CLOSE/DISTANT TALKER DISCRIMINATION BASED ON KURTOSIS OF LINEAR PREDICTION RESIDUAL SIGNALS

Kohei Hayashida[1], *Masato Nakayama*[1], *Takanobu Nishiura*[1],
Yoichi Yamashita[1], *Toshiharu Horiuchi*[2], *and Tsuneo Kato*[2]

[1]Graduate School of Information Science and Engineering, Ritsumeikan University,
1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577, Japan
[2]User Interface Laboratory, KDDI R&D Laboratories, Inc.,
2-1-15 Ohara, Fujimino, Saitama, 356-8502, Japan

ABSTRACT

Desired/undesired speech discrimination is as important as speech/non-speech discrimination to achieve useful applications such as speech interfaces and teleconferencing systems. Conventional methods of voice activity detection (VAD) utilize the directional information of sound sources to distinguish desired from undesired speech. However, these methods have to utilize multiple microphones to estimate the directions of sound sources. Here, we propose a new method to discriminate desired from undesired speech with a single microphone. We assumed that the desired talkers would be close to the microphone, and the proposed method could distinguish close/distant-talking speech from observed signals based on the kurtosis of the linear prediction (LP) residual signals. The experimental results revealed that the proposed method could distinguish close-talking speech from distant-talking speech within a 10% equal error rate (EER) in ordinary reverberant environments with less processing time.

Index Terms— Close/distant talker discrimination, kurtosis, linear prediction, linear prediction residual signal

1. INTRODUCTION

Speech/non-speech discrimination is important for various kinds of applications such as speech interfaces and teleconferencing systems. Desired/undesired speech discrimination is as important as speech/non-speech discrimination for these applications because the speech spoken by undesired talkers becomes noise for applications and it disrupts normal operations.

Conventional single channel voice activity detection (VAD) [1, 2] can distinguish speech from non-speech. However, single channel VAD cannot distinguish desired from undesired speech. Conventional multiple channel VAD [3, 4] distinguishes desired from undesired speech based on direction of arrival (DOA) estimates. However, these methods require multiple microphones. Moreover, directions of microphones and talkers may not be the same at each use in teleconferencing systems. Therefore, it can be expected that desired/undesired speech discrimination method based on the distance between the talker and the microphone obtains a stable performance than that of the DOA estimation. We propose a new method to discriminate desired from undesired speech with a single microphone in this study based on the distance. We assumed that the desired talker would be within a certain distance from the microphone, and the undesired talker would be beyond the certain distance. Conventional methods of measuring the distance with multiple microphones [5, 6, 7] utilize phase differences between observed signals [5, 6] or direct-to-reverberant energy ratios [7]. Although these methods can estimate the distance between the talker and the microphone, multiple microphones are required. Further, although conventional methods of measuring the distance with a single microphone [8, 9] can estimate the distance based on reflected waves measured with reference to the transmitted waves, these methods also require a loudspeaker and emit known signals. We propose a new method of the discrimination based on the kurtosis of linear prediction (LP) residual signals for speech signals to discriminate close from distant talkers with only a single microphone. The kurtosis of the LP residual signals decreases depending on the distance between the talker and the microphone [10, 11]. The proposed method utilizes this measure to distinguish close/distant-talking speech.

This work was partly supported by Grants-in-Aid for Scientific Research funded by The Japanese Ministry of Education, Culture, Sports, Science and Technology.

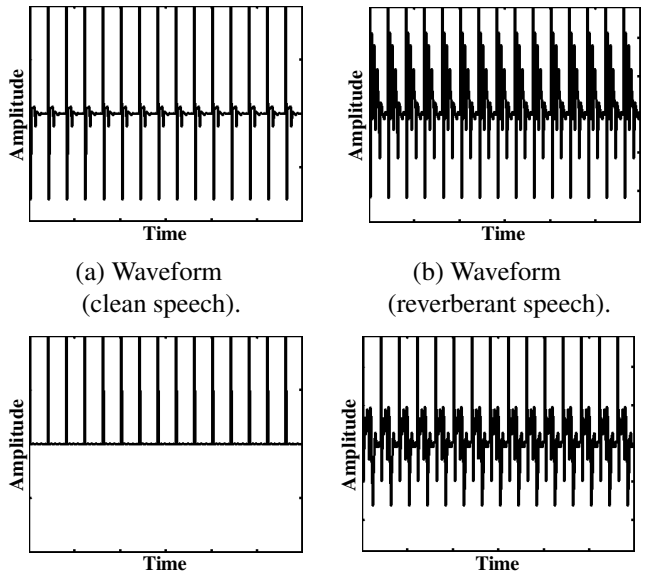
2. PROPOSED METHOD OF DISCRIMINATING CLOSE FROM DISTANT TALKERS BASED ON KURTOSIS OF LP RESIDUAL SIGNALS

LP residual signals of clean speech have strong peaks that correspond to pulses from the vocal cord, whereas those of reverberant speech have spreading peaks over time [10, 11].

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Figures 1 (a) and (b) show waveforms of clean speech and reverberant speech in the time domain. These signals were generated by simulation. Figures 1 (c) and (d) illustrate the LP residual signals of clean speech and reverberant speech. The clean speech signal in LP is modeled by pulses from the vocal cord and low-order finite impulse response (FIR) filters like those of the 10th order. Room reverberation is typically modeled by higher-order FIR filters like those of the order of thousands, and the LP residual signals of reverberant speech have spreading peaks over time. Therefore, the kurtosis of LP residual signals for reverberant speech is small, and that for clean speech is large. Hence, the kurtosis of LP residual signals is a reasonable measure of reverberation, and this has been conventionally utilized for speech dereverberation [10, 11].

The kurtosis of LP residual signals for close-talking speech is large in reverberant environments because of low reverberant distortion. The kurtosis of LP residual signals for distant-talking speech, on the other hand, is small because of high reverberant distortion. We applied this measure to close/distant talker discrimination in this study. The proposed method discriminated close and distant-talking speech from observed speech signals with a single microphone.



(a) Waveform (clean speech).

(b) Waveform (reverberant speech).

(c) LP residual signal (clean speech).

(d) LP residual signal (reverberant speech).

Fig. 1. Examples of LP analysis for simulated signals.

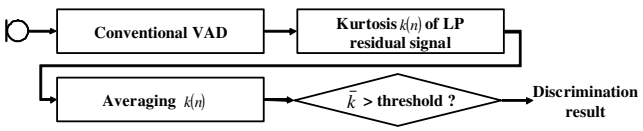


Fig. 2. Processing flow for proposed method.

2.1. Processing procedures

Figure 2 outlines the processing flow for the proposed method, which first extracts speech segment from observed signals with conventional VAD methods [1, 2]. The extracted signals are then weighted by a window function with a fixed frame length of:

$$\begin{aligned}\mathbf{x}(n) &= [x_1(n), \cdots, x_N(n)] \\ &= [s(nh - N + 1)w(1), \cdots, s(nh)w(N)]. \quad (1)\end{aligned}$$

The symbols $\mathbf{x}(n)$, $s(i)$, $w(i)$, n, h and N correspond to the windowed speech signals, the observed speech signal, the window function, the frame index in the time domain, the frame shift, and the frame length.

The LP value of the t-th windowed signal $\hat{x}_t(n)$ is defined as

$$\hat{x}_t(n) = -\sum_{i=1}^{p} a_i x_{t-i}(n), \quad (2)$$

where p is the number for the order of LP and a_i is the LP coefficient. LP residual signal $e_t(n)$ is calculated by:

$$e_t(n) = x_t(n) - \hat{x}_t(n). \quad (3)$$

The kurtosis of LP residual signal $k(n)$ is calculated by:

$$k(n) = \mathrm{E}\{e_t^4(n)\}/\mathrm{E}^2\{e_t^2(n)\} - 3. \quad (4)$$

The symbol, $\mathrm{E}\{\cdot\}$, denotes the expectation operator.

Averaged kurtosis is calculated by:

$$\bar{k} = \sum_{l=1}^{L} k(l)/L. \quad (5)$$

Symbol L denotes the number of analyzed frames. When the averaged kurtosis is larger than a threshold, the proposed approach determines that the observed speech is close-talking speech. Also, the observed speech is determined to be distant-talking speech when the averaged kurtosis is smaller than the threshold.

3. EVALUATION EXPERIMENTS

We evaluated the discriminating capabilities of the proposed method by varying the distance that separated close from distant-talking speech in four different reverberant environments.

Table 1. Clean speech signals for evaluation.

Sampling frequency	8 [kHz]
Quantization	16 [bit]
Number of speakers	104 (52 females, 52 males)
Total utterances	1001
Vocabulary	Eleven Japanese digits: “ichi,” “ni,” “san,” “yon,” “go,” “roku,” “nana,” “hachi,” “kyu,” “zero,” and “,maru.”

Table 2. Recording conditions in soundproof room.

Reverberation time	$T_{[60]} = 100$ [msec]
Size of room	2350 x 3250 x 2150 [mm]
Distance between microphone and speaker	100, 300, 500, 1000, 1200, 2000 [mm]
Distance between microphone and wall	250, 500, 1170 [mm]

Table 3. Recording conditions in laboratory.

Reverberation time	$T_{[60]} = 450$ [msec]
Size of room	3000 x 6000 x 2600 [mm]
Distance between microphone and speaker	100, 300, 500, 1000, 1200, 2000 2700, 3000, 4000 [mm]
Distance between microphone and wall	250, 500 [mm]

3.1. Experimental conditions

We evaluated the discriminating capabilities and the processing time for the proposed method with speech signals in real environments. We utilized clean speech signals in the Corpus and Environment for Noisy Speech RECognition (CENSREC-1-C) [12]. Table 1 summarizes the conditions for clean speech used in the evaluation. The clean speech consisted of one- to seven-digit utterances in Japanese. Each speaker made either nine or ten utterances. Recording was conducted in a soundproof booth using a head-set microphone (Sennheiser HMD25). The speech data were sampled at 16 kHz, quantized into 16-bit integers, and finally downsampled to 8 kHz. These signals had correct segmentations in speech periods that were manually prepared.

We measured room impulse responses in real environments. We recorded impulse responses in four kinds of reverberant environments such as a soundproof room, a laboratory, a conference room, and an elevator hall. Tables 2–5 list recording conditions in the four environments. We used a mouth simulator(Bruel & Kjaer Type 4227) as a loudspeaker to simulate the radiation characteristics of someone speaking to measure the impulse responses. A mouse simulator emitted sound sources toward the microphone. The impulse responses were sampled at 48 kHz, quantized into 16-bit integers, and finally downsampled to 8 kHz.

We designed evaluation signals by convolving the impulse responses with clean speech to simulate speech at various distances. Table 6 summarizes the analysis conditions we used in the evaluation. The proposed method calculated the kurtosis of LP residual signals from only the speech segments under these conditions. In other words, we evaluated the proposed method under conditions where its VAD worked optimally.

We evaluated the proposed method by using the false rejection rate(FRR) and the false acceptance rate(FAR). FRR and FAR are defined as:

$$\mathrm{FRR} = \frac{N_{\mathrm{FR}}}{N_{\mathrm{close}}} \times 100, \quad (6)$$

$$\mathrm{FAR} = \frac{N_{\mathrm{FA}}}{N_{\mathrm{dist}}} \times 100, \quad (7)$$

Table 4. Recording conditions in conference room.

Reverberation time	$T_{[60]} = 600$ [msec]
Size of room	8300 x 6800 x 2700 [mm]
Distance between microphone and speaker	100, 300, 500, 1000, 1200, 2000 2700, 3000, 4000, 5000 [mm]
Distance between microphone and wall	250, 500, 3350 [mm]

Table 5. Recording conditions in elevator hall.

Reverberation time	$T_{[60]} = 850$ [msec]
Size of room	9300 x 6300 x 2700 [mm]
Distance between microphone and speaker	100, 300, 500, 1000, 1200, 2000 2700, 3000, 4000, 5000 [mm]
Distance between microphone and wall	250, 500, 3000 [mm]

Table 6. Experimental conditions.

Frame length	512 [sample]
Frame shift	80 [sample]
Window function	Hamming
Number for order of LP	10
Threshold	$0 \sim 20$ in steps of 0.01

where $N_{\mathrm{close}}, N_{\mathrm{dist}}, N_{\mathrm{FR}}$, and N_{FA} correspond to the total number of close-talking utterances, the total number of distant-talking utterances, the number of close-talking utterances detected as distant-talking speech, and the number of distant-talking utterances detected as close-talking speech.

We also evaluated the processing time per frame with the proposed method. We measured the elapsed time of processing with the proposed method after conventional VAD. The measured times were then normalized by the number of processing frames, and finally averaged in all trials. We used a desktop PC that had a Core i5-2320 3.0 GHz CPU and 6 Gbytes of memory for the evaluation. The proposed method was implemented in C++.

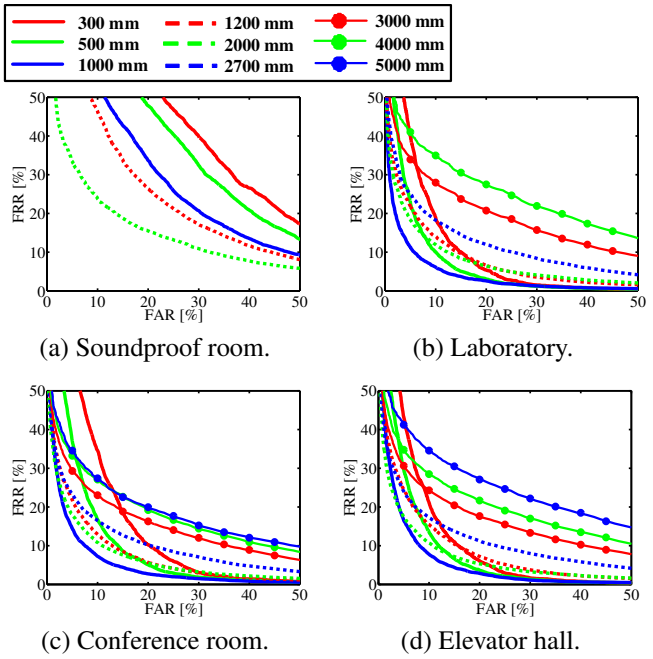
3.2. Experimental results

Figure 3 plots the FRR and FAR for the four environments, where each line indicates the boundary for separating close from distant-talking speech. Speech that was spoken under the boundary was defined as close-talking speech. Table 7 lists the ERRs of FRR and FAR. The discrimination capabilities of the proposed method in the soundproof room are worse in Fig. 3 (a) and Table 7 than those in the other three environments. The proposed method could distinguish close from distant-talking speech with less than 15% EER in environments other than the soundproof room when the boundary ranged from 500 to 2700 mm. The EER for the proposed method was less than 10% especially when the boundary was 1000 mm. These results indicated that the kurtosis of LP residual signals was an effective measure for distinguishing close and distant talkers in ordinary reverberant environments.

Table 8 summarizes the averaged processing time per frame for the proposed method. The processing time per

Table 7. EERs for each environment at each distance.

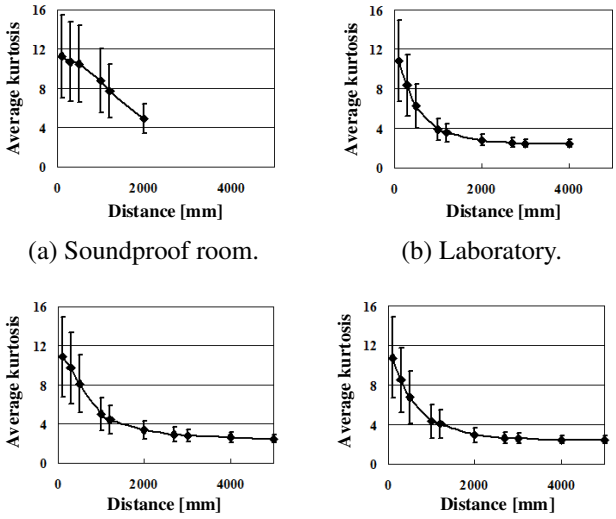
	300 mm	500 mm	1000 mm	1200 mm	2000 mm	2700 mm	3000 mm	4000 mm	5000 mm
Soundproof room	33.9 %	30.9 %	25.3 %	23.0 %	17.1 %	-	-	-	-
Laboratory	12.7 %	10.0 %	7.8 %	12.1 %	11.0 %	14.7 %	20.5 %	24.9 %	-
Conference room	16.1 %	12.2 %	8.7 %	11.4 %	10.5 %	13.7 %	17.5 %	19.6 %	19.8 %
Elevator hall	13.3 %	11.2 %	9.0 %	12.6 %	10.4 %	14.2 %	18.3 %	21.0 %	24.8 %



(a) Soundproof room. (b) Laboratory.

(c) Conference room. (d) Elevator hall.

Fig. 3. FRR and FAR for four environments.



(a) Soundproof room. (b) Laboratory.

(c) Conference room. (d) Elevator hall.

Fig. 4. Average kurtosis for each distance.

Table 8. Processing time per frame.

Proposed method	0.0297 [msec/frame]
Sampling interval	0.125 [msec/sample]

frame for the proposed approach is smaller than the sampling interval, meaning it could easily achieve real-time processing.

We concluded from these results that the new approach could accurately discriminate close from distant talkers with a single microphone in real time under ordinary reverberant environments.

3.3. Discussions

Figure 4 plots the average kurtosis at each distance in each environment, and the error bars indicate the standard deviation for average kurtosis. The decrease in average kurtosis that depends on distance in the soundproof room is less than that in the other three environments. This means that the decrease in the direct-to-reverberant ratio that depends on the distance in the lightly reverberant environment is less than that in the ordinary reverberant environments. Therefore, the FAR and FRR in the soundproof room was worse than that in the other three environments. These facts indicate that the proposed method made it difficult to distinguish close from distant talkers in lightly reverberant environments such as anechoic chambers and soundproof rooms.

The overall trend in Figs. 4 (b)–(d) suggests that the averages and the standard deviations of LP residual signals at each distance are similar values among ordinary reverberant environments. Moreover, the decrease of average kurtosis depending on distance is small beyond 1000–2000 mm. Hence, we could confirm that the proposed method could accurately distinguish whether the talker was within 1000–2000 mm from the microphone in ordinary reverberant environments.

4. CONCLUSION

Desired/undesired speech discrimination is required to construct a useful speech interface. We proposed a method of discriminating close and distant talkers in this study with a single microphone based on the kurtosis of LP residual signals. The results obtained from evaluation experiments indicated that the proposed approach could very accurately distinguish close and distant talkers in real time under ordinary reverberant environments. We intend to evaluate the new method in various kinds of noisy environments in future work. We will also try to determine a suitable threshold for the discrimination.

5. REFERENCES

[1] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, "ITU-T recommendation G.729 annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, pp. 64–73, 1997.

[2] ETSI Standard, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES 202 050 v1.1.5*, 2007.

[3] J.E. Rubio, K. Ishizuka, H. Sawada, S. Araki, T. Nakatani, and M. Fujimoto, "Two-microphone voice activity detection based on the homogeneity of the direction of arrival estimates," *Proc. ICASSP*, vol. IV, pp. 385–388, 2007.

[4] Y. Guo, K. Li, Q. Fu, and Y. Yan, "A two-microphone based voice activity detection for distant-talking speech in wide range of direction of arrival," *Proc. ICASSP*, pp. 4901–4904, 2012.

[5] J.M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filterling," *Robotics and Autonomous Systems*, vol. 55, pp. 216–228, 2007.

[6] A. Pourmohammad and S.M. Ahadi, "Real time high accuracy 3-D PHAT-based sound source localization using simple 4-microphone arrangement," *IEEE Systems Journal*, vol. 6, no. 3, pp. 455–468, 2012.

[7] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 8, pp. 2374–2384, 2011.

[8] M. Yang, S.L. Hill, B. Bury, and J.O. Gray, "A multifrequency AM-based ultrasonic system for accuracy distance measurement," *IEEE Trans. Instrum. Meas.*, vol. 43, no. 6, pp. 861–866, 1994.

[9] M. Nakayama, K. Suzuki, and N. Nakasako, "Acoustic distance measurement based on phase interference using the cross-spectral method with adjacement microphones," *Proc. ICASSP 2013*, pp. 423–427, 2013.

[10] B. Yegnanarayana and P.S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 267–281, 2000.

[11] B.W. Gillespie, H.S. Malvar, and D.A.F. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Proc. ICASSP*, vol. 1, pp. 3701–3704, 2001.

[12] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda, and S. Nakamura, "CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments," *Acoust. Sci. Technol.*, vol. 30, no. 5, pp. 363–371, 2009.