Voice Detection using Speech Energy Maximization and Silence Feature Normalization

In-Sung Han¹ and Chan-Shik Ahn²

¹ Dept. of The 2nd R&D Institute, Agency for Defense Development
   460,Songpa-gu, Seoul, 138-600, South Korea
   ishan78@add.re.kr

² Dept. of Computer Engineering, Kwangwoon University
   20, Gwangun-ro, Nowon-gu, Seoul, South Korea
   coolshah@gmail.com

Abstract. The estimation of noises is very important to remove noises from voices with background noises. This paper proposed a voice detection method using a silence feature normalization and voice energy maximize. In the high signal-to-noise ratio for the proposed method was used to maximize the characteristics receive less characterized the effects of noise by the voice energy. Cepstral feature distribution of voice/non-voice characteristics in the low signal-to-noise ratio and improves the recognition performance. Result of the recognition experiment, recognition performance improved compared to the conventional method.

Keywords: Speech Recognition, Voice Detection, Noise Reduction, Speech Energy Maximization, Silence Feature Normalization.

1 Introduction

In case of feature parameters used for voice detecting, in an environment that SNR is high, as the characteristics of feature parameter on speech and non-speech are relatively distinctive, their recognition function doesn’t remarkably go down. However, in an actual noise environment that there are various kinds of noise or in one that SNR is low, as feature parameter is sensitive against noise signal, its voice detecting function decline may arise as a problem.

Therefore, this paper suggests voice detecting method that has advantages in noise environment by using speech maximization and silence feature normalization. Suggestion made use of characteristic of the silence feature’s being less affected by maximizing speech energy in high SNR environment. On the other hand, in low SNR, recognition function got improved using the characteristics of cepstrum feature distribution of speech and non-speech. Function improvement compared to existing methods was verified through recognition experiments.

★ Corresponding author
This paper consists of 5 parts in total. Related studies will be mentioned in the second chapter and voice detecting method that has advantages against noise environment using speech energy maximization and silence feature normalization will be explained in detail in the third chapter. Systematic assessment is conducted in chapter 4 and this paper is concluded with chapter 5.

2 Related Work

2.1 Spectral Energy

In low SNR energy spectrum, speech band shows relatively high energy spectrum compared to non-speech band. Therefore, it can be assumed that speech energy spectrum has comparatively higher energy spectrum than non-speech energy spectrum. Entropy can be explained with below formula in spectral energy band [1].

Spectral energy probability can be expressed in below formula

\[ F(k; l) = \frac{|Y(k, l)|^2}{\sum_{k=1}^{K} |Y(k, l)|^2} \]

\[ (1) \]

\( k \) shows Frequency bin Index, \( l \) shows frame index. Spectral energy probability against Frequency bin can be deduced from Frame \( l \). Deduced each frequency bin probability is calculated to entropy by formula

2.2 Critical Band

Masking effect is the phenomenon of weaker sound’s being blocked by stronger sound and stronger sound’s called masker, and the weaker sound is called maskee(weaker sound). Maskee can be hear when it’s over masking threshold calculated by the masker(stronger sound) [2].

Frequency masking is that the maskee is being masked when the masker and the maskee simultaneously occur and the masking degree can be calculated through frequency analysis. The threshold of the sound being masked differ from frequency bandwidth, and frequency bandwidth with equivalent masking threshold is called ‘critical band’. The width of critical band is constant as 100Hz when the frequency is below 1kHz, and when the frequency is over 1kHz, it proportionally increases according to the frequency[3].
3 Voice Detection That Excel in Noise Environment

3.1 Voice energy maximization

Voice has pitch frequency for vowels, and the pitch frequency is called 'Basic frequency'. Basic frequency has the characteristics of showing the biggest energy throughout the full-band of voice band, so it’s expressed as the maximum energy and minimum energy is shown as noise signal unrelated to voice signal.

Voice energy has relatively bigger energy compared to noise and noise has smaller amount of energy, so calculation of the ratio of voice energy against noise energy and energy ratio on the output are shown in below formula.

\[ PSR(b_t, l) = \frac{E(b_t, l) - E_{\text{max}}(l)}{\mu(l)} \]  

(2)

3.2 Silence feature normalization

Silence feature normalization is the method of finding the silence interval with weak log energy and then cutting the value down under a certain value.

Voice interval contaminated by noise has broader frequency bandwidth compared to the band only with voice and the interval with big log energy is less affected by noise compared to the interval with small log energy, so used weighting function. The weighting function \( \omega(n) \) can be shown like below formula.

\[
\omega(n) = \begin{cases} 
1/(1 + \exp(-(\log \bar{E}(n) - T_0)/\beta \sigma_2)) & \text{if } \log \bar{E}(n) > T_0 \\
1/(1 + \exp(-(\log \bar{E}(n) - T_0)/\beta \sigma_2)) & \text{if } \log \bar{E}(n) \leq T_0 
\end{cases}
\]  

(3)

4 Experiment Result

This paper conducted the experiment on voice detection method that excel in noise environment by using voice energy maximization and silence feature normalization that this paper suggested. For assessment, Aurora 2.0 database was used. Aurora 2.0 consists of noise environment and each-noise level(including white gaussian noise,
h babble noise) and sorts each noise environment (street, airport, car noise etc.) so it’s used to verify voice improvement algorithm[4].

The experiment was conducted in each noise environment with 15dB, 10dB, 5dB and 0dB separately to verify its performance on SNR change, also to evaluate endpoint detection performance PHR (Pause Hit Ratio) and FAR (False Alarm Ratio) in sound area were used. As sound source, 8kHz sampling rate, 16bits were used and FFT size is 256 samples and 1/2 overlapping interval was used. Hamming Window was also used[5].

PHR used as performance scale shows 98% accuracy in low SNR of noise environment and showed 100% accuracy in high SNR. FAR (False Alarm Ratio) in sound area was 0% by showing outstanding performance in SNR of 15dB and 10dB and 2% performance were marked for each in low SNR interval of 5dB and 0dB

![Fig. 1. Result of (a) Input signal (b) MMSE-STSA (c) Proposed method at SNR 10dB Airport Noise](image)

5 Conclusion

This paper evaluated voice detection and recognition performance that excel in noise environment by using speech energy maximization and silence feature normalization. As feature parameter is sensitive against noise signal in actual noise environment with various kinds of environmental noise or for low SNR voice, so voice detection performance deteriorates. Therefore, voice detection method that excel in noise environment using voice energy maximization and silence feature normalization was suggested. Suggested method made use of characteristics of silence feature’s being less affected by noise by maximizing voice energy in high SNR and voice and non-voice cepstrum feature dispersion in low SNR.

In result, recognition performance has accuracy 99% in char noise environment. FAR performance is good (0%) in SNR 15dB and 10dB. However, in low level 5dB
and 0dB it just improved 2% of its performance. Improved recognition performance compared to existing method was verified through experiment result.

References