COMBINING EVIDENCE FROM AUDITORY, INSTANTANEOUS FREQUENCY AND RANDOM FOREST FOR ANTI-NOISE SPEECH RECOGNITION

Kun Liao

China Power Complete Equipment Co., Ltd, China

ABSTRACT

Due to the shortcomings of acoustic feature parameters in speech signals, and the limitations of existing acoustic features in characterizing the integrity of the speech information, This paper proposes a method for speech recognition combining cochlear feature and random forest. Environmental noise can pose a threat to the stable operation of current speech recognition systems. It is therefore essential to develop robust systems that are able to identify speech under low signal-to-noise ratio. In this paper, we propose a method of speech recognition combining spectral subtraction, auditory and energy features extraction. This method first extract novel auditory features based on cochlear filter cepstral coefficients (CFCC) and instantaneous frequency (IF), i.e., CFCCIF. Spectral subtraction is then introduced into the front end of feature extraction, and the extracted feature is called enhanced auditory features (EAF). An energy feature Teager energy operator (TEO) is also extracted, the combination of them is known as a fusion feature. Linear discriminate analysis (LDA) is then applied to feature selection and optimization of the fusion feature. Finally, random forest (RF) is used as the classifier in a non-specific persons, isolated words, and small-vocabulary speech recognition system. On the Korean isolated words database, the proposed features (i.e., EAF) after fusion with Teager energy features have shown strong robustness in the nosiy situation. Our experiments show that the optimization feature achieved in a speech recognition task display a high recognition rate and excellent anti-noise performance.

KEYWORDS

Cochlear filter cepstral coefficients; Teager energy features; Linear discriminate analysis; Random forest; speech recognition.

1. INTRODUCTION

Speech is the material shell and acoustic representation of language, and is one of the most direct, common and convenient carrier of information exchange for humans, and plays an important role in human-computer interaction and information transmission. With the advent of artificial intelligence (AI), it has always been the ideal of AI researchers to enable computers to simulate human consciousness and thinking information, so as to achieve human-computer interaction [1]. Speech recognition is equivalent to "machine's auditory system". It takes speech signal as the research object, combines signal processing technology and pattern recognition model to communicate with a computer, so that the computer can convert the speech signal into corresponding text or commands through the process of understanding and recognition [2]. Interpretation of human spoken language through technology has a diverse range of applications including in air transport, intelligent homes, disaster rescue, medical diagnostics, and other

David C. Wyld et al. (Eds): CSTY, SIGI, MaVaS - 2021
pp. 75-85, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.112207

human-computer interaction fields [3]. In the aviation transmission industry, speech recognition technology is used in the monitoring and response of the aviation air traffic control command, which makes the control command automatically respond, and then improves the training efficiency of the analog aircraft [4]. In smart home, combining speech recognition technology, wireless information transmission technology and embedded mobile computing technology, we can design a barrier-free intelligent home environment system for the disabled groups, so as to help them improve their quality of life and ability to participate in society [5]. In disaster rescue, speech recognition technology is integrated into the design of sensors to detect the calls of the victims, realize the auditory navigation of rescue robots, and achieve the purpose of search and rescue [6]. In medical diagnosis, according to the relationship between voice and depression, a depression recognition model based on voice variables is built to realize the correct diagnosis of the disease [7]. All the above studies are the practical application of speech recognition technology in social life, which shows that speech recognition technology has high research value and significance, and the improvement of the performance of speech recognition system has become a research hotspot of researchers.

As an important element of speech recognition, feature extraction has a large influence on the performance of the system [8]. Therefore, methods to extract the most information-capable, noise-less, easily classified and stable new features from speech must be developed, and integrate and optimize the different types of features that have been proposed also require further research to establish a speech recognition system with the best classification performance. Currently, the most mainstream speech feature is Mel frequency cepstral coefficient (MFCC), and MFCC is extracted based on Fourier transform, studies have shown that the Fourier transform is not suitable for the processing of non-stationary time-varying signals [9]. In recent years, a new method of auditory transformation compensates for the shortcomings of Fourier transform. The time-frequency analysis method of introducing auditory transformation into speech signal processing has attracted the attention of some researchers, and based on auditory transformation, a new feature model which is more in line with human ear auditory characteristics has been proposed. For instance, Peter Li try to extract cochlear filter cepstral coefficients (CFCC) features for robust speaker identification [10-11], and CFCC features show strong robustness in speech signal processing [12]. In [13], Yanyan proposed CFCCIF features based on Cochlear filter cepstral coefficients and instantaneous frequency, and combined CFCCIF features with Principal Component Analysis (PCA) for speech recognition. Although there are many applications of CFCC features, there are very few studies focused on the application of CFCCIF features to speech recognition in noisy environments. Hence, we attempt to apply CFCCIF features to noisy speech recognition systems. Spectral subtraction is then introduced in the front-end of CFCCIF features extraction to enhance features, so as to extract more robust feature parameters. And we combine enhanced CFCCIF and Teager energy features to form a fusion feature, and linear discriminate analysis (LDA) is used to optimize the fusion feature parameters to obtain the optimal feature to improve the recognition accuracy. Finally, random forest (RF) is used as the classifier in a speech recognition system.

2. PROPOSED CFCCIF FEATURES

2.1. Cochlear filter cepstral coefficients (CFCC)

The feature extraction procedure for CFCC consists of four parts: a series of cochlear filter banks model based on auditory transform, hair cell function, nonlinearity, and discrete cosine transform (DCT)[14]. The following subsection briefly describes auditory transform and procedure for estimating the CFCC features.

2.1.1. Auditory Transform

As a new method of processing non-linear signals, auditory transform is equivalent to converting time-domain signals into frequency-domain signals through cochlear filter banks. The cochlear filter function is used as the basis function of the wavelet, completing the whole process of sound transmission from the outer ear to the basement membrane, with an existing inverse transform [15].

Let $\boldsymbol{\psi(t)}$ be the impulse response of the basilar membrane of cochlear $\psi(t) \in L^2(R)$, in which the function $\psi(t)$ satisfies the following conditions:

①. It integrates to zero:

$$\int_{-\infty}^{+\infty} \psi(t) dt = 0 \tag{1}$$

②. It is square integrable or has finite energy:

$$\int_{-\infty}^{+\infty} \left| \psi(t) dt \right|^2 < \infty \tag{2}$$

③. It satisfies:

$$\int_{-\infty}^{+\infty} \frac{\left|\Psi(\omega)\right|^2}{\omega} d\omega = C \tag{3}$$

where $0 < C < \infty$, and

$$\Psi(\omega) = \int_{-\infty}^{+\infty} \psi(t) e^{-j\omega t} d\omega \tag{4}$$

Let $f(t)$ be any square integrable function. The auditory transform of $f(t)$, with respect to $\psi(t)$ as the impulse response of the basilar membrane in the cochlea, is defined as:

$$T(a,b) = \int_{-\infty}^{+\infty} f(t) \psi_{a,b}(t) dt \tag{5}$$

where $\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) dt$, a, b are real, and a is a scale or dilation variable. By changing a, the central frequency of an impulse response function can be shifted. Subscript b is a time shift or translation variable. If a is known, $\psi_{a,0}(t)$ moves a unit along the time axis to get $\psi_{a,b}(t)$. Note that $1/\sqrt{a}$ is an energy normalizing factor. It ensures that the energy stays the same for all a and b, providing:

$$\int_{-\infty}^{+\infty} \left|\psi_{a,b}(t)\right|^2 dt = \int_{-\infty}^{+\infty} \left|\psi(t)\right|^2 dt \tag{6}$$

A typical cochlear impulse response function or cochlear filter can be defined as:

$$\psi_{a,b}(t)=\frac{1}{\sqrt{a}}\frac{(t\text{-}b)^{\alpha}}{a}\exp\left[-2\pi f_L\beta\left(\frac{t\text{-}b}{a}\right)\right]\times\left[\cos 2\pi f_L\left(\frac{t\text{-}b}{a}\right)+\theta\right]u(t\text{-}b) \tag{7}$$

where $\alpha>0,\beta>0$, parameters α and β determine the frequency domain shape and width of the cochlear filter. Subscript α and β are taken as the generally empirical value, $\alpha=3,\beta=0.2$, $u(t)$ is the unit step function, and the value θ is the initial phase. The value of a can be determined by the current filter, the central frequency f_c, and the lowest frequency f_L of the cochlear filterbank, which is denoted as:

$$a=f_L/f_c \tag{8}$$

2.1.2. Other Operations in CFCC Extraction

As an important part of the auditory system, human cochlear inner ear hair cells transform the vibration signals transmitted from the basement membrane of the cochlea into analyzable nerve impulse signals of the brain, and then transmit them to the auditory nerve fibers[16]. The following nonlinear function of hair cell describes this motion:

$$h(a,b)=[T(a,b)]^2 \tag{9}$$

where $T(a,b)$ is the filterbank output of speech signal $f(t)$.

The hair cell output of each filterbank is converted into a representation of the nerve spike count density in a duration associated with the current band central frequency, which is computed as:

$$S(i,j)=\frac{1}{d}\sum_{b=1}^{l+d\text{-}1}h(i,b),l=1,L,2L\cdots;\forall i,j \tag{10}$$

where $d=max\{3.5\tau_i,20ms\}$ is the window length, τ_i is the period of the i band. $\tau_i=1/f_c$, and L is the window shift duration.

The output of the above formula is further applied to scales of loudness functions as logarithmic nonlinearity, providing:

$$y(i,j)=\log\left[S(i,j)\right] \tag{11}$$

Finally, discrete cosine transform (DCT) is applied to decorrelate the feature dimensions. It generates the cochlear filter cepstral coefficients as a new auditory-based speech feature, which is computed as:

$$cfcc(i,n)=\sqrt{2/M}\sum_{m=1}^{M-1}y(i,m)\cos\left(\frac{\pi n(m-1/2)}{M}\right),\quad 0\le m\le M \tag{12}$$

where M is the number of filters.

2.2. Instantaneous frequency (IF) estimation

Let $s(t)$ be speech signal, the IF of $s(t)$ is defined as the derivative of the unwrapped phase of the analytic signal derived from $s(t)$. For a real signal $s(t)$, its complex analytic representation is given by:

$$s_a(t) = s(t) + js_h(t) \tag{13}$$

where $s_h(t)$ is the Hilbert transform of $s(t)$, and the inverse Fourier transform(IFT) of $s_h(t)$ can be expressed as:

$$S_h(\omega) = \begin{cases} +jS(\omega) & \omega < 0 \\ -jS(\omega) & \omega > 0 \end{cases} \tag{14}$$

Then, the amplitude envelope of $s_a(t)$ is expressed as:

$$\left| s_a(t) \right| = \sqrt{s^2(t) + s_h^2(t)} \tag{15}$$

The instantaneous phase is $\phi(t) = \tan^{-1}\left(s_h(t)/s(t)\right)$, and IF is derived from unwrapped instantaneous phase, which can be expressed as:

$$IF = \frac{d}{dt}\left(\phi(t)\right) \tag{16}$$

2.3. CFCCIF Estimation

Similar to nerve spike density estimation, the IF is obtained as:

$$SIF(i,\mathrm{j}) = \frac{1}{d}\sum_{b=l}^{l+d-1} IF\left(h(i,b)\right), \qquad l = 1, L, 2L, \cdots; \forall i, j \tag{17}$$

To use both envelope structure and IF information, the IF features (eq.15) are multiplied with the corresponding nerve spike density envelope (eq.10). Thus, IF obtained in silence regions will be suppressed. To capture the transient information, the change in envelope and IF between consecutive frames is estimated through derivative operation followed by logarithm. Finally, DCT is applied framewise to get CFCCIF features.

3. PROPOSED EAF FEATURES

3.1. Spectral Subtraction

The principle of spectral subtraction is that the power spectrum of pure speech signal can be obtained by subtracting the power spectrum of noise from the power spectrum of speech signal with noise [17].

Let $y(n)$ be speech signal with noise, $s(n)$ is pure speech signal, $d(n)$ is noise, and the relationship between them is:

$$y(n)=s(n)+d(n),0\le n\le N\text{-}1 \tag{18}$$

where n the data points, and N is frame length.

Their representation in the Fourier transform domain is given by:

$$Y(\omega)=S(\omega)+D(\omega) \tag{19}$$

As speech is assumed to be uncorrelated with background noise, the short-term power spectrum of $y(n)$ has no cross-terms, hence:

$$E\left|\left|S(\omega)\right|^2\right|=E\left|\left|Y(\omega)\right|^2\right|-E\left|\left|D(\omega)\right|^2\right| \tag{20}$$

where $S(\omega)$, $D(\omega)$, $Y(\omega)$ is the short-term power spectrum of $s(n)$, $d(n)$, and $y(n)$.

For a short-time stationary process in a frame, use:

$$\left|S(\omega)\right|^2=\left|Y(\omega)\right|^2-\lambda_d(\omega) \tag{21}$$

in which $\lambda_d(\omega)$ is the statistical average of silent segment $\left|D(\omega)\right|^2$. Therefore, the amplitude of the speech signal after spectral subtraction can be expressed as:

$$\hat{S}(\omega)=[|Y(\omega)|^2\text{-}E(|D(\omega)|^2)]^{1/2}=[|Y(\omega)|^2\text{-}\lambda_d(\omega)]^{1/2} \tag{22}$$

3.2. EAF Features Extraction

To reduce the influence of noise on the CFCCIF features and further enhance the robustness of features, spectral subtraction is introduced in the front-end of CFCCIF features extraction. The speech signal is preprocessed first in a process which includes pre-emphasis, endpoint detection, and frame windowing. Formula (20) is then used to subtract the spectrum amplitude of noise from the spectrum amplitude of the noise signal, providing the spectrum amplitude of pure signal. Based on the phase insensitivity of speech, the phase angle information before spectral subtraction is directly used to reconstruct the signal after spectral subtraction to obtain the denoised speech. Finally, the denoised speech signal is extracted using process of CFCCIF features extraction, the enhanced auditory features (EAF)is obtained. The extraction process is illustrated in Figure. 1.

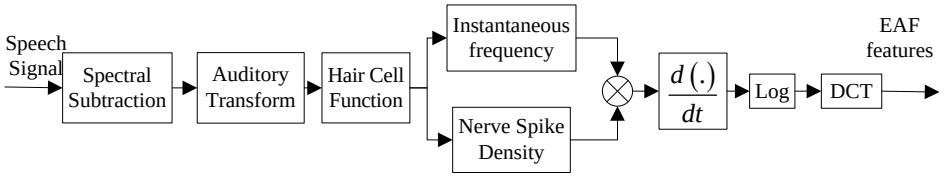

Figure 1. Extraction process

4. FUSION FEATURES EXTRACTION AND OPTIMIZATION

4.1. Teager Energy Features

Let $x(n)$ be a discrete-time signal, and the definition of TEO is:

$$\psi[x(n)] = x(n)^2 - x(n+1)x(n-1) \tag{23}$$

where $\psi[x(n)]$ is output of TEO, $x(n)$ is the sampling value of the discrete signal at n point.

Let $x(n)$ be a speech signal with additive noise, $s(n)$ be pure speech signal, and $\omega(n)$ be zero-mean additive noise. This relationship can be expressed as:

$$x(n) = s(n) + \omega(n) \tag{24}$$

The TEF of $x(n)$ is given by:

$$\psi[x(n)] = \psi[s(n)] + \psi[\omega(n)] + 2\tilde{\psi}[s(n), \omega(n)] \tag{25}$$

where $\tilde{\psi}[s(n), \omega(n)]$ is mutual Teager energy of $s(n)$ and $\omega(n)$, and

$$\tilde{\psi}[s(n), \omega(n)] = s(n)\omega(n) - 0.5s(n-1)\omega(n+1) - 0.5s(n+1)\omega(n-1) \tag{26}$$

Both $s(n)$ and $\omega(n)$ are zero mean and independent of each other, providing:

$$E\{\tilde{\psi}[s(n), \omega(n)]\} = 0 \tag{27}$$

$$E\{\psi[x(n)]\} = E\{\psi[s(n)]\} + E\{\psi[\omega(n)]\} \tag{28}$$

Compared with TEF of pure speech signal, the TEF of noise can be neglected, according to:

$$E\{\psi[x(n)]\} \approx E\{\psi[s(n)]\} \tag{29}$$

Thus, TEF can eliminate the influence of zero-mean noise and achieve speech enhancement [17-17]. The application of TEF in feature extraction can not only better track the non-linear energy of speech signal, can reasonably present the transformation of signal energy, but also suppress noise and enhance speech signals, achieving good results in speech recognition.

To construct a more effective subset of speech features, this paper combines above EAF features and TEF which reflects the change of signal energy. The fusion features not only represent the auditory perception characteristics and instantaneous frequency information of human ears, but also combine the characteristics of speech energy change, and suppress the zero-mean noise effect on speech signal to some extent, so as to more accurately describe the characteristics of speech.

4.2. Linear Discriminant Analysis (LDA)

To reduce the storage of feature data and further optimize the fusion feature, LDA is used to reduce the dimension of the fusion feature and further improve the performance of the recognition system. The principle of LDA is to project the high-dimensional pattern samples into the optimal discriminant vector space to extract classification information and compress the dimension of feature space. After projection, the pattern samples can have the largest class distance and the smallest intra-class distance in the new subspace, that is, the pattern has the best separability in the space [19-20]. It can transform the original feature set into a new feature subspace with a lower dimension, and compress the data while keeping as much relevant information as possible.

5. EXPERIMENTAL SETUP AND ANALYSIS OF RESULTS

The isolated words database is used for performing isolated word recognition from speech signals. The vocabulary sizes used here are 10 words and 20 words. The corpus consists of 10 digits and 40 command words. Random forest is used for speech recognition comparison experiments, and 10-fold cross-validation method is used to test the performance of feature and recognition network.

To verify the validity and robustness of the proposed fusion feature and optimization feature, and the optimized feature is defined as LDA-Features, the following experimental schemes are designed.

Table 1 Comparison of speech recognition based on five features (%)

Vocabulary	Experiment	Features	SNR (dB)					Average
			0	5	10	15	20	
10	Experiment 1	CFCCIF	80.79	86.63	87.12	90.41	90.96	87.18
	Experiment 2	EAF	82.94	87.38	89.08	91.00	91.80	88.44
	Experiment 3	EAF+TEF	83.48	88.02	89.08	91.80	92.06	88.89
	Experiment 4	LDA-Features	92.40	94.41	96.71	96.32	96.83	95.33
20	Experiment 1	CFCCIF	72.76	81.17	85.68	87.61	88.89	83.22
	Experiment 2	EAF	73.16	81.99	86.81	88.68	90.37	84.20
	Experiment 3	EAF+TEF	77.13	82.00	87.28	89.22	90.77	85.28
	Experiment 4	LDA-Features	90.30	94.20	94.27	96.05	95.55	94.07

(1) According to the recognition results of experiment 1 and experiment 2 shown in TABLE I, it can be observed that EAF features have a superior recognition effect compared with CFCCIF features. The average speech recognition rate under 10 words is increased by 1.26%, and under 20 words is increased by 0.98 %. This result proves that the feasibility and validity of the proposed features EAF in isolated word speech recognition systems. However, it can also been seen in TABLE I that the recognition effect of the CFCCIF features in a low SNR environment is not ideal.

(2) Comparing the recognition results of experiment 2 and experiment 3, it can be seen that after adding the TEF feature that embodies nonlinear energy characteristics, the recognition effect of the fusion feature is further improved compared to single feature. This result illustrates that TEF contains the effective information of speech signal and can be used as an auxiliary feature parameter to improve the performance of a speech recognition system.

(3) Comparison of experiment 2 and 3 confirms that after the fusion feature is optimized by LDA, the recognition rate have a significant improvement. This is because LDA can reduce dimension of feature while retaining important information in feature, thus improving the classification performance of speech recognition systems, and further verifying the effectiveness of fusion feature optimization.

(4) Figure. 2. depicts the speech recognition results of four groups of experiments in different SNR environments intuitively. It can be determined that fusion feature have certain advantages in recognition rate and robustness. It further illustrates the fusion feature can be constructed by combining auditory features with energy feature, and further verifying the effectiveness of LDA optimized feature method.

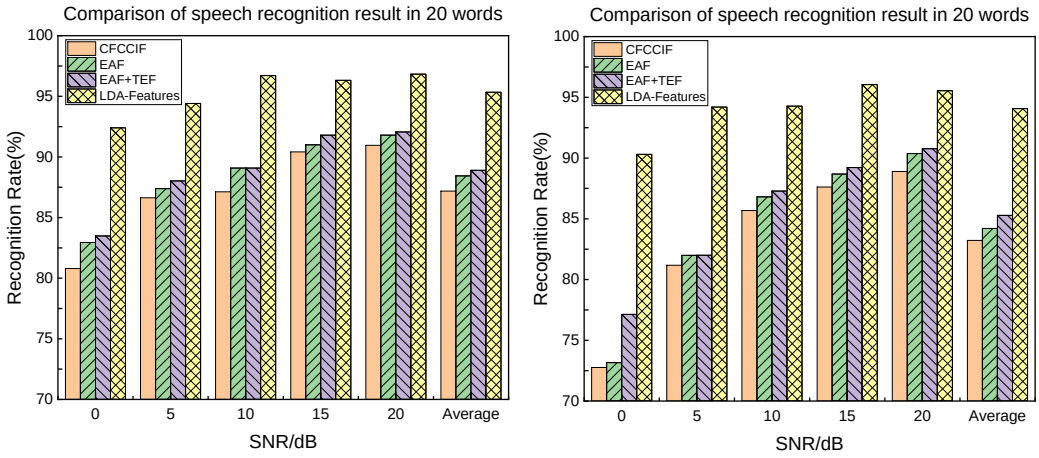

Figure. 2. Comparison of experimental results

6. CONCLUSION AND FURTHER STUDY

In this paper, CFCCIF features were extracted based on CFCC and instantaneous frequency information, and this feature was applied to speech recognition system in noisy environment. Spectral subtraction was then introduced into the front end of CFCCIF features extraction to improve the robustness of the feature to noise, and the extracted EAF features and TEF was combined in a fusion feature. It is proved that the fusion feature can effectively improve the recognition rate of speech recognition compared with the single feature. Finally, a feature optimization method of linear discriminate analysis was proposed, and its effectiveness was verified. In future research, we would like to consider finding a better speech enhancement method combined with feature extraction to achieve better speech recognition performance. In addition, the study of more better feature optimization are also the future research direction that needs to be further explored.

REFERENCES

[1] Ding I J, Yen C T . Enhancing GMM speaker identification by incorporating SVM speaker verification for intelligent web-based speech applications[J]. Multimedia Tools & Applications, 2015, 74 (14):5131-5140.

[2] Kłosowski, Piotr, Dustor A . Automatic Speech Segmentation for Automatic Speech Translation[M]// Computer Networks. Springer Berlin Heidelberg, 2013.

[3] Karpov A A, Yusupov R M . Multimodal Interfaces of Human–Computer Interaction[J]. Herald of the Russian Academy of Sciences, 2018, 88 (1):67-74.

[4] Cheung B . The resurgence of tactile display technologies.[J]. Aviation Space & Environmental Medicine, 2004, 75 (10):925.

[5] Han Y, Hyun J, Jeong T, et al. A smart home control system based on context and human speech[C]// 2016 18th International Conference on Advanced Communication Technology (ICACT). IEEE, 2016.

[6] Doostdar M, Schiffer S, Lakemeyer G . A robust speech recognition system for service-robotics applications.[J]. Lecture Notes in Computer Science, 2008, 5399:1-12.

[7] Lang H, Cui C . Automated Depression Analysis Using Convolutional Neural Networks from Speech[J]. Journal of Biomedical Informatics, 2018:S153204641830090X-.

[8] Wang G Y, Zhang Y M, Sun M L, et al. Speech signal feature parameters extraction algorithm based on PCNN for isolated word recognition[C]// International Conference on Audio. IEEE, 2017.

[9] Seyedin S, Gazor S, Ahadi S M . On the distribution of Mel-filtered log-spectrum of speech in additive noise[J]. Speech Communication, 2015, 67:8-25.

[10] Li Q L Q . An auditory-based transfrom for audio signal processing[C]// IEEE Workshop on Applications of Signal Processing to Audio & Acoustics. IEEE, 2009.

[11] Li Q, Huang Y. Robust speaker identification using an auditory-based feature[J]. IEEE International Conference on Acoustics Speech & Signal Processing, 2010, 23 (3):4514-4517.

[12] Li Z, Gao Y . Acoustic feature extraction method for robust speaker identification[M]. Kluwer Academic Publishers, 2016. 12 (75): 7391-7406.

[13] Yanyan, Shi, Jing, et al. Fusion Feature Extraction Based on Auditory and Energy for Noise-Robust Speech Recognition[J]. IEEE Access, 2019, 7:81911-81922.

[14] Wang D, Hansen J H L . F0 estimation for noisy speech by exploring temporal harmonic structures in local time frequency spectrum segment[C]// The 41st IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2016. 6510-6514.

[15] Li Q, Huang Y. An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification Under Mismatched Conditions[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, 19 (6):1791-1801.

[16] Haji T, Kitazawa S . Acoustic analysis of certain consonants using a computed model of the peripheral auditory system.[J]. Nippon Jibiinkoka Gakkai Kaiho, 1994, 97 (11):2055-2064.

[17] López-Oller, Domingo, Benamirouche N, Gomez A M, et al. Speech excitation signal recovering based on a novel error mitigation scheme under erasure channel conditions[J]. Speech Communication, 2018, 97:73-80.

[18] Kaiser J F. On a simple algorithm to calculate the 'energy' of a signal[C]// International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2002:381-384 vol.1.

[19] Senior A, Cho Y, Weston J . Learning improved linear transforms for speech recognition[C]// IEEE International Conference on Acoustics. IEEE, 2012.

[20] Gharsellaoui S, Selouani S A, Dahmane A O . Automatic emotion recognition using auditory and prosodic indicative features[C]// Electrical & Computer Engineering. IEEE, 2015.

AUTHORS

Short Biography

Kun Liao, Graduated from Northeast Electric Power University, research on Pattern Recognition, Artificial intelligence, image processing and speech recognition.