Journal of Signal and Information Processing, **2013, 4, 106-110**
http://dx.doi.org/10.4236/jsip.2013.42014 Published Online May 2013 (http://www.scirp.org/journal/jsip)

Research on Different Feature Parameters in Speaker Recognition

Qiyue Liu, Mingqiu Yao, Han Xu, Fang Wang

Department of Communication and Information System, Hebei University of Science and Technology, Shijiazhuang, China.
Email: yaomingqiu@126.com

Received March 15th, 2013; revised April 16th, 2013; accepted April 25th, 2013

ABSTRACT

Feature parameters extraction is critical for speaker recognition research. The paper presents the function of pitch, formant and Mel frequency central coefficient (MFCC) in speaker recognition. It can increase the identification rate effectively for feature parameter sorts the speech corpus. Using Euclid Distance to compare feature parameters is very effective.

Keywords: Pitch; Formant; MFCC; Euclid Distance

1. Introduction

People can distinguish different speakers through the ear, since people can know the difference, machine can also do it in some kind of method. Speaker recognition is to make a machine to identify different people, which are to let the machine know who is talking.

The ultimate purpose of speaker recognition is to identify who is speaking, while to ignore the content of speech. In fact, it is the recognition of the characteristics of speech.

The human voice is a natural property, each person's speech organs have their own characteristics and pronunciation habits. Therefore, to identify speaker exactly, the parameters that can fully reflect the personality characteristic must be extracted from the speech signal.

These feature parameters should have these characteristics [1,2]:

- These can fully embody the large difference between different people, and can keep stability relatively when the speaker's speech changes.
- These can maintain good health and stubbornness when voice suffers from outside interference.
- These cannot imitate easily.
- These are easy to extract and compute, and have favorable independence between each dimension of the characteristic parameter.

Voice is different from fingerprint, the fingerprint is fixed, but the voice is changing, so it has not been found in some kind of parameter which could fully meet all of these features we mentioned above. The sound is connected with human emotion, health and environment, etc. And also has a relationship with the voice content. Therefore, all of the characteristic parameters we applied to have some defects now, which cannot accurately stand for the speaker's personality traits.

2. Research on Different Feature Parameters

Speaker's characteristics are generally reflected in channel feature and the glottal feature.

In the case of ensuring the recognition rate, it should be very difficult to improve recognition time through reducing computational complexity. It has been used that expending computation time to improve the recognition rate. And in speaker recognition, with the increasing in the number of speakers, the time it takes to identify is increasing in a rectilinear fashion. Because every time recognition must be matched with every speaker model orderly, and then find the closest corresponding speaker model as the final recognition result. In this way, the more registered number, the longer discriminating time, it must be reached by a limit that leads to a very long time to identify, it cannot meet the requirement finally. In this case it could be nicely solved adopting classification.

2.1. Pitch Frequency

The pitch has aroused the periodicity through vocal cords vibration when madding voiced sound, pitch frequency is

a very important parameter using to describe the characteristic of voice excitation source. The variational range of pitch frequency is generally from 50 Hz to 500 Hz, the cycle of the male voice is 50 Hz - 300 Hz, and the female is 100 Hz - 500 Hz. Although each person's different vocal structure lead to different fundamental frequency, because of the pitch frequency's scope is a little small, the gap between different people is little, and the most important is pitch frequency is affected by a lot of factors, such as emotion, tone, it is very difficult to achieve accurate fundamental frequency. Thus, the recognition rate is very low using the fundamental frequency for speaker recognition now. But male fundamental frequency is generally lower than the female, it is a good argument as classification.

Since the research of voice signal analysis, pitch extraction is always an important research topic. Speech signal changes complexly, which is affected by channel and has an ample harmonic constituent. Although many methods have been proposed at present, they all have limitations, cannot delegate speaker's different characteristics, and can not adapt to different requirement and environment.

There are a variety of methods to extract the fundamental tone [1]. These can be roughly divided into three categories, wave form estimation, correlation process and converter technique [3]. This paper used a converter technique to extract pitch, it transforms the speech signal to the cepstrum domain, eliminates channel impact using homomorphic analytical method, then obtains the information of pumping part, and ermittelts fundamental frequency.

Only voice sound has pitch alternation. The glottal excitation is less energy and white noise of spectrum evenly distributed when madding voiceless; when madding voice sound, it is a shock sequence having a certain period. This period is the pitch alternation. A finite length sequence of periodic impulse has a periodic impulse sequence in cepstrum domain $s(n)=\sum_{r=0}^{M}\alpha_r\delta(n-rT_p)$, M is positive, α_r is crest factor, T_p is pitch alternation, and the period cannot change in cepstrum domain, the amplitude increases along with r and the rate of decay is faster than in the time domain. In this way, the method based on cepstrum can be used to extract fundamental frequency and it has a better effect.

Lab settings: Intel(R) Core(TM)2 Duo T6400, 2 GHz memory, Windows XP system, MATLAB7.0 development platform, the experiment's voice data use Cool Edit Pro to transcribe, sampling frequency is 16,000 Hz, sampling precision is 16 bit, single track, the age of recorded speaker is in 8 - 60 years old, speaking mandarin, everyone speaks 7 sentences, the time of every sentence is in 3 - 12 s, including vowel, consonant, Chinese, English and figure.

The experimental results were shown in **Table 1**, every speaker's pitch frequency could not be accurately achieved with this method. The result appears in a scope rather than is a exact value. The scope of different people's pitch frequency has a small gap and intersection. So it is clearly not feasible only with a frequency value in the speaker recognition. The male voice's pitch frequency is generally lower than the female, therefore, pitch can be used to distribute speakers.

2.2. Formant

Formant information include in spectral envelope. The formant is generally the maximum of spectral envelope, so the necessary procedure of extracting formant is to estimate spectral envelope.

Methods of fetching formant contain cepstrum method and linear forecasting method [1]. Formant generally defined as the attenuation sine component of sound channel impulse response. A primary question for extracting formant is that impulse response of the sound channel cannot measure directly. Voice signals are the convolution of all pole model and glottal quasiperiodic function, so when analyzing, it must solve convolution, separate impulse response and excitation function.

The paper adopts linear forecasting method to estimate formant, the specific method is peak detection. Analyzing formant with linear predictor coefficients is faster and better than others. The track function which is described by linear predictor coefficients (LPC) is computed firstly, the function is used to compute the spectrum, according to the spectrum, the formant's peak, frequency and bandwidth are computed [4].

The experimental environment is identical with pitch's.

The **Table 2** shows that each formant could change when the same person said different word. Even though the same person's value of alteration has a scope, this scope includes others'. Therefore formant parameter cannot be the effective one in speaker recognition. Experimental data proved that children's value of F1 are higher than adults', so the parameter can be used to distinguish between child and adult.

Talbe 1. The result of pitch frequency with cepstrum method.

	Voice 1	Voice 2	Voice 3	Voice 4	Voice 5
Woman 1	333	301	307	311	318
Woman 2	266	262	210	250	243
Woman 3	262	231	280	271	250
Woman 4	202	213	210	220	206
Man 1	183	195	183	178	181
Man 2	172	141	168	141	156
Man 3	121	133	124	132	108
Man 4	112	109	134	114	129

Talbe 2. Formants of speech signals.

		F1	F2	F3
Adult	Man 1	704	1174	2456
	Man 1 (different content)	616	1831	2891
	Man 2	652	1323	2721
	Man 3	581	1830	2519
	Man 4	438	1614	1780
	Woman 1	618	1814	2617
	Woman 2	544	1834	2960
	Woman 3	551	2653	2630
	Woman 4	590	1210	2279
Child	Girl 1	749	1405	1643
	Girl 2	1015	1733	2314
	Boy 1	904	1353	2990
	Boy 2	837	1379	2560

2.3. Mel Frequency Central Coefficient

In a noisy environment, people can also identify correctly different sound in the ear, the important reason is the cochlea played a role. The cochlear is equivalent to a set of filters, the filters are done to the signal on logarithmic frequency scale, and so human ear is more sensitive to low frequency signals [5].

A set of Mel filters of imitating the role of the cochlea are triangular filters, the center frequency is equispaced in the Mel frequency axis, and they have the same span on the Mel frequency scale. The number of filter bank is decided by cutoff frequency of signal, all of the filter bank collectively cover between 0 and 1/2 sampling frequency.

To emphasizing low frequency information of the signal, MFCC change the linear frequency scale into Mel frequency scale, so useful information for identifying is highlighted and the noise jamming is shielded effectively. If Mel cepstrum is used, filtering and weighting in the cepstrum domain are based on linear spectrum processing [6].

MFCC generally reflect the static characteristics, but the human ear is more sensitive to the dynamic characteristics of voice. ΔMFCC can reflect dynamic property. This parameter can be acquired by computing first-order difference and second. The paper uses the parameter combining 12 dimensions MFCC with ΔMFCC.

The experimental environment is identical with pitch's. Five methods of comparing to two MFCC were attempted.

1) Correlation coefficient

In theory, the correlation coefficient is the maximum when the same person speaks the same word, and it is the second highest when the same people speak different words. Only in this way, could the speaker be identified. Analyzing experimental result, the **Table 3** shows that the correlation coefficient of female speaker L cannot be identified, because the value of the same people speaking different words is lower than the different people speaking the same word. So this means cannot be resultful method in speaker recognition.

Table 3. Correlation coefficient of MFCC of two voice signal.

	y	L	x
Same people same content	0.5298	0.6947	0.6371
Same people different content	0.3665	0.4116	0.4446
Different people same content	0.4161	0.6666	0.4463
Different people different content	0.3932	0.5084	0.4544

2) Comparing to similarity of corresponding three-dimensional map

The data of the **Figures 1** and **2** are from the same person. Although they are similar in general, there are many data of MFCC, they don't have regularity, the drew three-dimensional map is intricate, after smoothing, it is difficult to compare the similarity.

3) Comparing related coefficient of each column

Because each MFCC dimension is uncorrelated, they can be compared independently.

As shown in **Table 4**, this method is not useful for comparing MFCC. The part of women in different people different content is larger than same people same content in related coefficient of the first dimension. It is a negative relationship in the first, second, fifth, sixth, tenth and twelfth dimension of same people same content. So this method cannot serve as the way of comparing MFCC.

4) Euclid distance

Table 5 shows that the euclid distance is minimum when the same person speaks the same word, and it is second smallest when the same person speaks different words. Regardless of what the speaker said, the minimum Euclidean distance corresponding to the speaker is the recognition results.

3. Conclusions

Pitch and formant are both the most important parameters of the speech signal. In theory, because of the differences of buccal structure and sound track, everyone should have their own different characteristics of pitch and formant. Speech signal changes in complex, sound channel and noise have an effect on the signal, and extracting methods are imperfect, so pitch or formant is not an effective parameter in speaker recognition recently, they can only play a supporting role. MFCC is effective for speaker identification, because it combines sensing features of the human ear with producing mechanism of voice.

Speaker's personality can not be represented well by a single parameter, using only one just describes part of

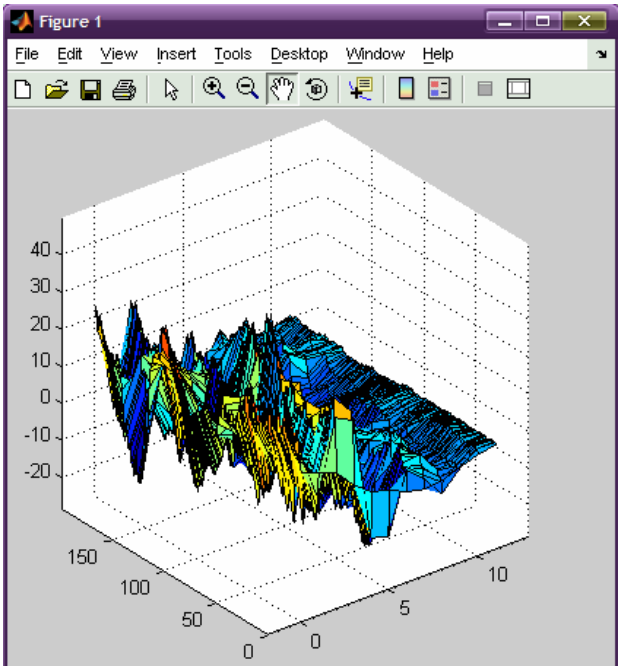

Figure 1. Three-dimensional map of female vowel.

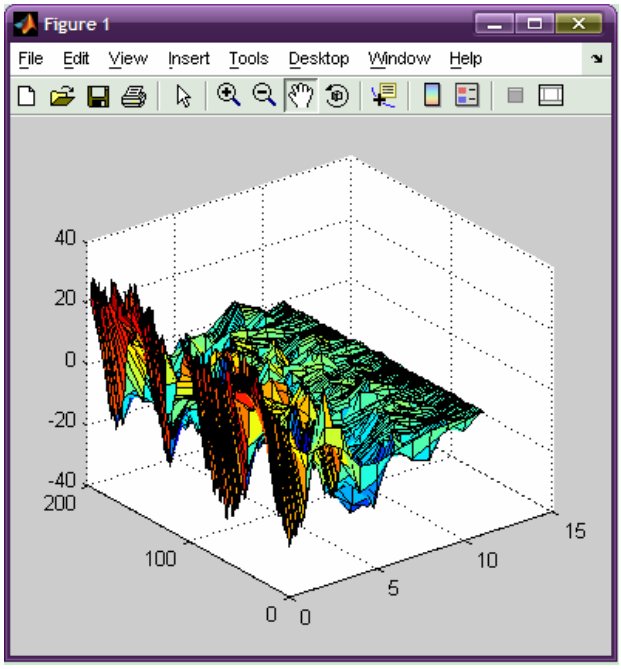

Figure 2. Three-dimensional map of female consonant.

Table 4. Correlation coefficient of each dimension of MFCC.

	Same people same content	Same people different content	Different people same content		Different people different content	
			woman	man	woman	man
1	−0.0053	−0.3237	−0.1266	−0.1800	0.0782	−0.2611
2	−0.3008	−0.1839	−0.2274	−0.3389	0.2223	0.1788
3	0.4592	0.3923	0.5468	0.3992	0.3423	−0.1024
4	0.3984	−0.1197	−0.0152	0.4367	0.2005	0.2433
5	−0.0324	−0.1635	0.3900	0.2537	−0.2118	−0.1130
6	−0.0547	0.0859	−0.2082	−0.0695	0.1711	−0.0897
7	0.0890	−0.0764	0.1685	−0.2432	−0.0870	0.1036
8	0.0187	0.2787	0.1434	0.1532	0.1763	−0.1018
9	0.4090	0.0681	0.2142	0.0786	0.2639	−0.0544
10	−0.0865	−0.1368	−0.0766	−0.1188	0.2791	0.3179
11	0.1750	−0.3922	0.4273	−0.2079	−0.2229	0.3064
12	−0.1299	0.0124	0.3571	0.0360	0.2908	0.0185

Table 5. Euclid distance of MFCC of two speech singnal.

		y	L	x
Same people same content		92,338	41,214	79,086
Same people different content		124,110	90,346	139,190
Different people same content	**woman**	141,340	94,724	199,120
	man	182,240	183,370	140,270
Different people different content	**woman**	176,860	92,334	219,090
	man	149,040	116,800	188,970

speaker's characteristics, therefore, to improve the speaker recognition rate, many parameters should be combined to identify.

REFERENCES

[1] H. Hu, "Introduction to Speech Signal Processing," Harbin Institute of Technology Press, Harbin, 2000.

[2] X. J. Yang and H. S. Chi, "Digital Processing of Speech Signals," Electronic Industry Press, Beijing, 1995.

[3] M. M. Sondhi, "New Methods of Pitch Extraction," *IEEE Transaction on AU*, Vol. 16, No. 1, 1968, pp. 262-266.

[4] K. Du, "LPC Analysis on Formant of Speech Signal," *Natural Science Journal of Harbin Normal University*, Vol. 2, 1998, pp. 49-52.

[5] N. Do Minh, "An Automatic Speaker Recognition System," Audio Visual Communications Laboratory Swiss Federal Institute of Technology, Lausanne, 2001.

[6] Y. Chen, Z. Y. Qu, Y. Liu, K. Jiu, A. P. Guo and Z. G. Yang, "Extraction and Application on One of Speech Parameters," *MFCC Journal of Hunan Agricultural University* (*Natural Science*), Vol. 35, No. 1, 2009, pp. 106-107.

