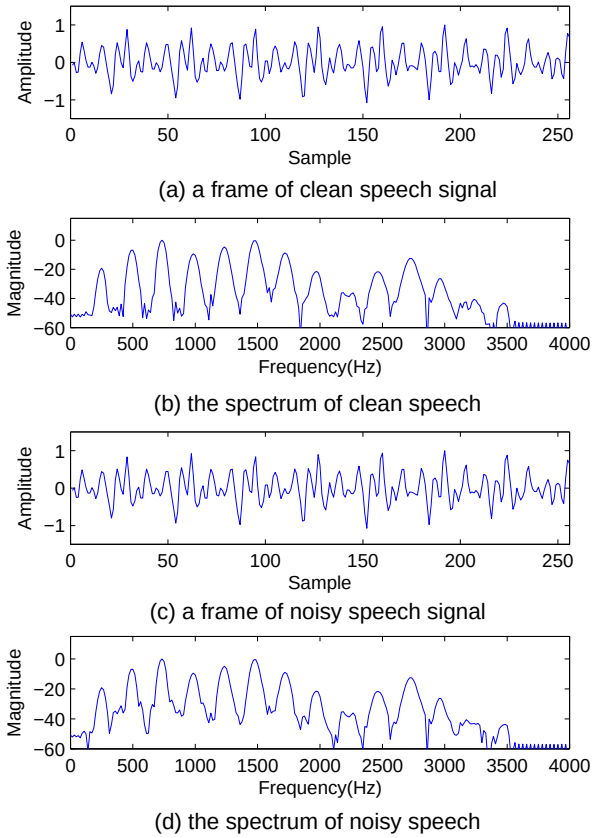


Figure 2: the waveform and the spectrum of one frame speech

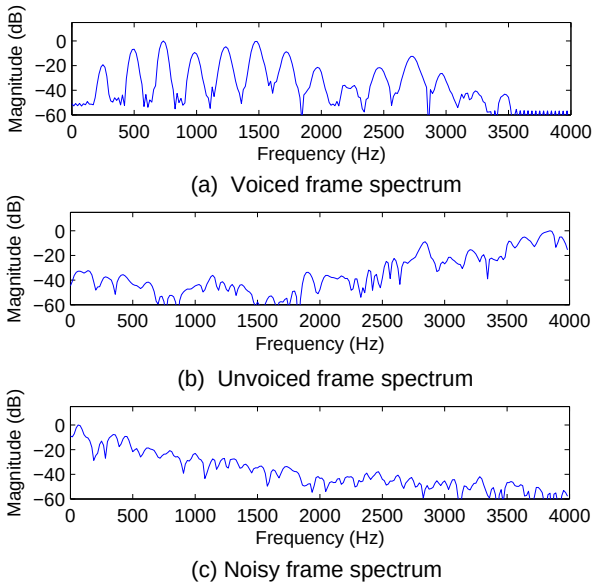


Figure 3: Log-spectrums of voiced, unvoiced and noisy frames

2.2. FHPD under Different SNR Values

Figure 4 shows the spectrums of the same voiced frame in different SNR values. Among them, figure (a) is the spectrum of the speech signal without any noise, figure (b) is the spectrum of the speech signal with Gauss White noise with SNR=15, and figure (c), (d) and (e) are the spectrum diagrams with SNR equal to 10, 5 and 0 respectively. From the diagrams, it can be concluded that the peak locations of lower harmonic waves are not been affected enormously by noise. That is to say, fundamental and harmonic peak data (FHPD) are not sensitive to noise. The experiments in other noisy conditions, such as babble noise, factory noise, car noise and airport noise, lead to the same conclusion.

We also record the fundamental and harmonic peak data to confirm that they are not sensitive to the noise in different SNR values. Table.1 shows FHPD parameters of a voiced frame selected from a female speaker voice under high level SNR values. It is obvious that the peak data of the same number harmonic waveform vary very little. To further confirm the conclusion that FHPD features are robust, the experiments in lower level SNR values are done, and the results are shown in Table.2.

Table.1 compares the FHPD data between clean voice and noisy speech with high SNR level noises, and table.2 compares the FHPD data between clean speech and noisy speech with lower SNR level noise. We can easily conclude that the FHPD features are robust not only to high level SNR noise but also to lower level SNR noise conditions.

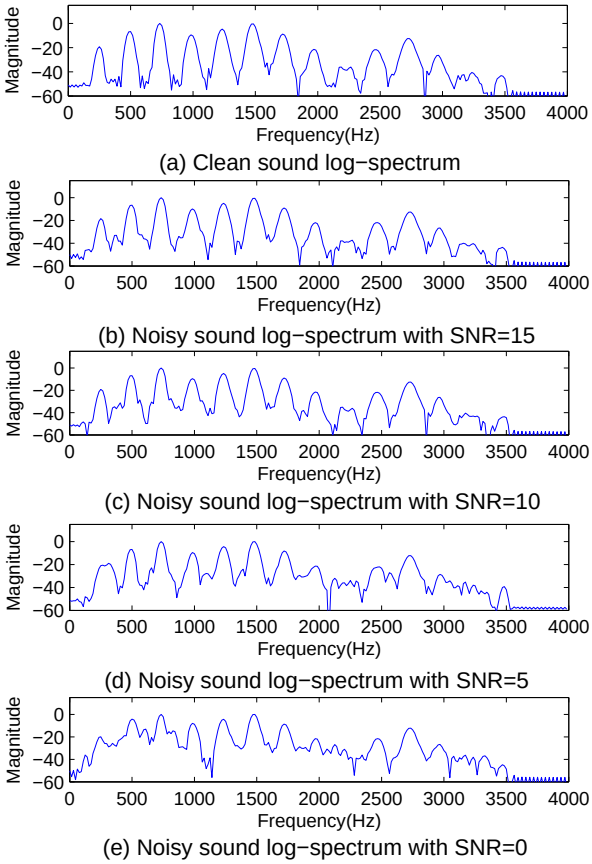


Figure 4: Log-spectrums of a voiced frame of a female speaker with different SNR values

The data in Table.1 and Table.2 can only improve the robustness of the FHPD characters. In order to confirm the FHPD parameters are speaker-dependent, the data of three different frames as shown in Table.3 are from one male speaker, and from Table.3 it can be concluded that FHPD features are speaker dependent. That means different speaker has different FHPD structure.

3. FHPD Feature Detection and performance Evaluation

In this section we discuss how to get accurate peaks of the fundamental and harmonic waveforms. The theoretical analysis of the effectiveness of the FHPD feature is present here too.

Table 1 FHPD parameters of a voiced frame in high level SNR values

harmonic waveform number or f_0	SNR=∞	SNR=30	SNR=20	SNR=15
f_0	-19.2600	-19.2633	-19.3064	-19.3145
1	-6.7836	-6.7815	-6.7859	-6.7989
2	0	0	0	0
3	-9.4267	-9.4345	-9.4386	-9.5622
4	-4.7002	-4.7322	-4.7314	-4.7369
5	-0.3625	-0.3632	-0.3573	-0.3763
6	-8.7705	-8.7743	-8.7724	-8.7764
7	-21.1233	-21.1258	-21.1276	-21.1253
8	-36.0425	-36.8622	-36.4123	-36.5135
9	-21.7099	-21.7642	-21.7242	-21.4257
10	-12.5823	-12.5932	-12.5894	-12.5845
11	-26.3985	-26.3934	-26.3810	-26.3904

Table 2 FHPD parameters of a voiced frame in low level SNR values

harmonic waveform number or f_0	SNR=∞	SNR=10	SNR=5	SNR=0
f_0	-19.2600	-19.1856	-20.6373	-19.9146
1	-6.7836	-6.8816	-6.9086	-4.1849
2	0	0	0	0
3	-9.4267	-9.4945	-9.5486	-7.9622
4	-4.7002	-4.9613	-4.3307	-4.3469
5	-0.3625	-0.5163	-0.0883	-0.0447
6	-8.7705	-9.0081	-8.2451	-8.6446
7	-21.1233	-21.5483	-21.4251	-21.6422
8	-36.0425	-35.2249	-32.0527	-26.8421
9	-21.7099	-21.8313	-22.0630	-21.5710
10	-12.5823	-12.6795	-12.2389	-12.3382
11	-26.3985	-26.3185	-28.6085	-26.7319

3.1. Calculation of Fundamental Frequency and harmonic peak data

As we know, the frequency location of the first waveform peak in spectrum is the fundamental frequency. Fundamental frequency, often abbreviated f_0 , is defined as the frequency at which the vocal cords vibrate during a voiced sound. Due to the non-stationary of speech signal, it seems clear that f_0 estimation is a difficult matter. There are many methods for evaluating f_0 estimation. Among these methods, autocorrelation function and average magnitude function methods are typical[13]. These two methods are simple and robust when the signal is noiseless. However, the accuracy of these methods is significantly decreased when the speech signal is degraded.

In this letter, we choose Harmonic Power Spectrum (HPS) algorithm[14] to detect f_0. $S_n(e^{j\omega})$ is set to be the short-time spectrum of signal $s(n)$, and then harmonic power spectrum is defined as follow:

$$P_n(e^{j\omega}) = \prod_{r=1}^{D} S_n(e^{jr\omega}), \tag{1}$$

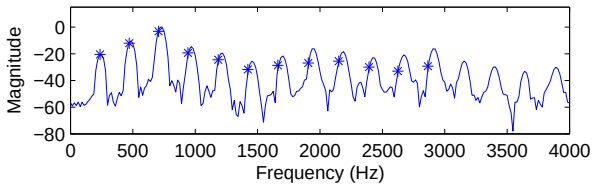
where D is the number of the harmonics that are used in the algorithm.

HPS algorithm works in frequency domain and thus, different from the algorithm based on autocorrelation, it is quite robust to additive and multiplicative noise.

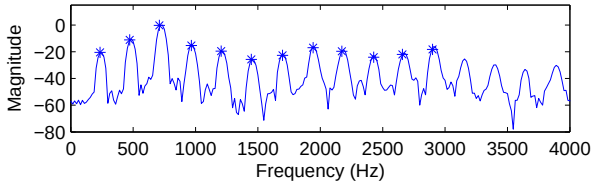
Successfully extracting the fundamental frequency f_0 , literature [15] put $f_0, 2f_0, 3f_0, \cdots, Nf_0$ into $S(f)$, and the results $S(f_0), S(2f_0), S(3f_0), \cdots, S(Nf_0)$ are FHPD parameters, where $S(f)$ is the short-time log spectrum of signal $s(n)$. As shown in figure 5(a), the peak location of fundamental waveform is almost accurate, but in other harmonic waveforms, the results are not exactly at the peak. To fix these errors, we propose an improved method. We use the maximum of $S(f)$ in the interval $[nf_0 - m, nf_0 + m]$ to be peak spectrum at frequency $nf_0 (n = 2 \cdots N)$, and the results are recorded as $\tilde{S}(nf_0)$, as defined in formula (2). Further more, the frequency locations of the harmonic waveforms peaks are used to adjust f_0 in turn. We determine the m value by the initial f_0 , because fundamental frequency of male and female are always in different intervals.

$$\tilde{S}(nf_0) = \max_{k=nf_0-m}^{nf_0+m} \{S(k)\}, (n = 2 \cdots N) \tag{2}$$

The result derived from improved method is shown as Figure 5(b).



(a) FHPD detection results using unimproved algorithm



(b) FHPD detection results using improved algorithm

Figure 5: Fundamental and Harmonic Peak Data results using unimproved and improved algorithm

Table 3 FHPD parameters of three different voiced frames from a male speaker

harmonic	frame A		frame B		frame C	
waveform	freq	magn	freq	magn	freq	magn
number or f_0	-uency	-itude	-uency	-itude	-uency	-itude
f_0	156	-29.5576	156	-28.3654	141	-26.9616
1	297	-7.3176	297	-7.2302	281	-7.0717
2	438	0.0000	438	0.0000	422	0.0000
3	578	-8.4685	563	-6.9983	563	-5.9306
4	719	-8.3196	703	-7.8032	703	-8.9742
5	859	-10.0237	859	-7.7098	844	-7.3376
6	984	-18.5008	984	-14.6200	969	-10.5227
7	1125	-26.9582	1109	-23.2654	1094	-20.6580
8	1266	-42.3121	1266	-35.4052	1266	-32.7561
9	1391	-38.3435	1375	-35.4980	1375	-34.1615
10	1547	-43.8533	1516	-43.0902	1531	-42.2465
11	1688	-36.8468	1688	-36.4351	1656	-36.8322

3.2. Performance Evaluation of FHPD Feature

Feature extraction method is of great value to speaker recognition systems. In this part, we will use F Ratio from Fisher identification theory to evaluate the performance of FHPD feature in each dimension for speaker recognition.

The F Ratio is defined as follow[16]:

$$F = \frac{mean\ variance\ of\ different\ speakers}{variance\ mean\ value\ of\ the\ same\ speaker} = \frac{< [\mu_i - \bar{\mu}]^2 >_i}{< [x_a^{(i)} - \mu_i]^2 >_{a,i}} \quad (3)$$

where

$x_a^{(i)} \rightarrow$ the feature parameters at time a of speaker i ;

$< \bullet >_i \rightarrow$ average computing for i ;

$< \bullet >_a \rightarrow$ the average of different speech segment of one speaker;

$\mu_i =< x_a^{(i)} > \rightarrow$ the estimation mean of all characters of speaker i;

$\bar{\mu} =< \mu_i >_i \rightarrow$ the mean of μ_i for all speakers.

Obviously, the features of larger F Ratio value are suit to present the personal character of the speaker. Experiments show that the middle and the lower dimension FHPD values are of more contributions to speaker recognition process. We choose the lower 12 dimensions of FHPD features to build a speaker model and complete the training and recognition processes.

We choose the peak amplitudes of fundamental waveform and the 11 low harmonic waveforms to form a 12-dimension feature vector.

Experiments show that different speaker has different harmonic structure. Table.4 shows FHPD features of 3 speakers (2 females and 1 male).

4. Experiments and Results

Speaker models such as Vector Quantization (VQ)[17], Gaussian Mixture Model (GMM)[18], Hidden Markov Model (HMM)[19], Artificial Neural Network (ANN)[20], and Support Vector Machine (SVM)[21] are commonly used in speaker recognition systems. As a generic probabilistic model, Gauss Mixture Model can simulate any continuous probability of the multi-dimensional vector, and thus it is suitable for text-independent speaker recognition system.

The proposed FHPD features as well as the classical MFCC features are used in a GMM-based speaker identification system. Both input speech features are 12-dimensional vectors, and the effectiveness of them is evaluated on the TIMIT database. We choose 30 speakers (16 males and 14 females) from TIMIT. Each speaker contributing ten utterances and each utterance has an average duration of about 3s. Seven utterances of each speaker are spliced together to be the training samples, and the other three are also spliced together as test samples. As we all know, several kinds of noise signals are frequently encountered in the real life. Therefore, we choose speech babble, volvo car interior and factory noises from the NOISEX-92 database[22] .

The babble noises are added to the testing speech utterances to obtain the signal-to-noise ratio (SNR) of 30, 20, 10, 5 and 0 dB. The results are shown in Table.5. Table 5 and Table 6 show the performance evaluated using MFCC and FHPD features with different order of GMM model.

Table.7 and Table.8 show the true recognition rates on volov car interior and factory noises at different levels.

Experiments show that MFCC features perform better than FHPD features with clean test speech or the speech with high level SNR noises. When the SNR level drops

Table 4 FHPD parameters of three speakers (2 females and 1 male)

harmonic waveform number or f_0	Speaker 1(female) freq -uency	Speaker 1(female) magn -itude	Speaker 2(female) freq -uency	Speaker 2(female) magn -itude	Speaker 3(male) freq -uency	Speaker 3(male) magn -itude
f_0	234	-16.5271	250	-19.3167	141	-29.4138
1	453	-18.2687	469	-8.87	297	-6.8593
2	672	-12.4713	703	-7.8785	438	0
3	891	-5.0369	938	-6.3114	578	-11.9867
4	1125	-4.8645	1156	0	734	-9.2196
5	1344	0	1391	-4.2599	875	-13.1855
6	1563	-4.9505	1609	-2.9724	1016	-22.9625
7	1781	-17.0043	1844	-7.5558	1156	-34.765
8	2000	-19.4903	2078	-30.6002	1297	-34.5652
9	2219	-16.7966	2328	-45.4616	1438	-37.9868
10	2375	-19.2894	2531	-24.0361	1578	-37.7772
11	2672	-22.6021	2766	-20.1938	1719	-38.2223

Table 5 True identification rates using 12-MFCC on babble noises with different order GMM

SNR	2-GMM	4-GMM	8-GMM	16-GMM	32-GMM	64-GMM
∞	50.00%	63.33%	76.67%	83.33%	90.00%	93.33%
30	50.00%	70.00%	80.00%	86.67%	86.67%	93.33%
20	23.33%	40.00%	40.00%	56.67%	50.00%	53.33%
10	3.33%	3.33%	6.67%	6.67%	6.67%	6.67%
5	3.33%	6.67%	3.33%	6.67%	6.67%	6.67%
0	0.00%	6.67%	3.33%	3.33%	3.33%	3.33%

Table 6 True identification rates using 12-FHPD on babble noises with different order GMM

SNR	2-GMM	4-GMM	8-GMM	16-GMM	32-GMM	64-GMM
∞	23.33%	36.67%	46.67%	60.00%	80.00%	83.33%
30	20.00%	33.33%	43.33%	53.33%	73.33%	76.67%
20	16.67%	30.00%	43.33%	46.67%	70.00%	70.00%
10	16.67%	26.67%	43.33%	43.33%	66.67%	63.33%
5	16.67%	20.00%	30.00%	50.00%	56.67%	60.00%
0	13.33%	16.67%	23.33%	40.00%	53.33%	53.33%

from 20 to 10, the identification rate using MFCC features drops quickly.

We also find that when the order of GMM model changes form 32 to 64, the identification results change little. Therefore we choose 32-GMM as the final speaker model. MFCC and FHPD features' performance comparison can be easily learned from Figure 6, which is plotted by the average identification rates of 32-GMM model under three kinds of noises.

Although at high SNR values, MFCC features work well, FHPD features can achieve much higher

Table 7 True identification rates under volvo car interior noises with different SNR levels

Features	SNR	8-GMM	16-GMM	32-GMM	64-GMM
MFCC	∞	80.00%	86.67%	90.00%	90.00%
	30	80.00%	86.67%	83.33%	86.67%
	20	40.00%	53.33%	53.33%	56.67%
	10	6.67%	10.00%	10.00%	10.00%
	5	6.67%	6.67%	6.67%	6.67%
	0	3.33%	3.33%	3.33%	3.33%
FHPD	∞	46.67%	60.00%	83.33%	83.33%
	30	43.33%	56.67%	76.67%	80.00%
	20	40.00%	46.67%	70.00%	70.00%
	10	40.00%	43.33%	63.33%	63.33%
	5	30.00%	40.00%	56.67%	60.00%
	0	23.33%	40.00%	53.33%	53.33%

Table 8 True identification rates under factory noises with different SNR levels

Features	SNR	8-GMM	16-GMM	32-GMM	64-GMM
MFCC	∞	80.00%	83.33%	90.00%	93.33%
	30	80.00%	86.67%	90.00%	90.00%
	20	36.67%	53.33%	53.33%	56.67%
	10	6.67%	6.67%	10.00%	10.00%
	5	3.33%	6.67%	6.67%	6.67%
	0	3.33%	3.33%	3.33%	3.33%
FHPD	∞	50.00%	60.00%	83.33%	83.33%
	30	46.67%	53.33%	76.67%	76.67%
	20	43.33%	50.00%	73.33%	76.67%
	10	40.00%	46.67%	63.33%	63.33%
	5	30.00%	46.67%	56.67%	60.00%
	0	23.33%	40.00%	53.33%	53.33%

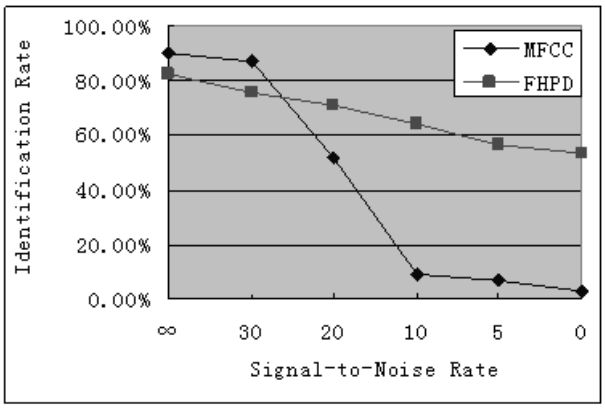


Figure 6: declining trend diagram of identification rate

identification rate than MFCC features when the SNR value is less than 30.

5. Conclusion and Future Work

In order to solve the robustness of speaker recognition systems in the feature domain, the short-time spectrum of pure and noisy voiced speech were analyzed in this paper. A new feature FHPD based on short-time spectrum was introduced for robust speaker identification system. In order to compare the performance of FHPD features, MFCC features were used as a baseline. In general, FHPD features outperformed standard MFCC features under babble noises at low SNR values. In summary, the FHPD features are robust and using them for speaker identification is a promising approach in the presence of additive noise.

Future work will be in the direction of combining FHPD parameters with other features to increase the identification rate.

Acknowledgements

This work is partially supported by the National Nature Science Foundation of China under Grant No. 61103141, and Jiangsu Provincial Natural Science Foundation of China under Grant No.CX2211_0261.

References

[1] T. Kinnunen and H. Li. An overview of text-independent speaker recognition: From features to supervectors. Speech communication. **52**(1):12-40. (2010).

[2] D. Zhang and X. Zhang. A New Service-Aware Computing Approach for Mobile Application with Uncertainty. Applied Mathematics & Information Sciences. **6**(1): 9-21. (2012).

[3] G. Lathoud, M. Magimai-Doss, B. Mesot and H. Bourlard. Unsupervised spectral subtraction for noise-robust ASR. in IEEE International Conference on Automatic Speech Recognition and Understanding. San Juan, China: IEEE. 343-348. (2005).

[4] L. M. Arslan. Modified wiener filtering. Signal processing. **86**(2): 267-272. (2006).

[5] S. Stan, T. Fingscheidt and C. Beaugeant. An evaluation of VTS and IMM for speaker verification in noise. in Proceedings of the 8th European Conference on Speech Communication and Technology. Geneva, Switzerland. 1669-1672. (2003).

[6] P. C. Loizou. Speech enhancement: theory and practice, Vol. **30**. London Taylor & Francis Publishers (2007).

[7] R. Saeidi, J. Pohjalainen, T. Kinnunen and P. Alku. Temporally weighted linear prediction features for tackling additive noise in speaker verification. Signal Processing Letters, IEEE. **17**(6): 599-602. (2010).

[8] A. A. Garcia and R. J. Mammone. Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping. in IEEE International Conference on Acoustics, Speech, and Signal Processing. Phoenix, AZ: IEEE. 325-328. (1999).

[9] H. Hermansky and N. Morgan. RASTA processing of speech. Speech and Audio Processing, IEEE Transactions on. **2**(4): 578-589. (1994).

[10] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. in Proc. Odyssey Speaker and Language Recognition Workshop. Crete, Greece. 213-218. (2001).

[11] M. J. F. Gales and S. J. Young. Robust continuous speech recognition using parallel model combination. Speech and Audio Processing, IEEE Transactions on. **4**(5): 352-359. (1996).

[12] J. A. Nolazco-Flores and L. Garcia-Perera. Enhancing acoustic models for robust speaker verification. in IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, NV: IEEE. 4837-4840. (2008).

[13] Li Jin, Jiang Cheng, Liu Fu. Improved algorithm for pitch detection[J]. Computer Engineering and Applications, 2011, **47**(3): 117-119.

[14] C. Llerena, L. lvarez and D. A. On. Pitch detection in pathological voices driven by three tailored classical pitch detection algorithms. in Proceedings of the 11th WSEAS international conference on Signal processing, computational geometry and artificial vision. Wisconsin, USA World Scientific and Engineering Academy and Society (WSEAS). 113-118. (2011).

[15] L. H. Zhang, Y. Bao and Z. Yang. Robust feature based on speech harmonic structure for speaker identification. **28**(10): 1786-1789 (2006).

[16] Zhao Li. Speech signal processing[M]. Beijing: China Machine Press: 219-220.

[17] A. E. Rosenberg and F. K. Soong. Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. Computer Speech & Language. **2**(3-4): 143-157. (1987).

[18] D. A. Reynolds, T. F. Quatieri and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. Digital signal processing. **10**(1-3): 19-41. (2000).

[19] H. Lei and N. Mirghafori. Word-conditioned HMM supervectors for speaker recognition. in Proceedings of the 8th Annual Conference of the International Speech Communication Association. Antwerp, Belgium. 746-749. (2007).
[20] K. R. Farrell, R. J. Mammone and K. T. Assaleh. Speaker recognition using neural networks and conventional classifiers. IEEE Transactions on Speech and Audio Processing. **2**(1): 194-205. (1994).
[21] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer and P. Torres-Carrasquillo. Support vector machines for speaker and language recognition. Computer Speech & Language. **20**(2): 210-229. (2006).
[22] A. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Documentation included in the NOISEX-92 CD-ROMs, 1992.

Shuiping Wang received the MS degree in System Analysis and Integration from Nanjing University of Information Science and Technology in 2000, and she is currently pursuing the Ph.D. degree in Pattern Recognition and Intelligent System from Nanjing University of Science and Technology. She is a associate professor at the Department of Computer and Software, Nanjing University of Information Science and Technology. Her research interests are in the areas of speech recognition and speaker recognition.

Zhenmin Tang received his Ph.D. degree in Pattern Recognition and Intelligent System from Nanjing University of Science and Technology. He currently is a professor at the Department of Computer Science, Nanjing University of Science and Technology, China. His research interests include signal processing, image processing and intelligent robot.

Ye Jiang is currently pursuing the Ph.D. degree in Pattern Recognition and Intelligent System from Nanjing University of Science and Technology. His research interests include speech recognition and speaker recognition.

Ying Chen received the MS degree in Applied Computer Technology from Northeast Dianli University in 2007, and she is currently pursuing the Ph.D. degree in Pattern Recognition and Intelligent System from Nanjing University of Science and Technology. Her research interests include information processing and speaker recognition.