Feature Extraction and Analysis of Speech Signal Based on Fractional Fourier Transform

Yueying Zhai

Wuhan Qingchuan University, Wuhan, 430000, China

Abstract: *The extraction of speech features is the basis of speech signal processing, and the extraction of speech features is to obtain the parameters representing speech signals through the analysis of speech signals. The shape information of the signal can be extracted and processed by selecting appropriate structural elements and adopting mathematical morphological transformation. Selecting different structural elements will result in different shape transformation results, thus extracting the shape information of different components. In this paper, the feature extraction and analysis of speech signals are further studied based on FRFT (fractional fourier transform). The simulation results show that in different noise environments, such as when the speech is bathed with signal-to-noise ratio of 0dB and 10dB respectively, the recognition rate of this method is higher than that of the traditional parametric method. In this paper, based on the strength of FRFT signal components, the component signals are detected one by one in order, and then according to the detection results, the strongest component signals are removed from the observed signals in order to reduce their influence on the weak component signals, thus improving the effectiveness and reliability of multi-component signal detection and parameter estimation.*

Keywords: *Fractional Fourier transform, Speech signal, Feature extraction*

1. Introduction

As a tool for spreading information, language is mainly manifested in pronunciation, which provides a more convenient way for communication and thinking than words. Human language originated from phonetics, which gradually evolved into words, rather than words followed by phonetics. The extraction of speech features is the basis of speech signal processing, and the extraction of speech features is to obtain the parameters representing speech signals through the analysis of speech signals. The shape of the signal is determined by its own sampling value. The shape information of the signal can be extracted and processed by selecting appropriate structural elements and adopting mathematical morphological transformation. Selecting different structural elements will result in different shape transformation results, thus extracting the shape information of different components [1]. Mathematical transformation has the characteristics of locality, parallelism and easy hardware realization. With the development of science and technology, people have higher and higher requirements for the speed and quality of information transmission. The research and exploration of modern speech signal processing technology can make people's speech information generation, transmission, storage and retrieval more efficient [2]. People use short-time processing technology for speech signals. Speech signal is a kind of short-term stable signal, which changes instantly and is very complex. It carries a lot of useful information, including semantics, personal characteristics, etc. The accuracy and uniqueness of its characteristic parameters will directly affect the speech recognition rate, and this is also the basis of speech recognition. Because the voice changes slowly, in a short period of time, it can be considered that the voice signal is stable and time-invariant. Based on this, the speech signal can be divided into some successive short segments for processing. When the framing is small enough, the nonlinear system can be approximated by the linear system [3-4]. Because the nonlinear frequency modulation signal is intercepted in a short time, when the window function is selected reasonably, each segment of the intercepted signal can be approximately regarded as a linear frequency modulation signal. Using the high time-frequency aggregation of the linear frequency modulation signal by FRFT, the high time-frequency resolution result of the nonlinear frequency modulation signal can be easily obtained. Therefore, short-time FRFT can improve the time-frequency estimation accuracy of nonlinear FM signals.

With the development of signal processing, computers can hear voices, see pictures and speak, among which voice has more advantages in future human-computer interaction, and it is also the direction with the greatest development potential. In signal processing, when short-time Fourier transform is used to estimate the instantaneous frequency of FM signal, the estimation accuracy may be low [5]. Therefore, this paper proposes a method of extracting the characteristic parameters of speech signals based on FRFT. Finally, the effectiveness of this method is verified by simulation experiments, which shows that this method can improve the recognition rate of speech signals to a certain extent.

2. Definition and Properties of Fractional Fourier Transform

In power electronic circuit fault diagnosis, Fourier analysis is a commonly used method. We try to use FRFT instead of conventional Fourier analysis to study its application effect in power electronic circuit fault diagnosis. The traditional Fourier transform is the most widely used and well researched mathematical tool in all signal processing tools. As a linear operator, the traditional Fourier transform can be seen as a signal rotating π/2 counterclockwise from the time axis to the frequency axis on the time-frequency plane. As a generalized form of FT, FRFT can be understood as a linear operator rotating the signal at any angle, Moreover, on the basis of retaining all the properties and advantages of traditional FT, new advantages have been added [6]. FRFT has been gradually used in feature extraction and fault diagnosis because it can subdivide the conversion process from time domain to frequency domain, gradually transform, and display all features from time domain to frequency domain [7-8]. There are many definitions of FRFT, but their starting points are different, but they are all equivalent. If a signal takes a non-zero value in a subset of the time axis or frequency axis, and the conditions for taking a non-zero value are limited to a finite interval, the signal is said to be compact on the time axis or frequency axis. However, the signals processed in practice are often time limited and band limited. The time bandwidth product of the signal can be used to determine the sampling frequency and sampling points of the signal, and to recover the original signal from the discrete signal. The selection of signal bandwidth is not required to be the minimum, as long as all the energy of the signal is included in it, so the sampling frequency can be used as the bandwidth. The scale factor S and normalized width x can be obtained as follows:

$$S = \sqrt{\Delta t / \Delta f} = T / f_s \tag{1}$$

$$\Delta x = \sqrt{\Delta t \bullet \Delta f} = T \bullet f_s \tag{2}$$

The original sampling interval of discrete data is $T_s = 1/f_s$, and the sampling interval obtained after scale transformation is:

$$T_s = 1/\sqrt{T \bullet f_s} = 1/\Delta x \tag{3}$$

Before normalizing the original signal, the definition of signal compactness is given first. If the non-zero value of a function is limited to an interval, the function is said to be compact. Theoretically, a signal cannot be compact simultaneously in time domain and frequency domain. The discrete scaling method is to realize the normalization by scaling the discrete data in the time domain. The scaling of the signal will inevitably cause some features of the original signal to be distorted. For example, scaling a chirp signal will make its frequency modulation rate larger or smaller. FRFT can be understood as a special time-frequency representation. This section will briefly introduce the internal relationship between FRFT and several common time-frequency representations. However, the signals studied in practice are generally considered to be finite in time and bandwidth. This theory does not conflict with the actual difference, because as long as most of the energy of the signal is concentrated in a limited area, we say that the signal is compact in the time domain and frequency domain [9].

3. Feature Extraction and Analysis of Speech Signal Based on Fractional Fourier Transform

3.1 Speech frequency feature extraction method

The process of extracting speaker's speech signal features is actually a process of removing redundant information in the original speech and reducing the amount of data. The features should have the following characteristics: First, it can effectively distinguish different speakers, and at the same time, it can keep the pronunciation of the same speaker relatively stable when it changes; Second, it is easy to extract from voice signals; Third, it is not easy to imitate. We know that speech can be divided into voiced and unvoiced. The vocal cords vibrate under the action of airflow, thus producing quasi-periodic acoustic excitation, which is sometimes called voiced sound through the resonance of oral cavity and nasal cavity. The air flow passes through the labial part of the oral cavity, and if it causes turbulence, it will produce fricative sound. Sometimes, because the lips suddenly open, it will form explosive sound. Any sound that the vocal cords do not vibrate is collectively called unvoiced sound, and sometimes it is also called no sound. As we know, when the traditional autocorrelation method is used to obtain the pitch features, because the fixed threshold is used for clipping in each frame, it is often impossible to effectively delete the non-main components in the waveform of the speech signal, especially at the beginning and end of the speech, and clipping with a fixed closed value becomes more difficult [10]. In order to solve this problem, we consider using threshold clipping which adaptively changes with the speech signal instead of fixed threshold clipping. The principle block diagram of speech signal feature extraction based on FRFT is shown in Fig. 1.

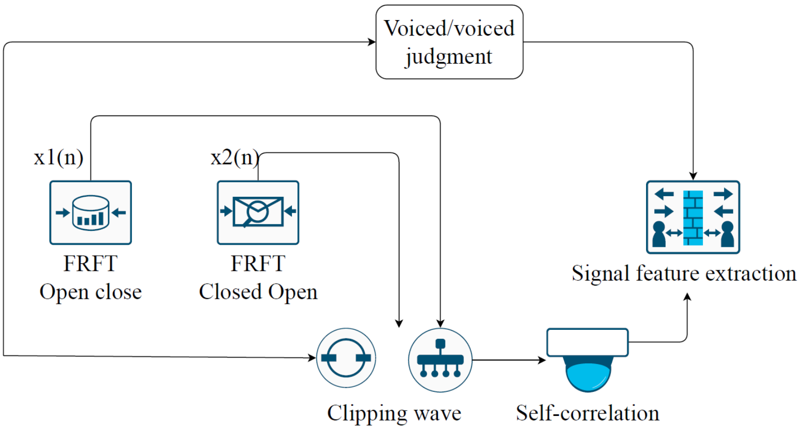

Figure 1: Principle block diagram of speech signal feature extraction based on FRFT

In the process of signal detection, the existence of strong component signal will inevitably affect the detection and parameter estimation of weak component signal. In order to obtain the adaptive cutoff value and two of positive and negative clipping of speech signal, we choose morphological opening and closing, morphological closing and opening as morphological filters of speech.

$$x1(n)=(x\otimes B) \tag{4}$$

$$x2(n)=(x\otimes B) \tag{5}$$

Here, B is a symmetrical line about the origin, which is equivalent to a window in linear signal processing.

Therefore, in order to effectively detect multi-component signals and estimate unknown parameters, certain measures must be taken to weaken the influence of strong signals on weak signals to weaken the influence of strong component signals The decomposition at all levels of FRFT can effectively decompose this type of target signal from different frequency bands. Next, use the decomposed coefficients at all levels of this type of target wavelet tree to extract the signal characteristics of a type of target [11-12]. Generally speaking, there are time domain, frequency domain, cepstrum domain and other analysis methods for speech signal analysis, but from the perspective of phonetics, it can be divided into model analysis method and non-model analysis method. Model analysis is based on the theory of mathematical model of speech signal to obtain these model parameters, while other analysis methods are generally classified into non model analysis methods. According to the time-frequency analysis principle, if the time-frequency distribution of the signal is highly concentrated in a certain dimension, the composite signal can be simply compressed and translated in this dimension and then sent through the sending end. If the translated sub channels are orthogonal to each other in this dimension, the receiving end will restore the transmitted signal without distortion, which is the essence of multiplexing technology. Based on the strength of FRFT signal components, component signals are detected one by one in order. Digitized voice signal sequences are stored in a data area in turn. In voice signal processing, these data are generally stored in a circular queue, so that a limited capacity data area can be used to deal with a large number of voice data. Processed voice data can be discarded in order to make room for storage to store new data.

Speech signal can be constructed by FRFT frame structure, decomposition scale and feature parameters. The frame structure refers to the sequence formed in the time dimension by framing and windowing after the speech signal passes through the DC removal, pre emphasis, endpoint detection and other pre-processing links; Decomposition scale refers to the wavelet decomposition of each frame of speech signal to obtain the approximate components of each frame of speech signal and the detail components on different scales. When processing in FRFT, data is taken from this data area by frame, and the next frame is taken after processing. Then, according to the detection results, the strongest component signal is removed from the observed signal in order to reduce its impact on the weak component signal, thereby improving the effectiveness and reliability of multi-component signal detection and parameter estimation [13].

3.2 Comparative analysis of speech features

In this chapter, simulation experiments will be conducted to further verify the effectiveness of the proposed method. First, record the voices of six people, each of whom recorded two sentences: "Student" and "I am very happy". The statement is gradually complicated, so that the effect of the new feature parameters can be better verified. The traditional parameters of speech signals and the new characteristic parameters proposed in this paper are extracted respectively. During the test, the speech signals are dyed with noise, and noise is added to each speech with signal-to-noise ratio of 0dB and 10dB respectively. After the similarity is calculated, it is normalized, and finally the recognition rate of each speech is obtained. Table 1 shows the recognition rates of speech signals under different signal-to-noise ratios.

Table 1: Recognition rate of speech signal under different signal-to-noise ratios

Voice signal	0db		10db	
	Traditional parameters	New feature parameters	Traditional parameters	New feature parameters
Student	95%	96%	93%	95%
I am very happy	89%	92%	88%	90%

It can be seen from Table 1 that in different noise environments, such as adding 0dB and 10dB SNR to speech respectively, the method in this paper can achieve higher recognition rate than the traditional parameter method. This shows that this method can improve the speaker's speech recognition rate to a certain extent, and can basically achieve the desired purpose.

Secondly, the audio signal "Ah Oh, please check if you have new messages" is collected. At the same time, the voice also contains some background noise, which has a certain impact on the extraction and transmission of information. The sampling frequency of the acquired signal is 22600 Hz. From the voice signal waveform, it can be roughly analyzed that the amplitude is mainly distributed between 0~1.5, the tone is high and low, including high frequency and low frequency components, and the duration is 276s, as shown in Fig. 2.

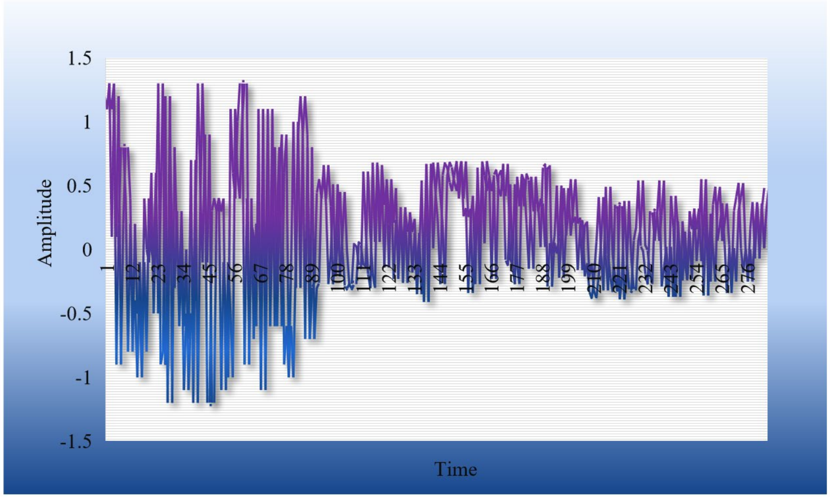

Figure 2: Waveform of Voice Signal

In the experiment, two kinds of additive noises and two different signal-to-noise ratios are used. It can be seen from Table 2 that the recognition rate of noisy speech recognition based on FRFT extracted features proposed in this paper is higher than that of feature parameters after tensor decomposition. It shows that the method proposed in this paper is robust in noisy speech recognition.

Table 2: Experimental results of noisy speech recognition

Experimental method	**Recognition result**			
	CAMRY		ROVER	
	5db	**10db**	**5db**	**10db**
Tensor decomposition feature	**90.00**	**90.62**	**83.65**	**89.40**
FRFT	**91.34**	**92.95**	**87.89**	**91.53**

In the speech recognition experiment, the non-Gaussian speech signal is affected by the Gaussian noise with the signal-to-noise ratio of 10 db. The effectiveness of the features after tensor decomposition and FRFT on each specific phoneme in the signal is studied. The performance comparison results of the features after tensor decomposition and FRFT are shown in Fig. 3.

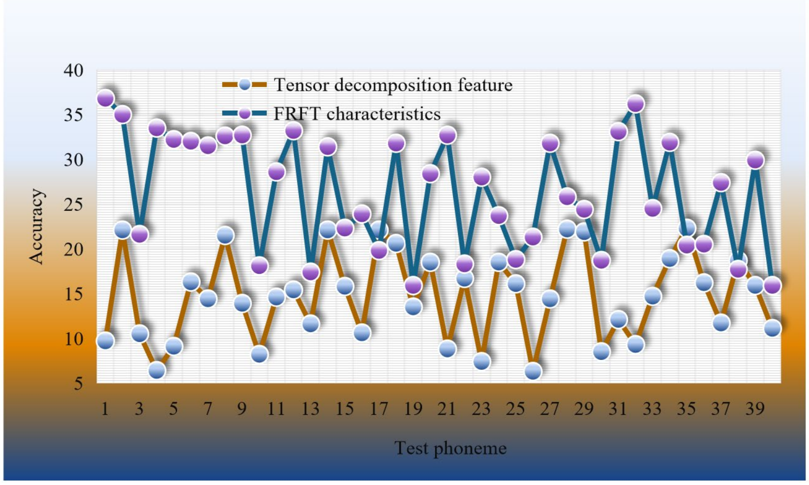

Figure 3: Speaker recognition accuracy of Gaussian noise speech signal with SNR of 10db

As can be seen from Figure 3, the recognition performance of the two features is low, which is mainly because other phonemes contain lower signal energy than vowels, so they are more obviously disturbed by noise. The FRFT features proposed in this paper are obviously better than those after tensor decomposition.

4. Conclusions

Speech signal is a short-time stationary signal, which is transient, very complex, and carries a lot of useful information, including semantics, personal characteristics, etc. The accuracy and uniqueness of its feature parameters will directly affect the level of speech recognition rate, and this is also the basis of speech recognition. In this paper, the feature extraction and analysis of speech signals are further studied based on FRFT. The simulation results show that in different noise environments, such as adding 0dB and 10dB SNR to speech respectively, the recognition rate of speech recognition using this method is higher than that of traditional parameter methods. This shows that this method can improve the speaker's speech recognition rate to a certain extent, and can basically achieve the desired purpose. Before normalizing the original signal, the definition of signal compactness is given first. If the non-zero value of a function is limited to an interval, the function is said to be compact. Based on the idea of finite automata, a model of fuzzy query control is established. On this basis, query modularization is realized. The whole control system is divided into two parts based on FRFT, namely data analysis and processing and specific query control. In this model, the introduction of the idea of finite automata is discussed, which enriches and develops the theory of computational intelligence in application.

Acknowledgements

Methods Study on the emotional feature extraction and analysis of speech signal (No. B2020319)

References

[1] Dreuw P. Method for recognizing a voice context for a voice control function, method for ascertaining a voice control signal for a voice control function, and apparatus for executing the method. vol. 67, no. 26, pp. 34-58, 2018.

[2] Ikuma T, Mcwhorter A J, Adkins L, et al. Time-varying harmonic models for voice signal analysis. vol. 49, no. 23, pp. 44-67, 2022.

[3] Yu xing. Li, Xiao. Chen, Jing. Yu, Xiao hui. Yang, and Hui jun. Yang, The data-driven optimization method and its application in feature extraction of ship-radiated noise with sample entropy. Energies, vol. 12, no. 3, p. 359, 2019.

[4] Yu xing. Li, Long. Wang, Xue ping. Li, and Xiao hui. Yang, A novel linear spectrum frequency feature extraction technique for warship radio noise based on complete ensemble empirical mode decomposition with adaptive noise, duffing chaotic oscillator, and weighted-permutation entropy. Entropy, vol. 21, no. 5, p. 507, 2019.

[5] Zhi yan Han and Jian Wang, Dynamic feature extraction for speech signal based on formant curve and MUSIC. Chinese Control and Decision Conference (CCDC), Chongqing, China, pp. 403-407, 2017.

[6] Zhao R B, Rui Y B. Micro-Doppler feature extraction method under low signal-to-noise ratio. Information Technology, vol. 63, no. 10, pp. 15-65, 2017.

[7] Yan zhu Hu, Song Wang, Lei yuan Li. Feature Extraction Research in Signal Process with Phi-OTDR Technology[J]. Journal of Beijing University of Posts and Telecom, 2017, 40(6): 109-114.

[8] Hao T, Xue Y, Zheng Z. Feature Extraction of Noisy Signals Based on Short-Time Fractional Fourier Transform. Journal of Naval Aeronautical and Astronautical University, vol. 22, no. 13, pp. 25-29, 2019.

[9] Zhang Y, Zhang J, Tao R. Key Frame Extraction of Surveillance Video Based on Fractional Fourier Transform. Journal of Beijing University of Technology: English, vol. 30, no. 3, pp. 11-18, 2021.

[10] Zhang Y, Zhang J, Tao R. Key Frame Extraction of Surveillance Video Based on Fractional Fourier Transform. Journal of Beijing Institute of Technology, vol. 30, no. 3, pp. 311-321, 2021.

[11] Shao Y, Di L U, Yang G X. Application of Fractional Fourier Transform in Fault Diagnostics of Rolling Bearing. Journal of Harbin University of Science and Technology, vol. 11, no. 7, pp. 8-14, 2017.

[12] Wang S, Guo Y, Yang L. Research on sparsity of frequency modulated signal in fractional Fourier transform domain. Editorial Office of Opto-Electronic Engineering, vol. 52, no. 11, pp. 20-32, 2020.

[13] Daulappa Guranna BHALKE, Betsy RAJESH, Dattatraya Shankar BORMANE. Automatic Genre Classification Using Fractional Fourier Transform Based Mel Frequency Cepstral Coefficient and Timbral Features. Archives of acoustics: journal of Polish Academy of Sciences, vol. 42, no. 2, pp. 213-222, 2017.