The separability analysis of pitch parameters in speaker recognition

Rong Rong He[1, a], Lan Tian[1], Fa Jiang Ma[1]

[1]*School of Information sci. and Eng., Shandong Univ., Jinan 250100, China*

Abstract. Aseparability research of prosodic parameters in speaker recognition is studied.The space distribution of the prosodic parameters is drawn using the mean and variance of the speaker pitch parameters,through which the separability of the speaker is obvious.We do a speaker identification experiment with a small corpus,and the experiment results show that speaker identification based on the prosodic parameters is possible. The first-level decision using the prosodic parameters can reduce the candidate corpus effectively,demonstrating the practicability of the pitch parameters.

Keywords: speaker recognition; pitch parameters; feature space

Introduction

Speaker recognition is a branch of the traditional speech recognition and it has wide application prospects in telephone financial transactions, network security, multimedia and other fields. According to their final tasks , it can be divided into two kinds: speaker identification and speaker verification two kinds. According to the testing speech inputted, it can be divided into text-related and unrelated categories. The correct rate and speed of the speaker recognition system is an important indicator to evaluate a system, but there're contradictions between accuracy and efficiency in the general recognition system.

At present, most of the speaker recognition models of the system are based on the channel characteristic parameters with little regard to the sound source characteristics. However, the sound source characteristics (such as pitch) is an important information source of the physical characteristics of the speaker and their speaking habits, which should be given adequate attention. This paper presents a method for preliminary speaker identification which takes the features of pitch into accountWe plotted the diagram of the speaker's pitch feature space distribution according to the statistical properties of the mean and variance. It can be observed that the pitch figure region doesn't coincide for different speakers, and that pitch parameters in the feature space are separable, which provides a basis for the design of speaker recognition based on the pitch feature. he As we used this method for primary judgment it can effectively narrow the range of candidate set[1], if jointly use the typical speaker recognition method for careful judgment, we can ensure that the system can have good performance in accuracy and efficiency.

Analysis of the characteristic parameters of pitch

The extraction of the feature parameters is crucial for speaker recognition[2]. Currently, speaker recognition rely mainly on the acoustic characteristics of the lower level, among which the relatively effective features[3] are: short-term speech energy, pitch period, short-term spectrum, BPFG voice characteristics, the linear prediction coefficients, formant frequencies and bandwidths, the LPC cepstrums, etc., and some linear regression coefficients that reflect the dynamic changes of those characteristics. The fundamental frequency parameters (also known as pitch parameters) are the most important parameters of the speech signal, which describes the important features of the speech excitation source, and has been widely used in many areas. People's fundamental frequency range is generally 60 ~ 500 Hz, men mainly in the range of 60 ~ 220 Hz, while women and children in the range of 150 ~ 450Hz. Different people have different pitch distributions, in spite of the disadvantages of being easily imitated, it can be used as an auxiliary judgment in some specific occasions (such as studio monitors) for speaker recognition.

Pitch features are always mixed in the track characteristics and the speech signals of the speaker instead of existing independently in the actual speech signals. To accurately extract the signal, it is necessary to adopt appropriate detection algorithm. Currently, there is a variety of pitch detection method. [4], such as the central clipping autocorrelation method, average magnitude difference function method, method of parallel processing, the wavelet method and cepstrum method, etc. In analysis, we using the cepstrum method, its principle is as follows. The speech sound s (n) is inspired by the glottis pulse source, responded by track e(n) and filtered by v (n), namely: s (n) = e (n) * v (n). If the three corresponding cepstrums are s (n), e (n) and v (n), then: s (n) = e (n) + v (n). Because of the separation of e (n) and v (n) in the cepstrum signal, specifically the pitch information and track information are distributed in different time section, so simply adopting the cepstrum filtering method [5]can separate the glottis excitation signal e (n), and then get its circle value.

In the cepstrum analysis, speech widowing is also of great importance. We choose gradual widows as for the

[a] Corresponding author: herongrong1993@163.com

widows function, Hamming window is the usual choice. To predict the pitch cycle from the high point in the cepstrum waveform, in other words, to estimate the position of the peak in it, we can get corresponding incentive source pitch position, which is pitch cycle.

In actual pitch detection, going through a filter is necessary to filter caused by the corresponding high fluency of the sampling theorem and the interference 50 Hz power fluency. After framing widowing the pitch data, the endpoint detection is conducted to get rid of the silence segments or surd segments. Then we use cepstrum method to calculate the pitch cycle. Because the channel response may produce harmonic interference, such as 1, 2, 1, 3, 2, 3 times frequency errors, so generally pitch trajectory smooth processing are carried out to remove the "singularity" pitch trajectory. Smoothing process consists of 3 kinds, namely, median smoothing and linear smoothing and combination smoothing. Median smoothing can effectively remove the small amounts of the singular points and at the same time, do not destroy the staged transformation between the two pitch tracks in the surrounding area of pitch cycle. Although linear smoothing can improve the effect of smooth, meanwhile it will lead to staged fuzzy between two smooth sections. So these methods can be chosen according to their effects. Figure 1 is the design sketch of smoothing process in the pitch analysis.

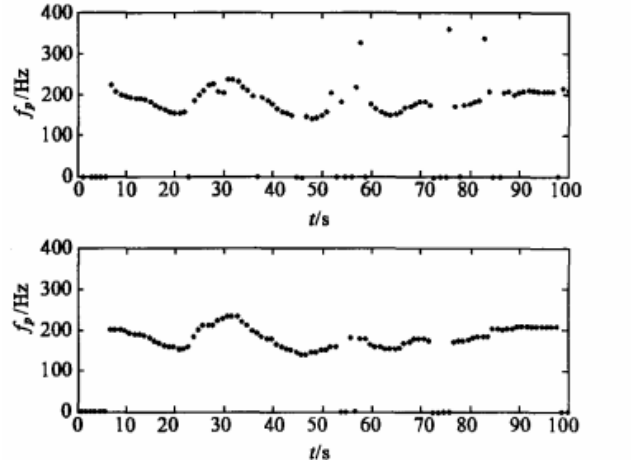


Fig.1 The smooth process of the pitch contour

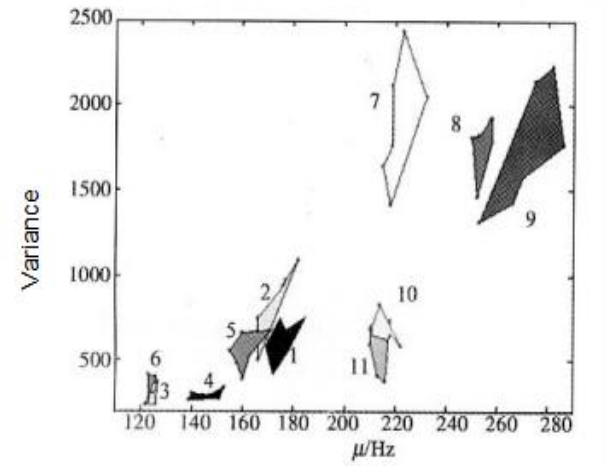
Speaker pitch feature space analysis

We established a group of 11-persons voice library for analysis, six men and five women, the training voice and the testing voice of which are the natural speech flow which are not limited in the content. The speech flow is collected and recorded in the clean speech environment (the mute condition at laboratory) by telephone voice card, which is 120 seconds in length. The first 60 s are set for training, and the rest for test. The sampling frequency is 11 kHz, and it's in 16-bit-quantization storage.

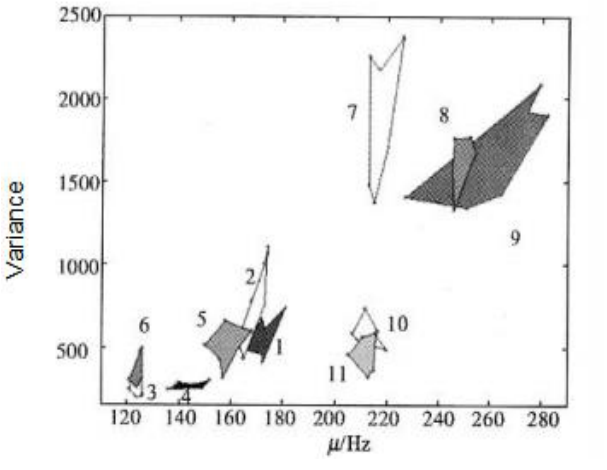
According to the method above, we obtained the pitch period for each frame. After that, we smooth the track, and then calculated their long-time average, deriving the average of the pitch period.

Due to the speaker's individual physiological features and habits, the range of the pitch change is also different. In order to fully reflect the characteristics of the distribution of different speakers' pitch period, we consider adding another statistical property – the variance of pitch period, therefore the mean and variance constitutes the pitch feature space. Figure 2(a) is a collection of the mean and variance of pitch feature space distribution (when the window length is 40 ms, and the frame shift is 20ms) of the 11-person group; Figure 2(b) is characteristic spatial distribution when the frame-shift is not considered.When drawing the pitch characteristic spatial distribution diagrams, each speaker's voice materials are divided into six segments and the mean and variance of each segment are calculated respectively. Then we can obtain the corresponding six points in the feature space, and connect them into a block area with the largest square, which corresponds to the general distribution of its pitch.

It can be seen from the figure that the male speakers are distributed at the left bottom of the basic characteristics of space, while female speakers are distributed on the right of the feature space. This phenomenon is because the male voice tone is low with a relatively small range, while female voice is higher with a larger voice range. Theoretically, speaker identification can be implemented by mapping the mean and variance of the speaker's voice segments into the distribution diagram.



(a)



(b)

Fig.2 The space distribution of the pitch with the frame shift(a) and without the frame shift(b)

The experiment on identifying the speaker

Based on the foregoing analysis, We use the characteristic of pitch in the experiment on identifying the speaker from the text in set. The length of time of training words is assumed as 60s and that of identifying words is assumed as 10s. The models for reference of the speaker are the mean and variance in pitch period. Given that the vnit for measures of pitch period and variance is different and the distance measure can not be used directly. We put them together into the range from 0 to 1, and then we do distance measure. The way to match is to adopt the simple model, and contrasting the parameter of the speakers to identify to all the speakers'. The speaker who have minimal distance is the likely one. The outcome of experiment suggest that the percentage of identifying correctly is up to 78%.

Certainly, the accuracy of identifying speakers depend on the character of pitch is not high. But because the draw of character of sipn can be tested by specific hardware and is not relevant to common track character parametric modeling identification, and can not replace the contribution stage of system. When use speaker to identify the telephone channel, the parameter has strong anti-noiseness and robustness because of its low occupancy frequency range. We can do anticipation judgement before in the fixed text and variable text's identifying speaker. For a speaker identification system of 60 people, we adopt a pitch anticipation strategy to test. We broaden preliminary gauge threshold and take 30%-close speaker enter into candidate set, and then we use method of VQ and HMM to do the final judgment of speaker. On the condition that the identification rate of system not reduce, it can largely decrease the range of candidate, and the efficiency of identification are increased up to over 60%, the reliability of the system is also improved.

Conclusions and suggestion

By analysing the space distribution of the character of speaker's pitch and the experiment of identifying actual speaker, this paper confirms that the application value of character of pitch in experiment of identifying speaker, and also the applicating methods are given. The methods are not perfect. Because the character of pitch of speaker's character of space distribution is likely to influenced by the emotion, body condition and applicating environment. This pretreatment method is not fit for a variety of reasons not to cooperate with the speaker's words used specification(for example, imitate other's voice). For this, we should add other measures to avoid the unrecoverable error. On the occasion of appropriate control using occasions, to avoid error of initial judgment, sometime we must broaden the initial judgment's scope. As a result the efficiency of identification of system will drop. If the statistical features of pitch can consist more whole feature vector with other feature(like voice channel feature), the forecasting error can be avoid easily, and effect of recognition will be better.

Acknowledgement

In this paper, the research was sponsored by the Nature Science Foundation of China (Nos.11474185,61271453) and the Fundamental Cross-decipline Research Foundation of Shandong University(No.2015JC029).

References

[1] MOU Xiao-long,HU Qi-xiu,WU Wen-hu.Text-independent speaker identification system based on multiple strategies[J]. Journal of Tsinghua University(Sci&Tech),1997,37(3):16-19.

[2]Kong Y Y, Zeng F G. Temporal and spectral cues in Mandarin tone recognition. J. Acoust. Soc. Am., 2006; 120:2830-2840

[3]TIAN Lan,BAI Shu-zhong.Based onMulti-features and multistages decision speaker identification in telecommunication network[J].Journal of Shandong University(Engineering Science),2003,33(6):648-651.

[4]YI Ke-chu,TIANBin.Speech signal processing[M].Beijing:National Defense Industry Press.200.1-10.\

[5]ZHAO Li.Speech signal processing[M].Beijing:Mechanical Industry Press,2003.1-10.