ADAPTIVE PITCH-BASED SPEECH DETECTION FOR HANDS-FREE APPLICATIONS

A. R. Abu-El-Quran and R. A. Goubran

Department of Systems and Computer Engineering, Carleton University
1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6, Canada
(arami, goubran)@sce.carleton.ca

ABSTRACT

This paper proposes a new algorithm for classifying an audio segment as speech or non-speech. The proposed algorithm is capable of handling reverberation and low signal-to-noise environments; therefore, it is suitable for hands-free applications.

The algorithm divides an audio segment into frames, estimates the presence of pitch in each frame, and calculates a pitch ratio parameter. This parameter is then used to classify the audio segment. The threshold used in calculating this parameter is adapted to accommodate different environments. The performance of the proposed algorithm is evaluated for different signal-to-noise ratios and different segment sizes using a library of audio segments. The library includes speech segments and non-speech segments such as fan noise and cocktail noise.

Using 0.4 second segments it is shown that the proposed algorithm can achieve a correct decision for 95.7% of the speech segments and 96.7% of the non-speech segments under reverberant conditions.

1. INTRODUCTION

In many applications such as hands-free conferencing and talker localization, it is necessary to determine whether an audio segment is speech or non-speech [1].

In hands-free conferencing, for example, this knowledge is desirable to adjust the parameters of several speech processing subsystems such as echo cancellation, noise cancellation, speech quality enhancement [2] or speech recognition [3]. Hands-free systems suffer from echoes, environment noise originating from the undesired sources, and reverberation as shown in figure 1.

Many modern video conferencing systems use a microphone array and a steerable camera [4]. The microphone array determines the location of the various sounds emanating from the room and steers the camera toward the talker. In these systems it is important to discriminate between speech and non-speech segments to ensure that the camera is pointing to a talker and not to a noise source [4], [5]. This task is also important when the microphone array is used to perform noise cancellation combined with near-field adaptive beamforming [6].

In the above applications a delay of 0.5 seconds is acceptable. However, the algorithm has to be capable of handling reverberation and low signal-to-noise (SNR) environments.

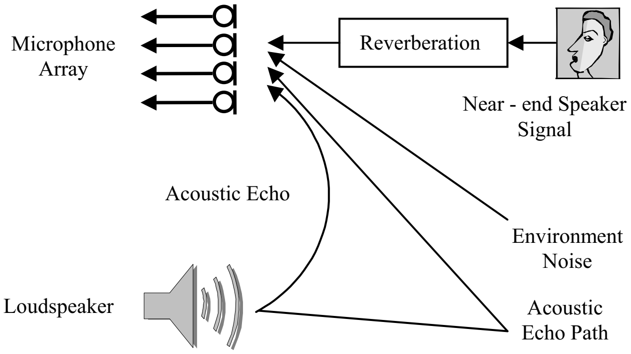

Figure 1. Hands-free conferencing environment

This paper presents a new feature that can be used in speech and non-speech classification under the previously mentioned conditions. The proposed method is based on the pitch ratio algorithm [1]. The algorithm is tested using reverberant with varying SNR audio library.

This paper is organized as follows; section 2 describes the adaptive pitch ratio algorithm. Section 3 describes the experimental results and the comparison with previous work. Section 4 presents the conclusion.

2. ADAPTIVE PITCH RATIO ALGORITHM (APR)

A block diagram of the proposed adaptive pitch ratio algorithm (APR) is shown in figure 2. An input audio segment is segmented into smaller frames then pitch detection for each frame is employed to calculate the pitch ratio. The pitch ratio is compared to a threshold to make the speech decision. The adaptive threshold is calculated by estimating the SNR for each input segment then

0-7803-8874-7/05/$20.00 ©2005 IEEE

calculating the corresponding threshold value. Modified autocorrelation pitch detection based on the center clipping method is employed as the pitch detection algorithm [7].

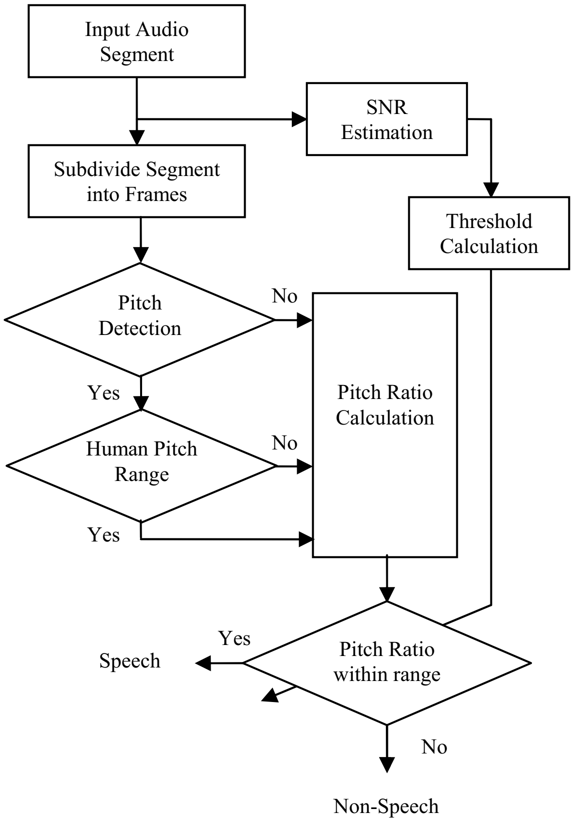

Figure 2. Adaptive pitch ratio Algorithm

The pitch ratio is the ratio of the number of frames that have a pitch to the total number of frames. The pitch ratio is calculated as follows:

$$\text{Pitch Ratio} = \frac{\text{NP}}{\text{NF}} \qquad (1)$$

where NP is the numbers of frames that have a pitch and NF is the total number of frames. The total number of the frames is calculated as follows:

$$\text{NF} = \left\lceil \left(\frac{\text{SD - FS}}{\text{1 - OR}} \right) / \text{FS} \right\rceil + 1 \qquad (2)$$

where SD is the audio segment duration, FS is the frame size (figure 3), and OR is the overlap ratio (eg. 50 %). $\lceil \bullet \rceil$ is the round down operator.

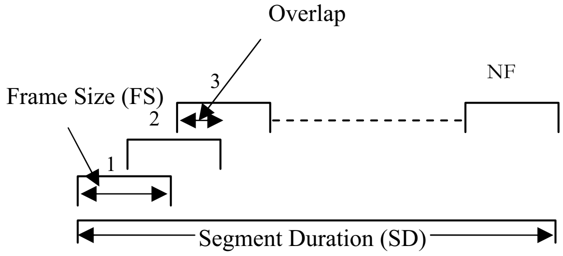

Figure 3. Parameters of the number of frames

The detected pitch is considered to be that of human speech if it falls in the [70-280 Hz] range [1]. The pitch ratio threshold is calculated to maximize the average correct speech and non-speech decisions.

Algorithms for SNR estimation for speech exist in the literature. SNR can be estimated using the kurtosis of noisy speech [8], adaptive filters [9], and filter banks [10].

The value of the SNR is used to determine the pitch ratio threshold which achieves the highest average correct decision. In this paper we recommend the use of the SNR estimation algorithm presented in [8] because of its simplicity and accuracy.

3. EXPERMINTAL RESULTS

A database containing a total of 911 speech and non-speech audio segments was collected. This database is the same database used in [1]. The sampling rate is 16 KHz with a resolution of 16 bits.

The impulse response of the room is generated using the image method to simulate reverberant rooms. The simulation program used is described in [11].

The proposed algorithm is tested using different SNR and reverberation environment. The same database is applied to different testing conditions.

The performance of the APR is compared to two previously used speech non-speech classification algorithms. These are the LPC algorithm described in [5] and the whole segment algorithm [12].

The LPC algorithm speech decision measures three values from the audio segment to be classified. These values are the change of the energy of the speech signal, speech duration, and the change of the pitch value using the LPC algorithm.

The whole segment algorithm's classification decision depends on the pitch detected in the whole input audio segment.

It is important to notice that the longer the segments size the longer the decision delay. The APR, LPC and whole segment algorithm will be compared. The results of an experiment relating the performance of these three algorithms as the segment size is varied are shown in figure 4.

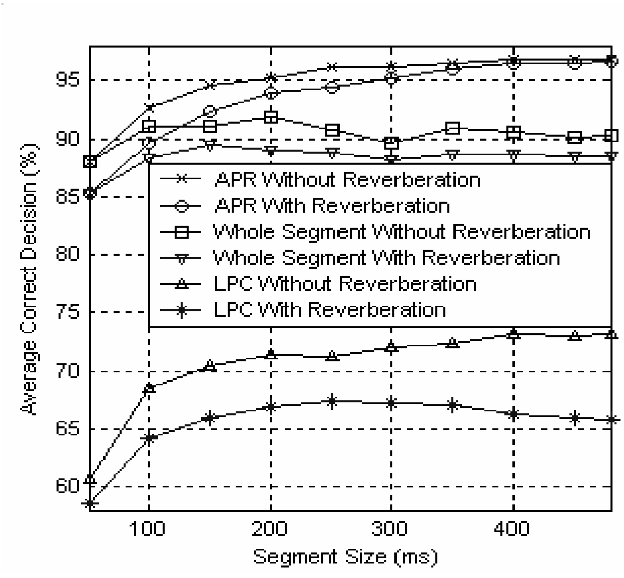

Figure 4. Average correct decision Vs segment size (ms) (APR: overlap ratio 0.75, and frame size = 50 ms)

The previous simulation shows that APR can achieve almost the same average correct decision for both the reverberant and original library at a segment size larger than 350 ms. Also it shows that the performance degradation of the other algorithms is higher than the degradation of APR in the reverberant environment (figure 5).

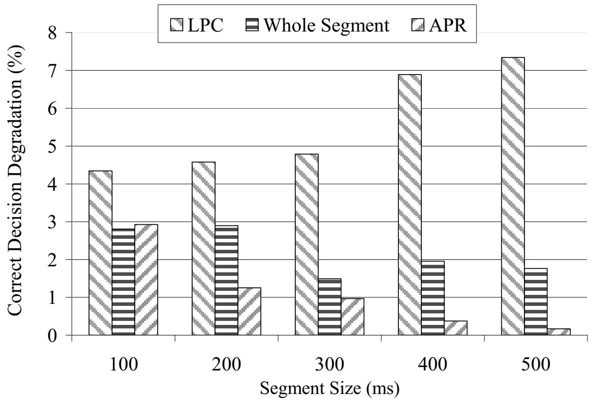

Figure 5. Performance degradation in reverberant environment

The adaptation of the pitch ratio threshold is done by estimating the values of the threshold that maximize the average correct decision corresponding to different SNR values. A piecewise linear curve is estimated from the relation between the pitch ratio threshold and the corresponding SNR.

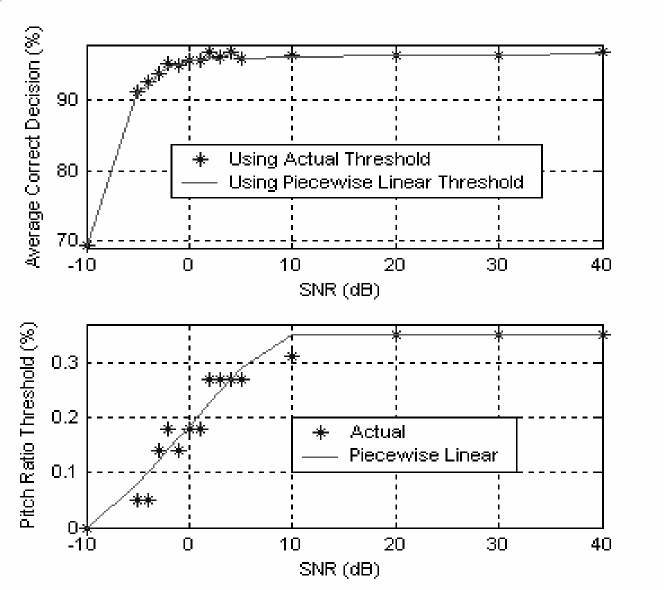

Figure 6. Adaptive threshold performance vs. SNR (dB)

As shown in figure 6 a piecewise linear model fit the curve, the use of the piecewise linear values instead of the actual values has almost no effect on the performance of the APR algorithm.

Simulation is done to compare the performance of the APR algorithm with the previous algorithms under different SNR values. This simulation is conducted using the reverberant audio library.

The simulation in figure 7 shows that the APR algorithm can achieve an average correct decision of 91% for a reverberant library with an SNR value of -5dB.

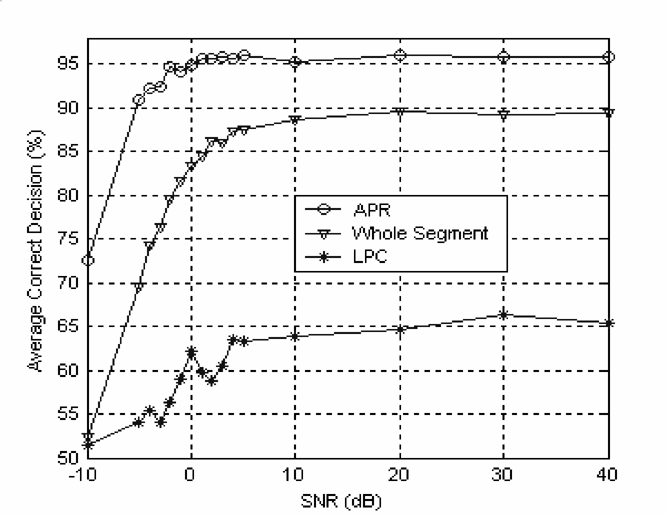

Figure 7. Comparison of performance under different SNR (dB)

Comparison between different algorithms for speech/non-speech correct decision using a segment size of 400 ms is shown in table 1

Table 1. Comparison between different algorithms for speech/non-speech correct decision

Method	Speech correct decision	Non-Speech correct decision
APR without reverberation	97.15 %	95.56 %
APR with reverberation	95.73 %	96.67 %
Whole segment without reverberation	86.65 %	94.46 %
Whole segment with reverberation	82.38 %	95.40 %
LPC without reverberation	74.78 %	71.75 %
LPC with reverberation	67.65 %	65.10 %

The APR shows a high performance for speech and non-speech audio types, also its average correct decision degradation in the reverberant environment is insignificant. The reverberation significantly affects the average correct decision of the LPC algorithm.

4. CONCLUSION

A new algorithm for classifying an audio segment as speech or non-speech has been introduced. Also it has been shown that the proposed algorithm is capable of handling reverberation and low signal-to-noise environments. It has been shown that the APR algorithm has better performance than the previous algorithms.

The simulation results showed that the APR algorithm is a good candidate to work in reverberant and low SNR environment, where the algorithm achieved 91% average correct decision for SNR equal to -5dB.

It is shown that the proposed algorithm achieves correct decision of 95.7% for speech and 96.7% for non-speech segments under reverberant conditions.

5. ACKNOWLEDGEMENT

The authors would like to acknowledge the funding from CITO, Mitel, NSERC, and OGS.

6. REFERENCES

[1] A.R. Abu-El-Quran, and R.A Goubran, “Pitch-based feature extraction for audio classification,” in *Proc., International Workshop Haptic, Audio, Visual Environments and their Applications*, Ottawa, ON, Canada, pp. 43-47, Sept. 2003.

[2] Wu Mingyang, and DeLiang Wang, “A one-microphone algorithm for reverberant speech enhancement,” in *Proc., International Conference Acoustics, Speech, and Signal Processing,* Hong Kong, pp. I-892 – I-895, April 2003.

[3] B.W. Gillespie, and A.E. Atlas, “Strategies for improving audible quality and speech recognition accuracy of reverberant speech,” in *Proc., International Conference Acoustics, Speech, and Signal Processing,* Hong Kong, pp. I-676 – I679, April 2003.

[4] D. Lo, R.A. Goubran, R Dansereau, G. Thompson, D. Schulz, “Robust Audio-Video Localization in Video Conferencing Using Reliability Information,” *IEEE Trans., Instrumentation and Measurement*, vol. 53, pp. 1132 – 1139, Aug. 2004.

[5] Qiyue Zou, Xiaoxin Zou, Ming Zhang and Zhiping Lin, “A robust speech detection algorithm in a microphone array teleconferencing system,” in *Proc., International Conference Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, USA, Vol. 5, pp. 3025–3028, May. 2001.

[6] Y. R. Zheng, R.A. Goubran, M. El-Tanany, “Robust Near-Field Adaptive Beamforming with Distance Discrimination”, *IEEE Trans., Speech and Audio Processing*, vol. 12, pp. 478-488, Sept. 2004.

[7] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, “A comparative performance study of several pitch detection algorithms,” *IEEE Trans., Acoustics, Speech, Signal Processing*, pp. 399–418, Oct. 1976.

[8] E. Nemer, R.A. Goubran, and S. Mahmoud, “SNR estimation of speech signals using subbands and fourth-order statistics,” *IEEE letters, Signal Processing,* pp. 171-174, July 1999.

[9] J. Rodriguez, F. Rios, R. Escano-Quero, and J.F. Martin, “Adaptive method for SNR estimation in speech signal,” *Electronics Letters*, pp. 421-422, Feb. 1996.

[10] C. Avendano, H. Hermansky, M. Vis, and A. Bayya, “Adaptive speech enhancement using frequency-specific SNR estimates,” in *proc., Interactive Voice Technology for Telecommunications Applications Workshop,* pp. 65-68, Sept. 1996.

[11] J. B. Allen, and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *Journal, Acoustic Society American,* pp. 943-950, Apr. 1979.

[12] G. Lu, and T. Hankinson, “A technique towards automatic audio classification and retrieval,” in *Proc., International Conference Signal Processing*, Beijing, China, Vol. 2, pp. 1142–1145, Oct. 1998.