Sensors & Transducers, Vol. 242, Issue 3, March 2020, pp. 1-5

www.sensorsportal.com

Sensors & Transducers

Published by IFSA Publishing, S. L., 2020
http://www.sensorsportal.com

Novel Speech Endpoint Detection Algorithm for Voice Detectors in Interaction of Intelligent Terminals

[1, *] Jixiang LU and [2] Xiao HAN

[1] School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[2] Hamburg University of Technology, Hamburg, 21107, Germany
[1] Tel.: +86 13319288957 (CHN)
E-mail: lujixiang@nwpu.edu.cn

Received: 29 February 2020 /Accepted: 25 March 2020 /Published: 31 March 2020

Abstract: The human-computer interaction technology of intelligent terminals has totally altered from computer-centered to natural-oriented human-centered paradigm. As the most natural interaction mode of human beings, speech interaction has become a hot topic in both research and application fields in recent years. In the practical application environment of speech interaction technology, a signal recognition and processing method is required to get clean speech from the noisy speech. This paper presents a speech enhancement technique which is based on improved spectrum subtraction by using auditory masking effect, and introduces a novel speech endpoint detection algorithm, which overcomes the influence of music noise and extracts pure speech.

Keywords: Intelligent terminals, Human-computer interaction, Voice detector, Cloud computing.

1. Introduction

The utility computing business model of cloud computing enables all kinds of enterprises and individual users in the society to rent computing and service software and hardware resources, with their own demands, from cloud computing providers anytime and anywhere, in a dynamic, elastic, and scalable way. The service mode of cloud computing has changed the traditional computing and service mode, leading to a large-scale revolution of information technology. However, no matter how powerful and advanced the computing capability and network technology is, the ultimate goal of cloud computing is providing good service for end user, or people. Users can utilize various computing resources and network services in a simple and efficient way. With the rise of 5G mobile Internet and the rapid development of artificial intelligence technology, the mode that how people use computing services has been significantly changed. Intelligent terminals such as smart phones and smart watches build a new model of the cloud era -- intelligent cloud terminals. Specifically, intelligent terminals provide numerous of content information services that people need, meanwhile they provide all kinds of information to people in the most natural way through modern human-computer interaction technology.

New user terminals, which are represented by intelligent cloud terminals, have been linked with people more and more tensely and naturally. As a hot research field in the intelligent era, human-computer interaction technology is also transformed into intelligent cloud interaction by combining with new computing modes of cloud computing. Modern human-computer interaction technology is fully turning towards the human-centered and human-computer interactive-oriented. Natural human-computer interaction mode studies the cognitive relationship of human beings in a certain environment from a user-centered perspective. Such an interaction mode uses novel interaction channels, equipment, and techniques such as vision, speech, posture, gesture, brain electricity, and muscle electricity, etc., to make

people communicate with computer naturally, concurrently, and collaboratively in a variety of matters. With the integration of precise and imprecise inputs from multiple channels, and intelligently capture the interaction intentions of users, the naturalness and efficiency of human-computer interaction of intelligent terminals can be comprehensively improved [1].

Typically, as the most natural interaction mode among human beings, voice detector has become a hot research and application topic in recent years [2]. The application of cloud computing make it feasible to encapsulate the complicated algorithms such as speech purification and speech enhancement in the cloud with virtualization techniques, realizing the cloudification of speech interaction. By doing so, the problem of unavailability of the algorithms due to the weak computing and storage capabilities of terminals can be solved, and consequently the recognition accuracy is improved, achieving much higher performance of whole system [3]. Accordingly, many hi-tech companies such as Apple, Google, Microsoft, IBM, and Amazon have considered the cloudification of speech interaction as the research hotspot and launched related applications on a large scale [4]. It is worth mentioning that during the execution of speech interaction which includes speech recognition, speech synthesis, and speech control in a natural environment, environmental noise will seriously affect the application performance, and even lead to a system failure [5]. Therefore, it is necessary to get clean speech signal from the noisy speech in the natural environment [6]. Herein, speech enhancement, which is one of key techniques in the voice detector, is able to remove the noise in the speech and extract the pure speech signal [7].

One of the classical algorithms in speech enhancement is spectrum subtraction [8]. However, traditional spectral subtraction ignores the analysis of speech spectrum and so damages the intelligibility of speech, especially in the case of low signal-to-noise ratio (SNR) or non-stationary noise environment, which produces serious music noise [9]. Since the performance of speech enhancement is mainly affected by the auditory perception of human ears, how to overcome the influence of music noise is always a hot topic in the study of speech enhancement technology. In this paper, an improved spectral subtraction method based on auditory masking effect is proposed, and a new speech endpoint detection algorithm is introduced, finally the effectiveness of the algorithm is verified with the simulation [10-11].

2. Principles of Basic Spectrum Subtraction

Noisy speech can be expressed as $y(n)=s(n)+d(n)$, where $s(n)$ is the pure speech, $d(n)$ is the additive noise, and they are uncorrelated. The speech short-time analysis is carried out frame by frame. The noisy speech in the i-th frame is

$$y(n,i)=s(n,i)+d(n,i),\quad n=0,1,\cdots,L-1,\quad i=1,2,\cdots, \tag{1}$$

where L is the length of the frame.

By using Fast Fourier Transformation (FFT) we have $Y(k,i)=S(k,i)+D(k,i)$, where $k=0,1,\cdots,N-1$ is the discrete frequency, and N is the length of FFT. Besides, let $\hat{s}(n,i)$ denote the enhanced speech in the i-th frame. The short time discrete spectrum and short time discrete power spectral densities of $d(n,i)$, $s(n,i)$, $y(n,i)$, $\hat{s}(n,i)$ are $D(k,i)$, $S(k,i)$, $Y(k,i)$, $\hat{S}(k,i)$, $D_p(k,i)$, $S_p(k,i)$, $Y_p(k,i)$, $\hat{S}_p(k,i)$, respectively. Then, we have

$$\begin{aligned} Y_p(k,i)&=\|Y(k,i)\|^2=[S(k,i)+D(k,i)]\times\overline{[S(k,i)+D(k,i)]}\\ &=S_p(k,i)+D_p(k,i)+\overline{S(k,i)}D(k,i)+S(k,i)\overline{D(k,i)} \end{aligned} \tag{2}$$

In spectrum subtraction, $D_p(k,i)$, $S(k,i)\overline{D(k,i)}$, $\overline{S(k,i)}D(k,i)$ are used for set mean approximation. Since $d(n,i)$ is the stationary noise with zero-mean, and the independent with $s(n,i)$, so $E\left[S(k,i)\overline{D(k,i)}\right]$ and $E\left[\overline{S(k,i)}D(k,i)\right]$ are equal to zero, where $E[\bullet]$ is the set mean. Therefore, when considering a short-time stationary process in an analytical frame, the estimated value of $S_p(k,i)$ in Eq. (1) can be approximately expressed as

$$\hat{S}_p(k,i)=Y_p(k,i)-E\left[D_p(k,i)\right], \tag{3}$$

Thus, the estimated value $\hat{S}(k,i)$ of original speech $S(k,i)$ is

$$\left|\hat{S}(k,i)\right|=\left[Y_p(k,i)-E\left[D_p(k,i)\right]\right]^{1/2}, \tag{4}$$

where $E\left[D_p(k,i)\right]$ can be obtained by prior knowledge or estimation of the noise during break period of the speech, which is the basic principle of spectrum subtraction.

3. Proposed Improved Algorithm

The main disadvantage of spectrum subtraction is the narrow range of applicable SNR. When the SNR is decreasing, it is easy to lose useful components of the signal, and the large music noise will be introduced while eliminating the noise, and the improved SNR of

the algorithm is proportional to the size of the introduced music noise. The auditory system of human beings has masking effect, in which the frequency components of speech signals below the auditory masking threshold can be masked by similar frequency components of speech signals above the auditory masking threshold [9]. If the masking effect of noise on pure speech can be reduced or even eliminated, the problem of residual music noise can be solved well while enhancing speech. Therefore, the auditory masking effect based spectrum subtraction is an improved spectrum subtraction which is exploiting the auditory masking threshold of each key frequency segment in each frame of speech signal.

Fig. 1 illustrates the work-flow of proposed algorithm. Let $S(l,k)$, $D(l,k)$, and $Y(l,k)$ denote k-th frequency component of $s(n)$, $d(n)$ and $y(n)$ in the l-th frame, respectively. As shown in the Eq. (5), we categorize the power spectrum of the received signals into three states: no-speech state $H_0^{(l,k)}$, and speech state $H_{1,0}^{(l,k)}$ and $H_{1,1}^{(l,k)}$, where $H_0^{(l,k)}$ represents the state in which the noise is not masked by speech, and $H_{1,0}^{(l,k)}$ and $H_{1,1}^{(l,k)}$ represent the states in which the noise is masked by speech. $T(l,k)$ is the masking threshold of k-th spectrum component of pure speech in the l-th frame.

$$\begin{cases} \text{Non-speech } H_0^{(l,k)}: & Y(l,k)=D(l,k) \\ \text{Speech } H_1^{(l,k)} \begin{cases} H_{1,0}^{(l,k)} & Y(l,k)=S(l,k)+D(l,k) \quad D(l,k)>T(l,k) \\ H_{1,1}^{(l,k)} & Y(l,k)=S(l,k)+D(l,k) \quad D(l,k)\le T(l,k) \end{cases} \end{cases} \tag{5}$$

Then we smooth the speech signal in each critical band $\hat{S}$ and noise $\hat{D}$.

For noisy speech signal processing, realizing the endpoint detection (speech/non-speech detection) of noisy speech signal is very important. If the speech processed correctly, an accurate determination of the activity of the speech signal is necessary. Generally, there are many methods of speech endpoint detection, but most of them require a certain training algorithm to obtain the statistical information of background noise and speech in the environment with stable background noise and relatively high signal-to-noise. However, in the natural environment, the statistical information of background noise and speech is hard to obtain in advance, and the noise changes slowly when the SNR is relatively low.

The existing speech endpoint detection algorithms use the ratio of energy to zero-crossing rate in time domain as the criterion, i.e., $F = amp / ZCR$ (or take logarithm), where amp is the short-term energy of speech signal in a certain frame, and ZCR is the short-term zero-crossing rate, which is the number of signal sample changes in a frame. Thus, the criterion is shown as follows

$$F = C_0 \times \text{spectrum entropy} \times \text{energy} / ZCR, \tag{6}$$

where C_0 in the cepstrum coefficient, which is given by

$$C_0 = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \log S(\omega) d\omega,$$

where $S(\omega)$ is the energy spectral density function of the signal.

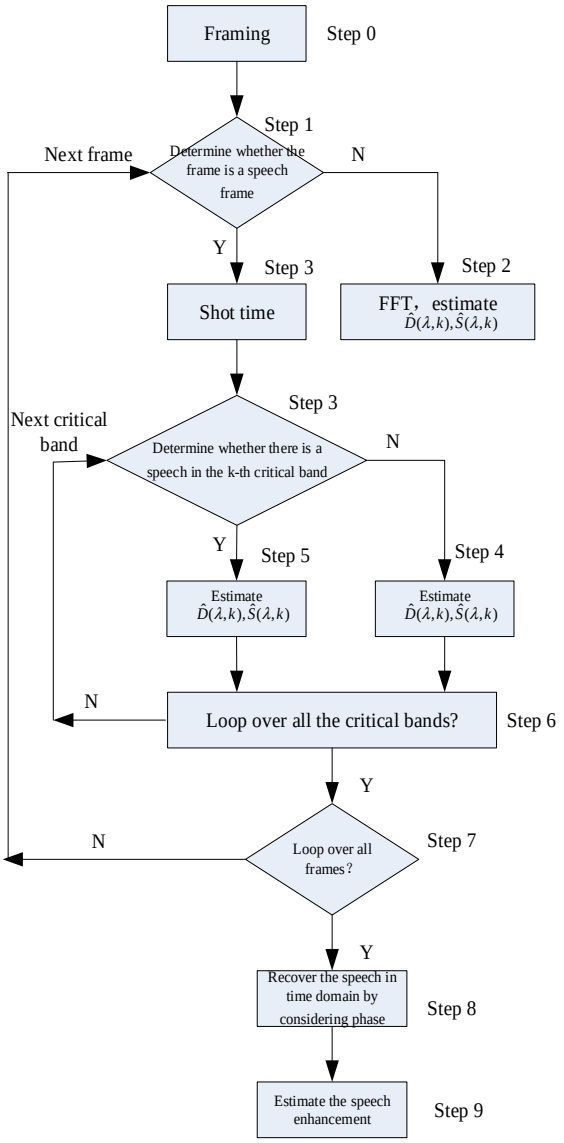

Fig. 1. Proposed improved algorithm.

$$C_0 = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \log S(\omega) d\omega, \tag{7}$$

We determine whether the signal is a speech frame in the time domain. If it is not a speech frame, we first use FFT, and then we have

$$D(\lambda,k)=\varepsilon\cdot D(\lambda-1,k)+(1-\varepsilon)\cdot\left|Y(\lambda,k)\right| \tag{8}$$

$$\hat{S}(l,k)=\beta_0\cdot\left|Y(\lambda,k)\right|, \tag{9}$$

where ε is a small constant. Then, we determine if there is a speech in the critical band of this frame.

$$\begin{aligned}D(\lambda,k)&=\alpha_0(\lambda,k)\cdot D(\lambda-1,k)\\&+(1-\alpha_0(\lambda,k))\cdot\left|Y(\lambda,k)\right|\end{aligned} \tag{10}$$

$$\hat{S}(l,k)=\beta_0\cdot\left|Y(\lambda,k)\right| \tag{11}$$

In Eq. (11), the value of $\alpha_0(\lambda,k)$ can be selected as ε. More specifically, $\alpha_1(\lambda,k)$ can be calculated based on the masking threshold represented by $T_i(i=1,2,\cdots,B)$ (B is the total number of critical band) or the energy of each critical band represented by $E_i(i=1,2,\cdots,B)$. For example, we can use $\frac{T_i}{\sum_i T_i}$ or $\frac{E_i}{\sum_i E_i}$ x $\max\{T_i\}-\min\{T_i\}$. If the reduction is used, the range of the coefficient is from 3 to 6, and γ is from 0.05 to 0.2.

Consequently, the speech signal and noise are

$$\begin{aligned}\hat{S}(l,k)&=\eta\cdot\hat{S}(l-1,k)+(1-\eta)\times\\&\times\max\left\{Y(\lambda,k)-\beta_1\cdot D(\lambda-1,k),\beta_0\cdot\left|Y(\lambda,k)\right|\right\}\end{aligned} \tag{12}$$

$$\begin{aligned}\hat{S}(l,k)&=\eta\cdot\hat{S}(l-1,k)+(1-\eta)\times\\&\times\max\left\{Y(\lambda,k)-\beta_1\cdot D(\lambda,k),\beta_0\cdot\left|Y(\lambda,k)\right|\right\}\end{aligned} \tag{13}$$

$$\begin{aligned}\hat{D}(\lambda,k)&=\\&=\alpha_1(\lambda,k)\cdot D(\lambda-1,k)+(1-\alpha_1(\lambda,k))\cdot\left|Y(\lambda,k)\right|\end{aligned} \tag{14}$$

$$\begin{aligned}\hat{D}(\lambda,k)&=\\&=\alpha_1(\lambda,k)\cdot D(\lambda-1,k)+(1-\alpha_1(\lambda,k))\cdot\beta_1\cdot D(\lambda-1,k)\end{aligned} \tag{15}$$

Specifically, β_0 and β_1 of spectrum subtraction's coefficient in each Bark frequency can be determined as follows

$$\frac{T_{\max}-T}{\beta_1(T)-\beta_{1,\min}}=\frac{T-T_{\min}}{\beta_{1,\max}-\beta_1(T)} \tag{16}$$

$$\frac{T_{\max}-T}{\beta_0(T)-\beta_{0,\min}}=\frac{T-T_{\min}}{\beta_{0,\max}-\beta_0(T)}, \tag{17}$$

where $\beta_{1,\min}=1$, $\beta_{1,\max}=6$, $\beta_{0,\min}=0$, $\beta_{0,\max}=0.02$, respectively.

Besides, let us define the changing trend of the noise as λ_m, which can be given by

$$\begin{aligned}\lambda_m&=\frac{\sum_k^k\left|D^{low}(m,k)\right|+\sum_k^k\left|D^{high}(m,k)\right|}{\sum_k\left|W^{low}\right|+\sum_k\left|W^{high}\right|}\\&=\frac{\sum_k^k\left|Y^{low}(m,k)\right|+\sum_k^k\left|Y^{high}(m,k)\right|}{\sum_k\left|W^{low}\right|+\sum_k\left|W^{high}\right|}\end{aligned} \tag{18}$$

$$\left|\hat{D}(m,k)\right|=\lambda_m\cdot|w| \tag{19}$$

Then we have

$$\hat{S}(m,k)=\left|Y(m,k)\right|-\left|\hat{D}(m,k)\right|,$$

when

$$\left|Y(m,k)\right|\geq\left|\hat{D}(m,k)\right|, \tag{20}$$

where $|w|$ represents the average estimation of the noise spectrum in "silent segment" by looping over value of each critical band.

$$\begin{aligned}\hat{D}(\lambda,k)&=\alpha_1(\lambda,k)\cdot D(\lambda-1,k)\\&+(1-\alpha_1(\lambda,k))\cdot\max\left\{Y(\lambda,k)-\hat{S}(\lambda,k),0\right\}\end{aligned} \tag{21}$$

4. Conclusions

The background noise in the speech not only seriously affects the speech communication, but also significantly decreases the performance of the whole voice detector. By considering the fact that the human auditory system itself has certain anti-noise ability, and combining the insensitivity of ears of human beings to speech phase with the stationary characteristics of short-term speech, this paper proposed an Novel improved speech enhancement method based on spectrum subtraction with auditory masking effect.

References

[1]. N. Radha, A. Shahina and A. N. Khan, A Study on Alternative Speech Sensor, in *Proceedings of the International Conference on Computer, Communication, and Signal Processing (ICCCSP)*, Chennai, 2018, pp. 1-6.

[2]. Kun-Ching Wang, A Novel Voice Sensor for the Detection of Speech Signals, *Sensors*, 13, 12, December 2013, pp. 16533–16550.

[3]. C. Demiroglu, D. V. Anderson, M. A. Clements and T. Barnwell, Multi-Sensor Spectro-Temporal Comb Filtering for Speech Enhancement, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, Honolulu, HI, 2007, pp. IV-589-IV-592.

[4]. D. Ayllón, R. Gil-Pita, M. Rosa-Zurera and H. Krim, Real-time multiple DOA estimation of speech sources wireless acoustic sensor networks, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2015)*, Brisbane, QLD, 2015, pp. 2709-2713.

[5]. S. Srinivasan and P. Kechichian, Utility of auxiliary sensor data for speech enhancement, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013, pp. 7294-7298.

[6]. Y. Avargel and I. Cohen, Speech measurements using a laser Doppler vibrometer sensor: Application to speech enhancement, *in Proceedings of the Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, Edinburgh, 2011, pp. 109-114.

[7]. C. Yin and F. Kuo, A Study of How Information System Professionals Comprehend Indirect and Direct Speech Acts Project Communication, *IEEE Transactions on Professional Communication*, Vol. 56, No. 3, Sept. 2013, pp. 226-241.

[8]. Q. J. Dai, Y. P. Chen, Optimizing speech enhancement based on noise masked probability, in *Proceedings of the 6th IEEE International Conference on Signal Processing (ICSP'02)*, 2002, pp. 448-451.

[9]. M. Shujau, C. H. Ritz and I. S. Burnett, Linear Predictive perceptual filtering for Acoustic Vector Sensors: Exploiting directional recordings for high quality speech enhancement, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2011)*, Prague, 2011, pp. 5068-5071.

[10]. Murakami, T., Namba, M., Hoya, T. Speech enhancement based on a combined higher frequency regeneration technique and RBF networks, in *Proceedings of the IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering (TENCON '02)*, Oct. 2002, Vol. 1, pp. 457-460.

[11]. Cai Hantian, Yuan Hongtao, A speech enhancement algorithm based on masking properties models. *Journal on Communications*, Vol. 23, No. 5, 2002, 8, pp. 93-98 (in Chinese).