A ROBUST SPEECH DETECTION ALGORITHM IN A MICROPHONE ARRAY TELECONFERENCING SYSTEM

Qiyue Zou, Xiaoxin Zou, Ming Zhang, Zhiping Lin*

School of Electrical and Electronics Engineering, Nanyang Avenue
Nanyang Technological University, Singapore 639798
*Centre for Signal Processing, Block S2, S2-B4b-05, Nanyang Technological University, Nanyang
Avenue, Singapore 639798. Email: emzhang@ntu.edu.sg

ABSTRACT

This paper describes a robust speech detection algorithm that can operate reliably in a microphone array teleconferencing system. High performance in a non-stationary noisy environment is achieved by combining the following techniques: (1) noise suppression by spectral subtraction, (2) silence detection by adaptive noise threshold and (3) non-stationary noise detection based on the availability of pitch signal. This algorithm can prevent the microphone-array-based speaker tracking system from being misguided by noises commonly present in a conference room. Real world experiments show that this algorithm performs very well and has the potential for practical applications.

1. INTRODUCTION

A microphone array has a number of advantages over a single microphone in teleconferencing applications. A microphone array can pass a high-quality speech signal from a desired source location by attenuating interference noises from other locations [1]. Also, the speaker's location can be precisely determined by time delay estimation techniques [2], so that a computer-controlled camera could track the movement of a speaker while transmitting concurrent images to remote listeners [3].

However, in a conference room, some interference noises may be produced accidentally, such as coughing, door knocking and hand clapping. These kinds of noises may cause the microphone array to locate the wrong speaker, resulting in unnecessary embarrassment. For example, the person who is coughing may distract the camera from focusing on the major speaker. Noises will also affect the accuracy of beamforming, which will degrade the quality of the transmitted signal.

Most speech detection techniques used in microphone array systems are similar to those developed for telecommunications and speech recognition [4], [5]. Noise energy threshold technique is widely used because of its reliability. Rabiner combines the short-time energy levels and the zero-crossing rate to detect the endpoint of a speech when the signal-to-noise ratio (SNR) is good [6]. Some other methods require the prior knowledge of background noises or speech signals to perform frequency domain analysis [7]. Recently, it is shown that the SNR estimated from a multi-channel microphone array can be used to predict the presence of speech [8]. However, most of these developed algorithms assume that the noise is stationary or its spectrum can be estimated accurately. Therefore, these techniques may not work well for non-stationary noises.

To overcome this problem, our paper proposes a new speech detection algorithm for a non-stationary noisy environment by combining the spectral subtraction, adaptive energy threshold and pitch detection techniques. Firstly, spectral subtraction aims at attenuating those stationary noise components, such as the sound of a noisy electrical fan. Secondly, the adaptive energy threshold method helps to select potential speech signals from input signals based on their energy levels. Thirdly, a pitch detection algorithm checks whether a voice signal is present or not. Some heuristic but important rules make a final decision.

The details of the speech detection algorithm will be presented in the next section. Section 3 gives some impressive real world experimental results. Section 4 comes with the conclusion.

2. THE SPEECH DETECTION ALGORITHM

The proposed speech detection algorithm for a microphone array teleconferencing system is illustrated in the following block diagram.

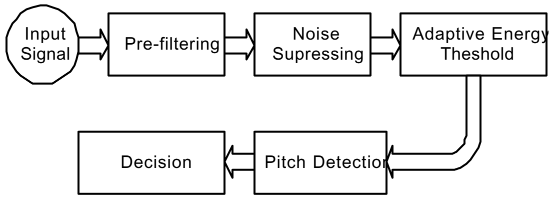


Fig. 1. Block diagram of the proposed speech detection algorithm

2.1 Pre-filtering

The spectrum of speech signal is ranging from 200Hz to 3400Hz. A bandpass filter is used to reject those unwanted frequency components, such as the 50Hz electrical noise.

2.2 Noise suppression

In a meeting room, some background noise always comes from electrical fans, air-conditioners, typewriters and other equipment. Part of that noise cannot be effectively removed by filtering, as its spectrum occupies the same passband as speech signal. The spectral subtraction technique could be used to improve the quality of speech signal [9].

The clean speech signal corrupted by stationary noise can be mo deled in the time domain as

$$y(t) = s(t) + n(t) \qquad (1)$$

where *y(t)* denotes the degraded speech signal, *s(t)* the clean speech signal and *n(t)* the stationary noise.

In the frequency domain, the model is

$$Y(\omega) = S(\omega) + N(\omega)\,. \qquad (2)$$

Noise is assumed to be uncorrelated to the speech signal, so

$$|Y(\omega)|^2 = |S(\omega)|^2 + |N(\omega)|^2\,. \qquad (3)$$

Due to the stationary characteristic of such background noise, its spectral power density, $\left|N(\omega)\right|^2$, can be estimated during those periods when no speech signal is present. The estimated power density of clean speech signal, $|\tilde{S}(\omega)|^2$, is obtained by subtracting the estimated noise spectral power density $\left|\tilde{N}(\omega)\right|^2$ from the noisy signal spectral power density $\left|Y(\omega)\right|^2$.

$$\left|\tilde{S}(\omega)\right|^2 = \left|Y(\omega)\right|^2 - \left|\tilde{N}(\omega)\right|^2\,. \qquad (4)$$

The estimated clean speech signal can be reconstructed as

$$s(t) = F^{-1}[\left|\tilde{S}(\omega)\right| \measuredangle jY(\omega)]\,, \qquad (5)$$

where $\measuredangle jY(\omega)$ denotes the phase of $Y(\omega)$.

This noise suppression technique requires a prior process to distinguish between background noise segments and speech segments. In a low noise environment, adaptive energy threshold can be exploited to select those stationary noise segments to update the background noise estimation (see Fig. 2). In a high noise environment, the result of this speech detection algorithm determines when to re-estimate the power spectral density of background noise.

2.3 Adaptive energy threshold

To detect potential speech segments, static energy threshold cannot adapt to varying environment automatically, so the adjustment of energy threshold based on the current noise level becomes reasonable [10]. Fig. 2 illustrates the procedure of speech detection by the adaptive energy threshold.

At startup, the energy threshold is initialized by assuming that there are no speech activities in the first few seconds. In process, a stack is maintained to record the speech and noise energy levels in the last few minutes.

For each input signal segment, its energy is computed and compared with the threshold. If the input segment is regarded as "silence", the threshold, η_{speech}, is updated based on the past noise energy E_n and speech energy E_s as

$$\eta_{speech} = \tau\left(\sum_{k=1}^{M} \frac{1}{k} E_n\right) + \mu\left(\sum_{l=1}^{M} \frac{1}{k} E_s\right), \qquad (6)$$

where *M* represents the length of stack, factor *1/k* is used to incorporate the timing influence, τ and μ are pre-defined para meters for different environment. Our experiments show that this detector performs very well in a meeting room when τ=1 and μ=0.3, which may need to be adjusted in a new environment.

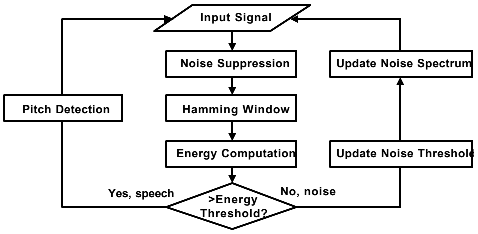


Fig. 2. Adaptive energy threshold

2.4 Pitch detection

Pitch signal is a very important feature that makes human speech signal distinct from common interference noises, especially non-stationary noises such as coughing. The sounds made by human being can be grouped into two categories according to the mechanism of generation: voiced and unvoiced [11]. Voiced speech is produced by the periodic vibration of the vocal cord when the air flows from lungs; unvoiced speech is made by turbulent airflow and is random in nature. The basic synthesis model (see Fig. 3) for speech signal can be modeled as an all-pole linear, time-varying system excited by quasi-periodic pulses (voiced signal) or random noises (non-voiced signal).

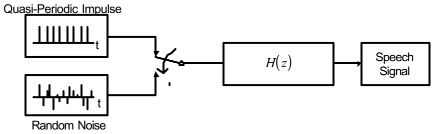
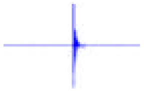
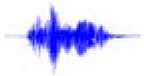


Fig. 3. Speech production model

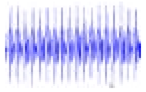
The generation of non-stationary noises present in a conference room is hard to be fitted into such a model. For example, the production of human cough is very similar to some of unvoiced sounds, in which the airflow is abruptly forced going through the mouth. The waveform of the sound of hand clapping appears as a short time impulse. Obviously, most kinds of the noises don't have the quasi-periodic driven sources. This feature is exploited by the pitch detection algorithm to determine whether a speech signal is present or not based on the availability of pitch.



Hand Clapping



Cough



Voice

The all-pole, linear and time-varying system can be expressed as

$$H(z)=\frac{S(z)}{U(z)}=\frac{G}{1-\sum_{k=1}^{n}a_k z^{-k}}\,. \qquad (7)$$

So, the speech signal generated by this model has a linear prediction form in time domain.

$$s(n)=\sum_{k=1}^{n}a_k s(n-k)+Gu(n)\,, \qquad (8)$$

where $a_1,...a_n$ are the linear prediction coefficients, $s(n)$ the short-time speech signal, $u(n)$ the excitation function. For a voiced signal, $u(n)$ should contain significant periodic impulses. For noise signal, the periodicity will not be so obvious.

For a short-time speech signal segment, the linear prediction coefficients, $a_1,...a_n$, can be estimated by Durbin algorithm [6]. The estimation error signal $u(n)$ can be obtained by inverse filtering $s(n)$ by $H^{1}(z)$. The autocorrelation of the error signal will show obvious periodic peaks for voiced signal.

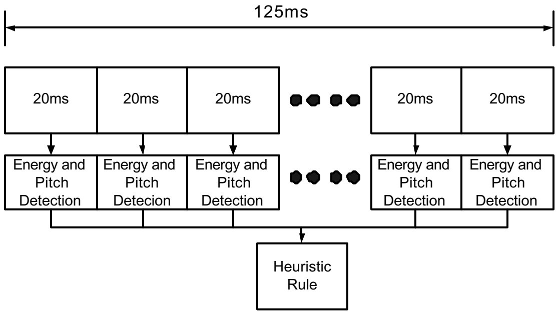
2.5 Heuristic determination

In this subsection, some important heuristic rules are proposed to ensure that the overall performance of the speech detection algorithm based on the three techniques mentioned above is accurate, robust and reliable. These rules exploit the special features of speech signals and noises.

Speech signals are locally stable during a short time, around 20-30ms. Thus, a hamming window with a length of 20ms is chosen to segment the input signal for speech detection. Some important parameters about energy and pitch can be measured within a single segment using those techniques mentioned above. However, making a speech/non-speech decision from only one segment is not reliable because some noises are locally periodic in a very short duration. An examination over a longer duration applies those heuristic rules to combine several single detection results can achieve a good performance. However, a longer duration may result in a longer time-delay. Our experiment chooses a period of 125ms for decision-making, which has a negligible delay but high accuracy. Fig. 4 shows the basic flowchart of decision-making based on heuristic rules.

The four rules used in this algorithm are listed below:

a) Energy



125ms
20ms
20ms
20ms
20ms
20ms
Energy and Pitch Detection
Energy and Pitch Detecion
Energy and Pitch Detection
Energy and Pitch Detection
Energy and Pitch Detection
Heuristic Rule

Fig. 4. Decision-making based heuristic rules

The energy level, $E(t)$, of speech signal should change continuously without abrupt peaks such as impulse noise. The first derivative, $\Delta Energy$, of short-time energy is used to measure its changing rate over time. Those segments with large derivatives are regarded as "noise" and discarded.

$$\Delta Energy = E(t+\Delta t)-E(t)\,. \qquad (9)$$

b) Duration of speech signal

The time duration of one English word is usually longer than 70ms. A confidence measurement is taken to ensure that a certain length of speech is detected during a 125ms interval.

c) Pitch tracking

The pitch signal must be available frequently during a piece of speech. Also, the allowable pitch frequency is limited to between 200Hz and 1300Hz. The pitch frequency during a short time cannot change too fast.

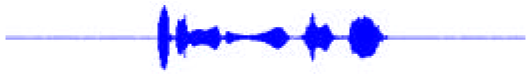
d) End-point of word

The end-point of a word is easily detected as "noise" because of its low energy and unvoiced components. Therefore, the detected "speech" interval is extended a little as the real speech boundary.

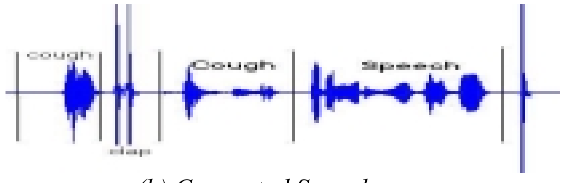
An input signal segment that can pass all the four rules is regarded as a speech signal.

2.6 Simulation

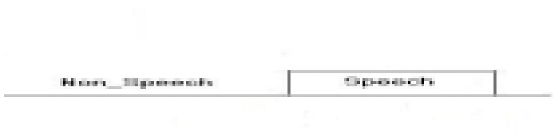
In the simulation, a piece of clean American accent speech is recorded using a microphone array. Different noises collected from a conference room are mixed with this clean speech to generate a corrupted speech signal.



(a) Clean Speech



(b) Corrupted Speech



(c) Detection Results

3. EXPERIMENTAL RESULTS

The test teleconferencing system uses an 8-element linear microphone array with a camera mounted in the middle. The size of conference room is 2.5m×5m, and the array is located near the front wall (see Fig. 5).

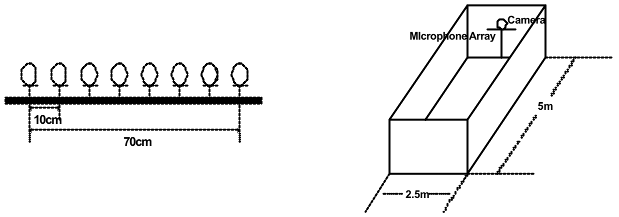


Fig. 5. Microphone array system

A speaker is standing 2-meter away from the camera, and a loudspeaker works as a noise source. Several typical cases in a conference room are selected to evaluate this algorithm. A speaker is required to say some words or continuous sentences with interference noise around. If a speech signal is detected, camera will rotate to the direction of the speaker; otherwise, camera will remain at its original position. The total times of success and failure are accumulated to check its reliability. Six male and six female of different ages are invited to take part in this test. The following tables show the recorded results.

Case 1: Only noise is present.

Noise Type	**Test Times**	**Results**		**Ratio**
		Success	**Fail**	
White Noise	100	100	0	100%
Cough	100	97	3	97%
Hand Clapping	100	95	5	95%
Knocking Door	100	96	4	96%

Case 2: Only speech is present.

Speech	**Test Times**	**Results**		**Ratio**
		Success	**Fail**	
Word	100	97	3	97%
Sentence	100	95	5	95%

Case 3: Speech and noise are both present.

Speech + Noises	**Times**	**Results**		**Ratio**
		Success	**Fail**	
White Noise	100	95	5	95%
Cough	100	93	7	93%
Hand Clapping	100	98	2	98%
Knocking Door	100	93	7	93%

For the purpose of comparison, other common used speech detection methods are implemented and simulated in the same conditions. The selected methods include those based on energy feature, short-time crossing rate, signal-to-noise ratio and Cepstral feature. It is convinced that our algorithm has a better performance in a non-stationary noisy environment.

Success Ratio of Other Speech Detection Methods

Noise Type	**Energy Feature**	**Short-time Cross Rate**	**SNR**	**Cepstral Feature**
White (10dB)	92%	86%	91%	87%
Cough	10%	44%	30%	51%
Hand Clapping	37%	85%	46%	79%
Knocking Door	47%	73%	43%	52%

4. CONCLUSION

In this paper, we have proposed a new speech detection algorithm for a microphone array teleconferencing system. The actual experiments show that it can reliably distinguish speech from interference noises, especially non-stationary noises. However, in the room with loud reverberation, its performance is somewhat degraded. In the future, we plan to combine the algorithm with an effective reverberation cancellation method to improve its robustness.

REFERENCES

[1] J.L. Flanagan, "Autodirective Microphone Systems", Acoustica, vol. 73, pp. 58-71, 1991.

[2] M.S. Brandstein, "A Framework for Speech Source Location", Ph.D thesis, Brown University, May 1995.

[3] H. Wang and P. Chu, "Voice Source Localization for Automatic Camera Pointing System in Videoconferencing", In Proc. of IEEE ASSP Workshop, pp. 187-190, 1997.

[4] L. Lamel and L. Rabiner, "An Improved Endpoint Detector for Isolated Word Recognition", IEEE Trans. ASSP, vol 29, pp. 777-785, 1981.

[5] D. Wu, M. Tanaka, R. Chen, L. Olorenshaw, M. Amadar and X. Menendez-Pidal, "A Robust Speech Detection Algorithm for Speech Activated Hands-free Applications", In Proc. of ICASSP, vol 4, pp. 2407-2410, 1999.

[6] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, N.J, 1978.

[7] J.A. Haigh and J.S. Mason, "Robust Voice Activity Detection Using Cepstral Features", In Proc. of IEEE TENCON'93, vol 3, pp. 321-324, 1993.

[8] J.F. Chen and W. Ser, "Speech Detection Using Microphone Array", Electronics Letters, vol. 36(2), pp. 181-182, Jan 2000.

[9] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. ASSP, vol 27, pp. 113-120, 1979.

[10] K.G. Cheong and R.W. Donaldson, "Adaptive Silence Detection for Speech Storage and Voice Mail Applications", IEEE Trans. ASSP, vol 36(6), pp. 924-927, June 1988.

[11] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, N.J., 1993.