Journal of Southeast University (English Edition) Vol. 37, No. 4, pp. 343 - 349 Dec. 2021 ISSN 1003—7985

Speech detection method based on a multi-window analysis

Luo Xinwei[1] Liu Ting[1] Huang Ming[1] Xu Xiaogang[1]
Cao Hongli[1] Bai Xianghua[2] Xu Dayong[2]

([1]Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China)
([2]The 1st Military Representative Office in Nanjing, Shanghai Bureau, PLA Navy Military Equipment Department, Nanjing 210000, China)

Abstract: Aiming at the poor performance of speech signal detection at low signal-to-noise ratio (SNR), a method is proposed to detect active speech frames based on multi-window time-frequency (T-F) diagrams. First, the T-F diagram of the signal is calculated based on a multi-window T-F analysis, and a speech test statistic is constructed based on the characteristic difference between the signal and background noise. Second, the dynamic double-threshold processing is used for preliminary detection, and then the global double-threshold value is obtained using K-means clustering. Finally, the detection results are obtained by sequential decision. The experimental results show that the overall performance of the method is better than that of traditional methods under various SNR conditions and background noises. This method also has the advantages of low complexity, strong robustness, and adaptability to multi-national languages.

Key words: voice activity detection; multi-window spectral analysis; K-means clustering; threshold adjustment; sequential decision

DOI: 10.3969/j.issn.1003-7985.2021.04.001

The purpose of voice activity detection (VAD) is to detect whether the current acoustic signal contains a voice signal and identify voice segments in it. VAD distinguishes voice signals from various background noise signals for subsequent processing. VAD can be applied on many occasions, such as speech recognition system[1-2]; signal noise reduction and speech enhancement[3-4], speech signal extraction and recognition in an interference environment[5], acoustic scene analysis[6], and data preprocessing of speech database sample construction[7]. Because acoustic signals are changeable and complex, it is difficult to find stable features for effective detection, especially under the condition of a low signal-to-noise ratio (SNR).

Received 2021-04-29, **Revised** 2021-08-20.

Biography: Luo Xinwei (1978—), male, doctor, associate professor, luoxinwei@seu.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 12174053, 91938203, 11674057, 11874109), the Fundamental Research Funds for the Central Universities (No. 2242021k30019).

Citation: Luo Xinwei, Liu Ting, Huang Ming, et al. Speech detection method based on a multi-window analysis[J]. Journal of Southeast University (English Edition), 2021, 37(4): 343 - 349. DOI: 10.3969/j.issn.1003-7985.2021.04.001.

In recent years, research in the fields of acoustic signal feature extraction and detector design has been extended to improve the performance of VAD. The feature extraction of speech signals has been deeply studied. The features of speech signals are usually extracted in the time-frequency (T-F) domain, including short-time energy with a zero-crossing rate[8], higher-order statistics[9], and autocorrelation[10]. More refined features are proposed based on these frequency features, such as Mel-frequency cepstral coefficients[11-12]. Ref. [13] proposed a wavelet subband-based VAD algorithm. Refs. [14-15] proposed long-term spectral variability (LTSV) and multi-band LTSV as features for VAD. In addition, other methods have been proposed to extract features, including multi-resolution cochleagram[16] and single-frequency filtering (SFF)[17-18]. As for classifier design, some innovations have been proposed based on the statistical data of speech and noise[19]. Supervised learning systems, such as support vector machines[20] and neural network algorithms[21], can improve the performance of statistical methods using labeled training data. Ref. [19] proposed a semi-supervised learning method that uses a noise model derived from training data in the initialization process. Ref. [22] pointed out that it is better to use unsupervised classification without training data in the practical applications of VAD. Although new techniques, such as neural networks[21] and deep neural networks[23-24], are introduced into the classification system and usually have good results in the pre-specified noise conditions and corpus, the performance cannot be guaranteed when the environmental noise changes and new speech signals appear. Meanwhile, the performance greatly depends on its parameter values and hyperparameter (e.g., number of layers, number of neurons, and coefficient values) settings[25-26].

Therefore, VAD applications still need a simple and efficient method. According to the above research, most speech signal detection features are constructed in the T-F domain. T-F features have been proven effective, but traditional T-F feature extraction methods have some problems, such as insensitivity to speech, poor resolution, and poor adaptability to different environmental noise[17]. To improve the detection ability of the signal envelope and harmonic structure, based on the traditional T-F anal-

ysis, a multi-window spectrum analysis method based on Hamming windows was designed to improve the signal resolution in the T-F domain. Based on the T-F diagram, envelope detection statistics are constructed, and a secondary detection framework is designed to realize VAD.

1 Feature Extraction of Speech Signals

Generally, formants are important features of speech signals. In the vowels excited by the human vocal cords, the sound energy is highly concentrated in specific spectral bands, thus forming spectral peaks. At a low SNR, the characteristics of energy aggregation in the frequency domain are still good detection features for vowels.

To transform the signal from the time domain to the T-F domain, short-time Fourier transform (STFT) is the most commonly used method. The traditional STFT processing has some shortcomings in detecting time-varying features. In this section, the traditional STFT will be improved to enhance the T-F analysis capability for speech signals.

1.1 Time-frequency transform of signals

For signal $x(t)$ and time window $w(t)$, the STFT is defined as

$$\mathrm{STFT}(t,f) = \int_{-\infty}^{\infty} [x(\tau) w^*(\tau - t)] \mathrm{e}^{-\mathrm{j}2\pi f\tau} \mathrm{d}\tau \quad (1)$$

The time window $w(t)$ has a decisive influence on the T-F resolution of the STFT analysis. To obtain the high-resolution energy distribution, it is necessary to find a window with good energy concentration in the T-F plane. This energy concentration is constrained by the Heisenberg-Gabor uncertainty principle, which states that for a given signal, the product of its time width and bandwidth is a constant.

Time width Δ_t and frequency width Δ_Ω are parameters used to describe the energy concentration characteristics of a window. It is not easy to analytically calculate the corresponding Δ_t and Δ_Ω for various windows commonly used in digital signal processing. For the convenience of calculation, the effective time width Δ_t of a signal can be defined as the time from the signal center as the symmetry center to both sides until it contains 80% of the energy area, so is the effective frequency width Δ_Ω. The analysis results of the different windows are shown in Tab. 1.

Tab. 1 Performance of various windows

Window function	Δ_t/s	Δ_Ω/Hz	$\Delta_t\Delta_\Omega$/(s·Hz)
Rectangular	1	1.14	1.140
Hanning	0.380	1.52	0.578
Hamming	0.404	1.26	0.549
Blackman	0.430	1.42	0.611
Gaussian ($\alpha=1$)	1.794	0.32	0.574
Gaussian ($\alpha=5$)	0.812	0.66	0.536
Gaussian* ($\alpha=1$)	1	1.08	1.080
Gaussian* ($\alpha=5$)	1	1.12	1.120

Note: * refers to truncated Gaussian.

The following conclusions can be drawn from Tab. 1: 1) Although the time-bandwidth product of the Gaussian window can reach the minimum value, the time-bandwidth product of truncated Gaussian windows in practice is much larger than that of ideal Gaussian windows. 2) Hanning, Hamming, and Blackman windows have similar performance. Among these windows, the Hamming window has the minimum time-bandwidth product, which means that the Hamming window has the best energy concentration characteristics in practical applications.

The resolution of the traditional STFT is fixed. The multi-window spectrum analysis (MWSA) is an effective method used to improve the adaptability of the T-F transforms to various time-varying signals in the T-F domain. MWSA is an extension of windowed spectrum analysis methods, which multiplies signals by a set of window sequences with different time widths. The MWSA produces several spectra, each with a different T-F resolution. For each point in the T-F domain, the best spectrum from the whole spectrum set can be chosen, or several spectra can be combined into one spectrum to get a better estimation of the signal in the T-F domain. In MWSA, the DFT of each data frame is replaced by the following formula:

$$P(k) = \sum_{i=1}^{H} \alpha_i \left| \sum_{n=-N/2}^{N/2-1} x(n) h_i(n) \mathrm{e}^{-\mathrm{j}2\pi \frac{n}{N} k} \right|^2 \quad (2)$$

where H is the number of window functions; N is the number of signal points, and N takes the even number; $h_i(n)$ is the i-th window function. The selection of $h_i(n)$ is quite flexible. In this study, to achieve the observation of multiple T-F resolutions and to make the data used in multiple windows align in the time domain, a group of windows is constructed based on the Hamming window. The formula of $h_i(n)$ is as follows:

$$h_i(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi i(n - N/2)}{N/i - 1}\right) \\ \qquad -\dfrac{N}{2} \leqslant n - \dfrac{N}{i} < \dfrac{N}{2} \\ 0 \qquad \text{otherwise} \end{cases} \quad (3)$$

where α_i is the weighted coefficient for each window function used to reduce the influence of different lengths of $h_i(n)$.

$$\alpha_i = \frac{1}{H} \frac{\sum_{n=-N/2}^{N/2-1} h_0(n)}{\sum_{n=-N/2}^{N/2-1} h_i(n)} \qquad i = 1,2,\cdots,H \quad (4)$$

From Eqs. (2) to (4), the spectrum estimation of a single frame signal can be obtained. In this study, the weighted average result is selected as the final spectrum, which has less noise fluctuation and good adaptability of time-varying signals. Hence, the MW_STFT transformation is defined as

$$\mathrm{MW\_STFT}(n,k) = \sum_{i=1}^{H} \alpha_i \left| \sum_{m=n-N/2}^{n+N/2-1} x(m) h_i(m-n) \mathrm{e}^{-\mathrm{j}2\pi \frac{m}{N} k} \right|^2 \quad (5)$$

Considering the time variation of speech signals, the MW_STFT with three windows is selected. The time lengths of the window functions h_0 to h_3 are 16, 64, and 512 ms. Fig. 1 shows the T-F diagrams of a speech signal (SNR = −6 dB) by the STFT with different window lengths and by MW_STFT. Compared with the traditional STFT, the harmonic structure of vowel signals is clearer, and the fluctuation of the background noise is smaller in the T-F diagram obtained by MW_STFT processing. The speech signal can also be identified easily from the T-F diagram by MW_STFT.

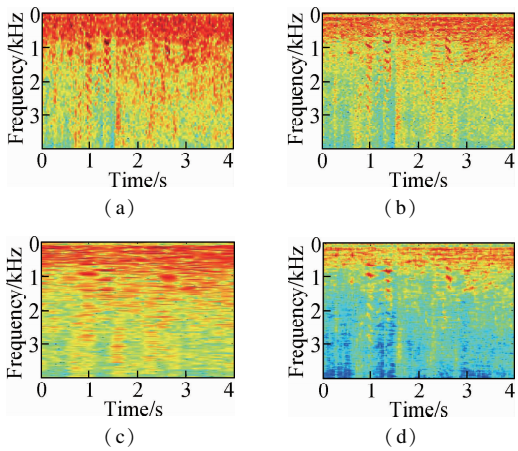

Fig. 1 The T-F diagram of speech signal (SNR = −6 dB) by windows with different lengths and MW_STFT. (a) STFT by h_1 (T = 16 ms); (b) STFT by h_2 (T = 64 ms); (c) STFT by h_3 (T = 512 ms); (d) MW_STFT

1.2 Construction of speech features

Usually, noise signals are colored. If energy is directly used as the test statistic, then it is often difficult to find speech from the signal due to the strong noise at low frequencies. In the T-F diagram of the signal, noise provides a floor envelope at each frequency. To compensate for the influence of the noise intensity, the floor value is used to calculate the weight value at each frequency. For $\mathrm{MW\_STFT}(n,k)$, the reciprocal μ_k of the mean value of the lower 15% data of the envelope value at each frequency index k is calculated. These μ_k are used to calculate the normalized weight values of each frequency. The choice of 15% is based on the assumption that there is at least 15% silence in speech. The weight of each frequency is given by

$$w_n(k) = \frac{\mu_k}{\sum_{i=1}^{K} \mu_i} \quad (6)$$

where K is the number of channels. The window length of the MW_STFT analysis for 16 kHz sampled signals is 2 048 points, the corresponding channel number K is 1 025, and the frequency resolution is 7.8 Hz. Now, the modified T-F diagram $\mathrm{MW\_STFT}_1(n,k)$ can be obtained by the weight value $w_n(k)$.

$$\mathrm{MW\_STFT}_1(n,k) = \mathrm{MW\_STFT}(n,k) w_n(k) \quad (7)$$

Using this weighting process, the noise level of each frequency is adjusted to approach a similar level.

To highlight the components of the speech signal, the weight of each frequency is calculated again using the ceiling value. For $\mathrm{MW\_STFT}_1(n,k)$, the average value of the top 20% data with higher envelope values on each frequency k is used to calculate the speech enhancement weight value. Similarly, the choice of a value of 20% is based on at least 20% of the phonetic assumptions in the language. The normalized enhancement weight and the modified T-F diagram are given by

$$w_s(k) = \frac{\rho_k}{\sum_{i=1}^{K} \rho_i} \quad (8)$$

$$\mathrm{MW\_STFT}_2(n,k) = \mathrm{MW\_STFT}_1(n,k) w_s(k) \quad (9)$$

Using this weighting process, the signal components of each frequency are enhanced. Now, the weighted energy can be constructed by $\mathrm{MW\_STFT}_2(n,k)$. For speech signals, the part below the speech frequency band can be ignored, which means that the signal components significantly lower than the speech frequency should not be included in the calculation of energy. Considering that the signal composed of speech and background noise has a wide dynamic range, the speech test statistics $S(n)$ based on the weighted energy are defined as

$$S(n) = \lg\left(1 + \sum_{k=k_0}^{k_1} \mathrm{MW\_STFT}_2(n,k)\right) \quad (10)$$

where k_0 and k_1 correspond to the indices of the cut-off frequency, which are set to 8 and 513, respectively, in this study, corresponding to the frequencies of 60 and 4 kHz. Through a logarithmic operation, the signal with a large dynamic range can be compressed, which is more convenient for subsequent detection processing.

2 Framework and Implementation of the Speech Detection Method

In general speech detection, signals are required to start with a segment of pure background noise, which is used to estimate the background noise characteristics of the whole signal. The performance of this kind of method is closely related to the selection of the segment of pure background noise. To automatically adapt to the environmental noise, a secondary processing structure is adopted in the detection framework, which can efficiently realize VAD using dynamic double thresholds. The framework of

the proposed VAD is shown in Fig. 2. In the preliminary detection, there is no guidance of labeled samples. However, the detection system lacks an understanding of the overall distribution of data, so it is difficult to select accurate thresholds. Therefore, the high threshold T_1 and low threshold T_2 are adopted. The initial value of T_1 and T_2 can be set arbitrarily, such as taking the first value of $S(n)$ in Eq. (10), because the initial value has little effect on the results due to the existence of a dynamic update mechanism. The update methods of T_1 and T_2 are as follows:

$$\left.\begin{aligned} T_1 &= \alpha T_1 + S(n)(1-\alpha) \\ T_2 &= \alpha T_2 + S(n)(1-\alpha) \end{aligned}\right\} \tag{11}$$

where α is a constant. The parameter α corresponds to the speed at which the parameter adapts to the change of noise. The smaller the value of α, the faster the change and the greater the fluctuation of T_1 and T_2. In this study, $\alpha = 0.95$.

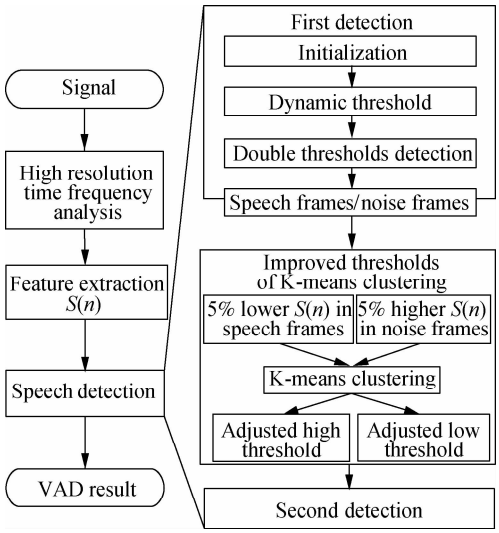

Fig. 2 Block diagram of the proposed VAD

After the preliminary detection, we obtained the labels (speech or non-speech frames) of the data. Then, according to the experience, 5% lower features from speech frames and 5% higher features from non-speech frames are selected, and K-means clustering is used to obtain more accurate double thresholds for speech detection.

Sequential detection is used in this study to make the decision process robust. When the dynamic thresholds T_1 and T_2 are ready, a second detection is performed. The processing flow is shown in Fig. 3.

3 Experiments and Discussions

In this experiment, the TIMIT test corpus is used for evaluation. Four kinds of noises, namely, Factory1, Volvo, white, and babble, are selected from the NOISEX-92 database. Volvo noise has the least interference on speech signals, and the other three kinds of noise have the greatest

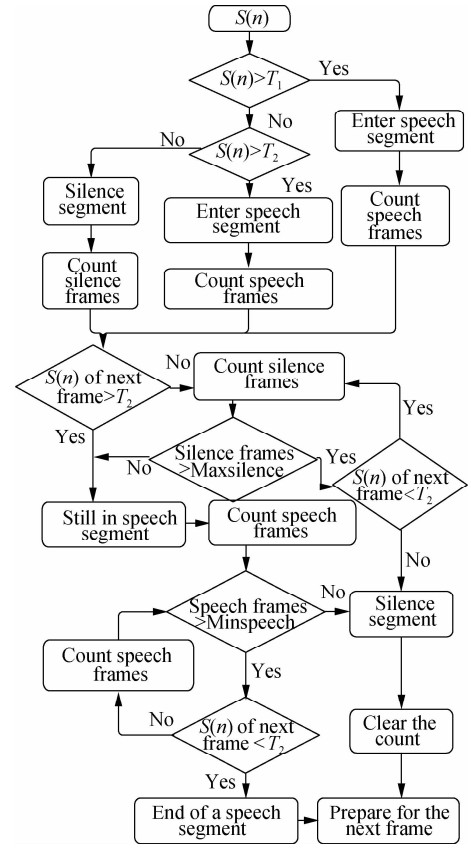

Fig. 3 Decision process of the proposed method

interference on clean speech signals[23], which are selected to test the performance of the proposed method in adverse environments. These noises are added to clean speech signals to generate degraded speech signals. The range of the SNR in this study is from −20 dB to 5 dB.

Fig. 4 shows the detection results of the proposed method on the voice signal in the different noise backgrounds at SNR = 5 dB and −10 dB. As shown in Fig. 4, when SNR = 5 dB, the proposed method has good results for various noises. When SNR = −10 dB, the proposed method maintains good performance against the Volvo noise interference and can effectively detect the other three kinds of noise interference, but the performance significantly decreases.

The same evaluation parameters as those in Ref. [18] are used for comparison with the SFF method: detection accuracy, FEC (missed detection rate in front-end clipping) and MSC (missed detection rate in mid-speech clipping), OVER (false alarm rate in carry-over of voice), NDS (percentage of noise detected as speech). FEC and MSC are true rejections, whereas OVER and NDS are false acceptance. Tab. 2 shows the experimental results of the proposed method under different kinds of noises at five SNR levels.

As shown in Tab. 2, with the decrease in the SNR, the detection accuracy of the proposed method also decreases. At a high SNR, the detection accuracy under various noise

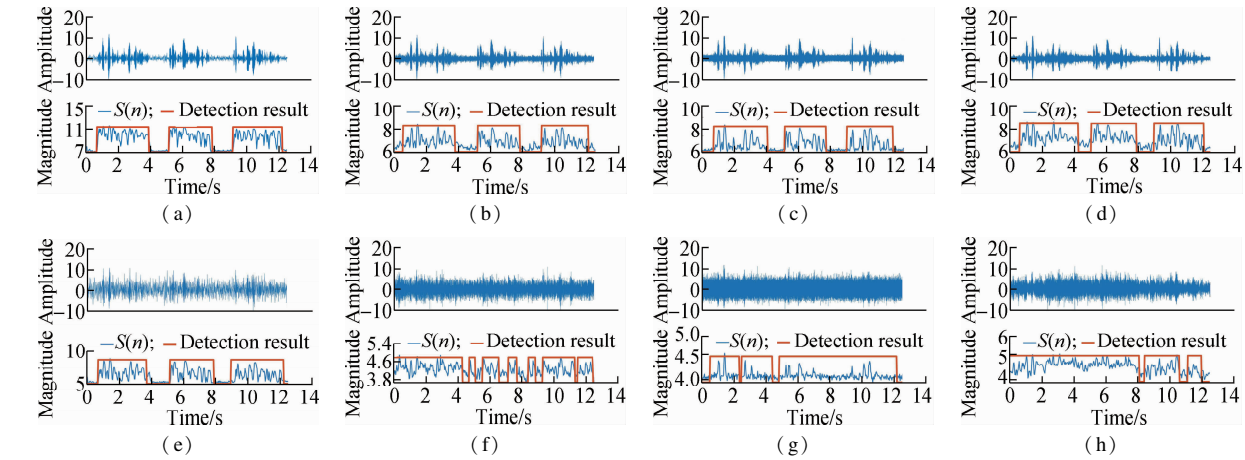

Fig. 4 Results of the proposed method in different noise backgrounds and SNRs. (a) Volvo noise and SNR = 5 dB; (b) Factory1 noise and SNR = 5 dB; (c) White noise and SNR = 5 dB; (d) Babble noise and SNR = 5 dB; (e) Volvo noise and SNR = −10 dB; (f) Factory1 noise and SNR = −10 dB; (g) White noise and SNR = −10 dB; (h) Babble noise and SNR = −10 dB

Tab. 2 Experimental results of the proposed method under different kinds of noises at five SNR levels %

Noise	SNR/dB	Detection accurary	FEC	MSC	OVER	NDS
Volvo	5	97.7	0.3	1.3	0.8	0.0
	0	98.2	1.3	0.0	0.5	0.0
	−5	96.2	0.3	3.1	0.0	0.0
	−10	94.1	0.3	2.8	2.3	0.0
	−20	92.8	1.5	1.3	0.0	1.0
White	5	94.1	1.0	0.8	0.3	1.0
	0	92.1	0.5	4.6	2.3	0.5
	−5	89.2	0.5	1.5	3.3	5.4
	−10	81.8	2.3	1.0	3.1	11.8
	−20	75.9	0.0	0.0	5.1	18.9
Factory1	5	94.9	1.5	2.3	1.3	0.0
	0	87.4	0.5	4.9	5.9	1.3
	−5	82.8	1.3	0.0	5.4	9.5
	−10	72.6	7.4	9.2	3.8	6.7
	−20	54.4	2.8	27.9	2.3	10.0
Babble	5	94.6	1.5	0.8	1.3	1.5
	0	90.8	0.5	2.1	3.6	3.1
	−5	78.0	1.5	0.0	9.7	10.8
	−10	72.6	1.5	13.1	5.4	8.5
	−20	66.7	3.6	5.4	5.4	18.9

backgrounds exceeds 90%. With the decrease in the SNR, the babble noise and factory noise have the greatest impact on speech. The detection accuracy under the condition of SNR = −20 dB is 66.67% and 54.36%. Under the background of the white noise, the detection accuracy is 75.90% when the SNR is −20 dB. The energy of the Volvo noise is mainly concentrated in the low frequency and has little impact on speech components. Under the condition of −20 dB SNR, the detection accuracy is 92.82%.

A comparison of the results of the proposed method, SFF method, and adaptive multi-rate2 (AMR2) method under different noise conditions at SNR = 5 and −10 dB is shown in Tab. 3.

Tab. 3 shows that under the conditions of the Volvo background noise (SNR = −10 dB) and white noise (SNR = 5 dB), the SFF and the proposed method obtain

Tab. 3 Comparison of the proposed method, SFF, and AMR2 method under different kinds of noises %

Noise	SNR/dB	Method	Detection accurary	FEC	MSC	OVER	NDS
Babble	−10	MSTFT	72.6	1.5	13.1	5.4	8.5
		SFF	67.7	0.1	12.1	0.1	20.0
		AMR2	61.7	0.1	13.1	0.1	25.0
Babble	5	MSTFT	94.6	1.5	0.8	1.3	1.5
		SFF	93.3	0.0	2.6	0.1	4.0
		AMR2	72.4	0.0	0.5	0.1	26.8
Factory1	−10	MSTFT	72.6	7.4	9.2	3.9	6.7
		SFF	67.6	0.1	13.5	0.1	18.7
		AMR2	58.8	0.1	17.4	0.1	23.6
Factory1	5	MSTFT	94.9	1.5	2.3	1.3	0.0
		SFF	91.7	0.0	1.9	0.1	6.3
		AMR2	74.1	0.0	1.4	0.1	24.2
Volvo	−10	MSTFT	94.1	0.3	2.8	2.3	0.0
		SFF	98.0	0.0	0.5	0.1	1.3
		AMR2	95.9	0.0	0.2	0.1	3.6
Volvo	5	MSTFT	97.7	0.3	1.3	0.8	0.0
		SFF	96.4	0.0	2.4	0.1	1.0
		AMR2	94.4	0.0	0.5	0.1	4.9
White	−10	MSTFT	81.8	2.3	1.0	3.1	11.8
		SFF	77.6	0.1	21.9	0.0	0.2
		AMR2	63.2	0.1	34.3	0.0	2.2
White	5	MSTFT	94.1	1.0	0.8	0.3	1.0
		SFF	97.0	0.1	1.8	0.1	1.0
		AMR2	87.5	0.1	8.6	0.1	3.7

nearly 95% detection accuracy, but the SFF shows better performance. Under the two conditions, the speech components in the signal are not significantly disturbed, and various VAD methods can obtain good results. In six of the eight cases, the proposed method shows advantages. Particularly, in the low SNR signals with babble, Factory1, and white noises, the proposed method improves the accuracy by more than 4% as compared with the other methods, which shows that the proposed method has better performance under the condition of a low SNR.

The proposed algorithm is programmed by MATLAB 2020a and runs on the workstation with an 8-core CPU (I7 9700K) and 16 GB RAM. It takes an average time

of 150 ms to process 12 s long data. This processing efficiency is acceptable for general tasks.

To illustrate the adaptability of the proposed method to different speech signals, speech signals from speech corpora of different languages are selected, including Mandarin, Arabic, Japanese, Russian, and Portuguese. Four sentences are randomly selected and spliced into a long speech signal. The performance is shown in Tab. 4.

Tab. 4 Comparison of the accuracy of the proposed method for different languages under the same conditions as those in the TIMIT corpus %

Noise	SNR/dB	Mandarin	Arabic	Japan	Russian	Portuguese
Volvo	5	98.0	97.7	96.1	95.5	98.0
	0	96.7	96.1	96.5	95.8	95.7
	−5	93.6	94.6	95.0	90.1	95.3
	−10	87.2	94.6	96.3	89.5	93.5
White	5	93.6	97.6	90.2	94.6	92.2
	0	92.3	97.1	87.7	92.4	85.4
	−5	88.5	96.9	80.9	87.7	78.0
	−10	82.1	90.2	75.9	85.8	70.2
Factory1	5	93.1	96.6	92.8	95.1	90.6
	0	80.5	89.9	82.3	86.4	75.6
	−5	74.4	70.2	72.9	73.1	66.4
	−10	62.8	64.1	63.9	65.9	53.8
Babble	5	93.1	92.5	88.2	91.0	91.0
	0	84.9	75.7	81.0	79.3	80.3
	−5	70.8	70.6	72.0	65.1	77.6
	−10	63.3	69.3	71.1	61.7	74.9

For the other five languages, the performance of the proposed method is comparable to that of English speech detection. The performance is good at a high SNR. More specifically, when the SNR is 5 dB, the detection accuracy is greater than 90%, regardless of the noise and language. In this case, the noise has no great influence on the characteristics of speech. With the decrease in the SNR, the performance of this method for all languages decreases. In the environment of −5 dB white noise, the proposed method maintains a correct detection rate of more than 80% for all languages, which is acceptable. Further analysis of the results shows that the detection accuracy of the proposed method is close to or greater than 90% for Mandarin, Arabic, and Russian, and approximately 80% for Japanese and Portuguese. This result is attributed to the frequency distribution characteristics of the different languages.

4 Conclusions

1) The proposed MW_STFT method can obtain a better T-F diagram than the traditional STFT through the multi-window STFT analysis, which is helpful for improving the detection ability of characteristic signals.

2) The use of the difference between signal and background noise to construct speech detection statistics can effectively improve the speech signal detection ability under colored noise.

3) The proposed detection method improves the performance and robustness of the detection system through a multi-window analysis and dynamic double-threshold processing. Experimental results show that the performance of the proposed method is generally better than that of traditional methods.

References

[1] Lamel L, Rabiner L, Rosenberg A, et al. An improved endpoint detector for isolated word recognition[J]. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1981, **29**(4): 777 − 785. DOI: 10.1109/TASSP.1981.1163642.

[2] Lu L, Jiang H, Zhang H J. A robust audio classification and segmentation method[C]//*Proceedings of the Ninth ACM International Conference on Multimedia*. Ottawa, Canada, 2001: 203 − 211. DOI: 10.1145/500141.500173.

[3] Song J L, Meng Y, Cao J M, et al. Research on digital hearing aid speech enhancement algorithm [C]//2018 37*th Chinese Control Conference* (*CCC*). Wuhan, 2018: 4316 − 4320. DOI: 10.23919/chicc.2018.8482732.

[4] Çolak R, Akdeniz R. A novel voice activity detection for multi-channel noise reduction[C]//*IEEE Access*. IEEE, 2021: 91017 − 91026.

[5] Jaiswal R. Speech activity detection under adverse noisy conditions at low SNRs[C]//2021 6*th International Conference on Communication and Electronics Systems* (*ICCES*). Coimbatre, India, 2021: 97 − 101. DOI: 10.1109/ICCES51350.2021.9488934.

[6] Masumura R, Matsui K, Koizumi Y, et al. Context-aware neural voice activity detection using auxiliary networks for phoneme recognition, speech enhancement and acoustic scene classification[C]//2019 27*th European Signal Processing Conference* (*EUSIPCO*). Coruna, Spain, 2019: 1 − 5. DOI: 10.23919/EUSIPCO.2019.8902703.

[7] Moldovan A, Stan A, Giurgiu M. Improving sentence-level alignment of speech with imperfect transcripts using utterance concatenation and VAD[C]//2016 *IEEE* 12*th International Conference on Intelligent Computer Communication and Processing* (*ICCP*). Cluj-Napoca, Romania, 2016: 171 − 174. DOI: 10.1109/ICCP.2016.7737141.

[8] Rabiner L R, Sambur M R. An algorithm for determining the endpoints of isolated utterances[J]. *The Bell System Technical Journal*, 1975, **54**(2): 297 − 315. DOI: 10.1002/j.1538-7305.1975.tb02840.x.

[9] Nemer E, Goubran R, Mahmoud S. Robust voice activity detection using higher-order statistics in the LPC residual domain[J]. *IEEE Transactions on Speech and Audio Processing*, 2001, **9**(3): 217 − 231. DOI: 10.1109/89.905996.

[10] Marzinzik M, Kollmeier B. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics[J]. *IEEE Transactions on Speech and Audio Processing*, 2002, **10**(2): 109 − 118. DOI: 10.1109/89.985548.

[11] Shi L, Ahmad I, He Y J, et al. Hidden Markov model

based drone sound recognition using MFCC technique in practical noisy environments[J]. *Journal of Communications and Networks*, 2018, **20**(5): 509 - 518. DOI:10.1109/JCN.2018.000075.

[12] Al-Ali A K H, Dean D, Senadji B, et al. Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions [J]. *IEEE Access*, 2017, **5**: 15400 - 15413. DOI:10.1109/ACCESS.2017.2728801.

[13] Pham T V, Stark M, Rank E. Performance analysis of wavelet subband based voice activity detection in cocktail party environment[C]//*The* 2010 *International Conference on Advanced Technologies for Communications*. Ho Chi Minh City, Vietnam, 2010: 85 - 88. DOI:10.1109/ATC.2010.5672718.

[14] Ghosh P K, Tsiartas A, Narayanan S. Robust voice activity detection using long-term signal variability [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, **19**(3): 600 - 613. DOI:10.1109/TASL.2010.2052803.

[15] Tsiartas A, Chaspari T, Katsamanis N, et al. Multi-band long-term signal variability features for robust voice activity detection[C]//14*th Annual Conference of the International Speech Communication Association* (*INTERSPEECH* 2013). ISCA, 2013: 718 - 722. DOI: 10.21437/interspeech.2013-201.

[16] Haider F, Luz S. Attitude recognition using multi-resolution cochleagram features[C]//2019 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). Brighton, UK, 2019: 3737 - 3741. DOI: 10.1109/ICASSP.2019.8682974.

[17] Aneeja G, Yegnanarayana B. Single frequency filtering approach for discriminating speech and nonspeech[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **23**(4): 705 - 717. DOI:10.1109/TASLP.2015.2404035.

[18] Makowski R, Hossa R. Voice activity detection with quasi-quadrature filters and GMM decomposition for speech and noise[J]. *Applied Acoustics*, 2020, **166**: 107344. DOI:10.1016/j.apacoust.2020.107344.

[19] Sohn J, Kim N S, Sung W. A statistical model-based voice activity detection[J]. *IEEE Signal Processing Letters*, 1999, **6**(1): 1 - 3. DOI:10.1109/97.736233.

[20] Dey J, Bin Hossain M S, Haque M A. An ensemble SVM-based approach for voice activity detection[C]// 2018 10*th International Conference on Electrical and Computer Engineering* (*ICECE*). Dhaka, Bangladesh, 2018: 297 - 300. DOI:10.1109/ICECE.2018.8636745.

[21] Krishnakumar H, Williamson D S. A comparison of boosted deep neural networks for voice activity detection[C]// 2019 *IEEE Global Conference on Signal and Information Processing* (*GlobalSIP*). Ottawa, ON, Canada, 2019: 1 - 5. DOI:10.1109/GlobalSIP45357.2019.8969258.

[22] Germain F G, Sun D L, Mysore G J. Speaker and noise independent voice activity detection [C]//*Interspeech* 2013. France, 2013: 732 - 736. DOI:10.21437/interspeech.2013 - 204.

[23] Tachioka Y. DNN-based voice activity detection using auxiliary speech models in noisy environments[C]//2018 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). Calgary, AB, Canada, 2018: 5529 - 5533. DOI:10.1109/ICASSP.2018.8461551.

[24] Paseddula C, Gangashetty S V. DNN based acoustic scene classification using score fusion of MFCC and inverse MFCC[C]//2018 *IEEE* 13*th International Conference on Industrial and Information Systems* (*ICIIS*). Rupnagar, India, 2018: 18 - 21. DOI:10.1109/ICIINFS.2018.8721379.

[25] Sun Y N, Yen G G, Yi Z. Evolving unsupervised deep neural networks for learning meaningful representations[J]. *IEEE Transactions on Evolutionary Computation*, 2019, **23**(1): 89 - 103. DOI:10.1109/TEVC.2018.2808689.

[26] Long J Y, Zhang S H, Li C. Evolving deep echo state networks for intelligent fault diagnosis[J]. *IEEE Transactions on Industrial Informatics*, 2020, **16**(7): 4928 - 4937. DOI:10.1109/TII.2019.2938884.

基于多窗口分析的语音检测方法

罗昕炜[1] 刘 婷[1] 黄 铭[1] 徐晓刚[1] 曹红丽[1] 柏祥华[2] 徐大勇[2]

([1] 东南大学水声信号处理教育部重点实验室, 南京 210096)

([2] 海装上海局驻南京地区第一军事代表室, 南京 210000)

摘要:针对低信噪比条件下的语音信号检测性能差的问题,提出了一种基于多窗口时频图的语音活动检测方法.首先,基于多窗口的时频分析计算得到语音信号的时频图,根据语音信号与背景噪声的特征差构造语音测试统计特征.其次,采用动态的双门限处理进行初步检测,在此基础上通过K均值聚类得到全局的双门限值.最后,根据序贯判决的思想进行检测得到该方法的判决结果输出.实验结果表明,低信噪比条件下,基于多窗口时频图的语音活动检测方法获得了良好的检测性能,并且在各种信噪比条件和背景噪声下,其性能总体优于传统的检测方法.该方法具有复杂度低、鲁棒性强、对不同种类语言适应性强等优点.

关键词:语音活动检测;多窗口谱分析;K-means 聚类;阈值调整;序贯决策

中图分类号:TN911.72