Voice activity detection in MTF-based power envelope restoration

Masashi Unoki[1], *Xugang Lu*[2], *Rico Petrick*[3],
Shota Morita[1], *Masato Akagi*[1], *Rüdiger Hoffmann*[3]

[1]School of Information Science, JAIST, Japan
[2]National Institute of Information and Communications Technology, Japan
[3]Laboratory of Acoustics and Speech Communication, Dresden University of Technology, Germany
[1]{unoki,s-morita,akagi}@jaist.ac.jp, [2]xugang.lu@nict.go.jp,
[3]{Rico.Petrick,Ruediger.Hoffmann}@ias.et.tu-dresden.de

10.21437/Interspeech.2011-667

Abstract

This paper reports comparative evaluations of conventional voice activity detection (VAD) methods in reverberant environments. Both conventional and standard (G.729) methods are discussed. In general, these methods work well under clean conditions, but their performance is drastically affected by reverberation. Preliminary comparative evaluations showed that the false acceptance rate (FAR) is significantly increased due to the false rejection rate (FRR) being moderately increased by reverberation. We therefore developed a method using MTF-based power envelope restoration to improve the robustness of VAD in reverberant environments. This restoration method can blindly restore the power envelope of reverberant speech based on the MTF concept. The proposed method consists of an MTF-based restoration method as the front end and a conventional VAD method as the final decision. Experimental results demonstrated that the proposed method is superior to conventional methods with regard to robustness and providing accurate VAD (reducing both FAR and FRR) in reverberant environments.

Index Terms: voice activity detection, reverberation, modulation transfer function, power envelope restoration

1. Introduction

Voice activity detection (VAD) is used to detect periods of speech and non-speech in observed signals. This is a key technology for various speech signal processes such as robust speech recognition systems, speech enhancement, and effective speech coding [1, 2]. The main challenge at present is ensuring whatever VAD technique we use is robust enough.

There have been many previous studies on robust VAD, and many methods/algorithms have been proposed over the last few decades [1, 2]. Although conventional features such as signal energy and zero-crossing rate are indeed effective under clean (noiseless) conditions, they are drastically smeared due to noise. Features based on periodicity/aperiodicity and higher order statistics are robustly effective under both clean and noisy environments [2]. However, another cause of disturbance is reverberation. Reverberation is independent of noise, there have not been any studies on robustness in reverberant conditions.

We can speculate that features based on signal energy and zero-crossing rate are also smeared due to reverberation, making them ineffective in terms of a robust VAD. Moreover, since none of the conventional methods for estimating fundamental frequency work well in reverberant environments [3], the useful feature of periodicity is not effective in reverberant conditions. Higher order statistics of speech signals are affected by reverberation, so speech enhancement techniques based on independent component analysis fail in these conditions. This is another important issue in terms of robust VAD.

In previous work, we developed a blind speech dereverberation method [4]. Our approach is based on the concept of the modulation transfer function (MTF) and does not require any measurements of room impulse responses (RIRs). MTF-based dereverberation can be used to restore the power envelope of reverberant speech. We are currently expanding this approach to include denoising and dereverberation techniques, i.e., speech enhancement in noisy reverberant environments. Reverberation has a consistent diffusion effect on the temporal power envelopes and we can remove this effect from the temporal power envelope with the MTF-based dereverberation. We therefore feel that MTF-based dereverberation can be applied in a front end manner for robust VAD to solve the issue.

In this paper, we investigate how well conventional VAD methods work in reverberant environments and determine how the errors due to reverberation occur. We then propose a robust VAD method based on MTF-based power envelope restoration to eliminate the error we found out.

2. VAD in reverberant environments

Let us consider an example of a typical VAD method in a reverberant environment. We used one clean utterance $x(t)$ (FAK speaker, 8-kHz sampling frequency, 16-bit quantization, spoken numbers from 1 to 9, 0 (maru and zero), and silence) from CENSREC-1-C (a Japanese continuous data corpus designed for testing VAD algorithms in noisy environments) [6] for the VAD test. A reverberant speech $y(t)$ was obtained by convolving the original speech $x(t)$ with the RIR $h(t)$ from SMILE2004 datasets [7]. The RIR used in this example is that of a concourse with a reverberation time of about 2 s.

We investigated the VAD of reverberant speech by using a conventional method (feature of signal energy, decision at the threshold of -30 dB from the normalized level). Figures 1(a) and (b) show the first 14 s of original speech $x(t)$ and reverberant speech $y(t)$. Dashed and dashdot lines indicate the speech periods of original and reverberant speech, respectively, detected by this method. The dashed line indicates the same speech periods as these obtained from CENSREC-1-C.

We then investigated the effect on VAD in the power envelope due to reverberation in the same condition. The power envelope $e_x^2(t)$ can be obtained as

$$e_x^2(t) = \mathbf{LPF}\left[|x(t) + j \cdot \mathbf{Hilbert}(x(t))|^2\right], \qquad (1)$$

where $\mathbf{Hilbert}()$ is the Hilbert transform and $\mathbf{LPF}[\cdot]$ is a low-pass filtering with a cut-off frequency of 20 Hz [4]. The power envelope of reverberant speech $e_y^2(t)$ can be also obtained by the same method from $y(t)$. Figures 1(c) and (d) show the power

Copyright © 2011 ISCA

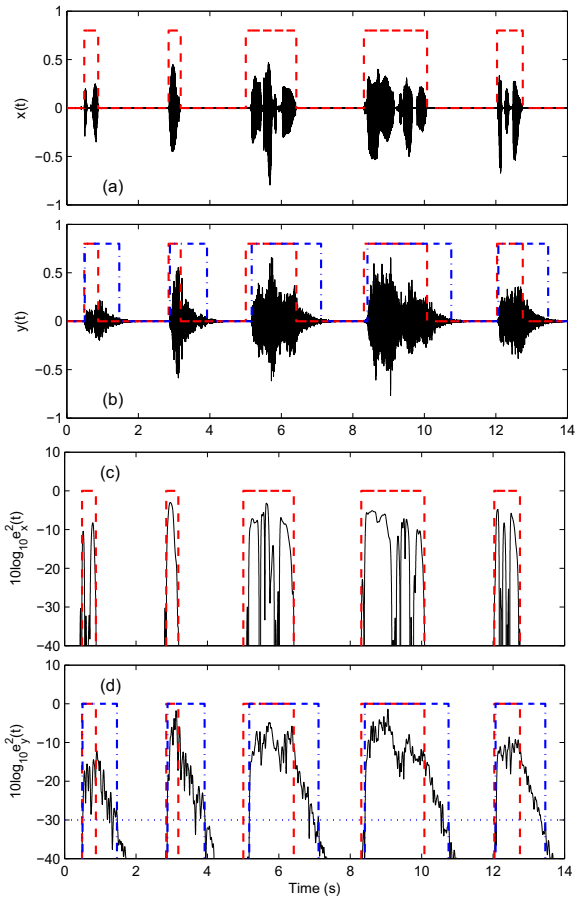

Figure 1: Example of VAD for reverberant speech: (a) original speech (first 14 s of FAK speech), (b) reverberant speech, (c) power envelope of (a), and (d) power envelope of (b).

envelope of the original $x(t)$ in Fig. 1(a) and the reverberant speech $y(t)$ in Fig. 1(b) (in decibels). The dashed and dashdot lines indicate the speech periods of these signals (in the same manner as in Fig. 1(b)).

In a general case, the VAD technique was evaluated in terms of its false rejection rate (FRR) and false acceptance rate (FAR). FRR indicates the rate of mis-judgments of non-speech periods in correct speech periods while FAR indicates the rate of mis-judgments of speech periods in correct non-speech periods. Both types of VAD had the same errors in non-speech periods caused by reverberation (Fig. 1), which suggests that the FAR was strongly affected by the reverberation while the FRR was not. This is just one result, but the same trend can occur in all stimuli with SMILE dataset RIRs.

3. VAD in MTF-based restoration

In the MTF-based power envelope restoration [4], $y(t)$, $x(t)$, and the stochastic-idealized RIR $h(t)$ in the room acoustics are modeled based on the MTF concept: $y(t) = x(t) * h(t)$, $x(t) = e_x(t)c_x(t)$, and $h(t) = e_h(t)c_h(t)$, where $e_h(t) = a\exp(-6.9t/T_R)$ and "*" denotes the convolution operation. Here, $e_x(t)$ and $e_h(t)$ are the envelopes of $x(t)$ and $h(t)$, and $c_x(t)$ and $c_h(t)$ are the mutually independent respective white noise functions (random variables) $< c_k(t), c_k(t-\tau) >= \delta(\tau)$. The parameters of the impulse response, a and T_R, are a constant amplitude term and the reverberation time, respectively. Based on this result, $e_x^2(t)$ can be recovered by deconvoluting $e_y^2(t)$ with $e_h^2(t)$. The transfer functions of power envelopes $E_x(z)$, $E_h(z)$, and $E_y(z)$ are then assumed to be the z-transforms of $e_x^2(t)$, $e_h^2(t)$, and $e_y^2(t)$, respectively. Here, $E_x(z)$, can be determined from

$$E_x(z) = \frac{1}{a^2}\left\{1 - \exp\left(-\frac{13.8}{T_R \cdot f_s}\right) z^{-1}\right\} E_y(z), \quad (2)$$

where f_s is the sampling frequency. Finally, $e_x^2(t)$ can be obtained from the inverse z-transform of $E_x(z)$[4]. Here, two parameters ($\hat{T_R}$ and $\hat{a}$) are obtained as

$$\hat{T}_R = \underset{0 \le T_R \le T_{R,\max}}{\arg\min} \left\{\frac{dT_P(T_R)}{dT_R}\right\}, \quad (3)$$

$$T_P(T_R) = \min\left(\underset{t_{\min} \le t \le t_{\max}}{\arg\min} \left|\hat{e}^2_{x,T_R}(t) - \theta\right|\right), \quad (4)$$

$$\hat{a} = \sqrt{1/\int_0^T \exp(-13.8t/\hat{T}_R)dt}. \quad (5)$$

We set θ to -20 dB from the maximum power envelope.

A block-diagram of the MTF-based method is shown in Fig. 2. Based on this processing, $e_y^2(t)$ can be reasonably restored as $\hat{e}_x^2(t)$ without measuring RIRs. Thus, it is possible to construct robust VAD in reverberant environments by combining conventional VAD with the MTF-based restoration as a front end. In the case shown in Figs. 1(c) and (d), $\hat{e}_x^2(t)$ was obtained as shown in Fig. 3. Then, the same conventional VAD (threshold of -30 dB for decision) was applied to detect speech periods indicated by the dashdot line in Fig. 3. The effectiveness of the proposed VAD can be clearly observed by comparing Figs. 1(d) and 3. In this case, we found that FAR can be drastically reduced while FRR is only a little reduced.

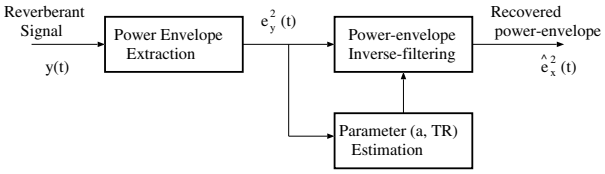

Figure 2: Block diagram of power envelope inverse filtering.

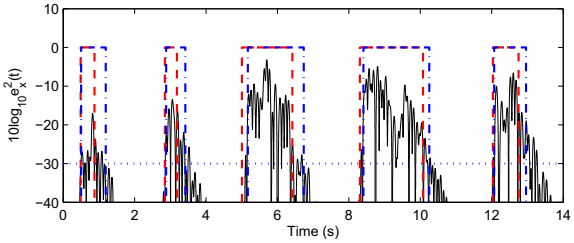

Figure 3: VAD in the restored power envelope.

4. Comparative evaluations

We evaluated five VAD methods to evaluate how robust the detection of speech/non-speech periods was in artificial reverberant environments. These methods were the VAD in G.729 [9], the conventional VAD (thresholds of signal energy and power envelope) we previously described, and the proposed method. Another method, conventional VAD with cepstrum-mean-normalization (CMN) as a front end, was also used to

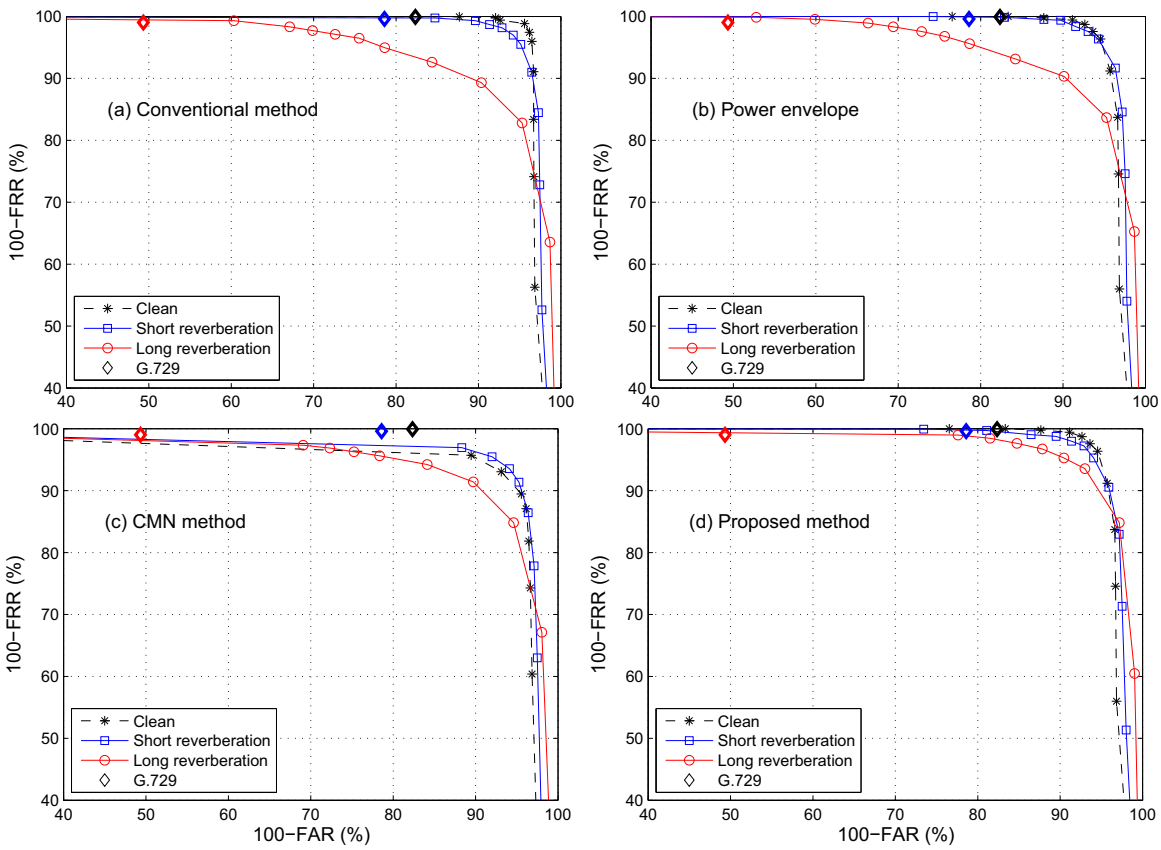

Figure 4: ROC curves of the VAD algorithms for clean and short and long reverberation: (a) conventional method with signal, (b) conventional method with power envelope, (c) conventional method with CMN, and (d) proposed method.

compare the effectiveness of MTF-based restoration with CMN as the front end.

The speech signals we used were three Japanese sentences (/aikawarazu/, /shinbun/, and /joudan/) uttered by ten speakers (five males: Mau, Mht, Mnm, Mtm, and Mtt, and five females: Faf, Ffs, Fkn, Fsu, and Fyn) from the ATR database [8]. We used 100 types of RIR $h(t)$ and ten reverberation times (TRs) ($T_R = 0.1, 0.2, 0.3, 0.4, 0.5, 1.0, 1.5, 2.0, 2.5,$ and 3.0). All stimuli $y(t)$s were composed through $30,000$ ($= 3 \times 10 \times 10 \times 100$) convolutions of $x(t)$s with $h(t)$s.

To measure VAD performance, we used FRR and FAR, defined as $\mathrm{FRR} = N_{\mathrm{FR}}/N_{\mathrm{s}} \times 100$ (%) and $\mathrm{FAR} = N_{\mathrm{FA}}/N_{\mathrm{ns}} \times 100$ (%), where N_{s}, N_{ns}, N_{FR}, and N_{FA} are the total number of speech frames, total number of non-speech frames, number of speech frames detected as non-speech frames, and number of non-speech frames detected as speech frames, respectively. We measured these by varying the threshold and then averaged the results for three categories: clean, short reverberation (0.1 to 0.5 s), and long reverberation (1.0 to 3.0 s). The final performance evaluation was represented as a receiver operating characteristics (ROC) curve.

Figure 4 shows the ROC curves of all methods except VAD in G.729. For comparison, the G.729 result is plotted in all panels as a diamond symbol. In ROC curves, an ideal performance shows an edged shape located on the top-right ROC curve. In each panel, there are three lines that indicate clean condition, short reverberation, and long reverberation. In the clean condition, all methods worked well. However, in reverberant conditions, the conventional methods performed poorly (Figs. 4(a) and (b)). The shape of the ROC curves became especially broad-edged toward the bottom-left location. In the CMN-method, the ROC curves in reverberation conditions were improved, particularly when the reverberations were short. Three ROC curves in the proposed method were almost identical shapes which is the best result of all the methods.

Next, we evaluated the five VAD methods in actual reverberation environments. We used the same speech stimuli as in the first evaluation with 43-RIRs from the SMILE datasets [7]. The RIR conditions are shown in the left column in Tab. 1, and pairs of FAR/FRR in the five methods are listed in the right. Bold and italic fonts indicate best and worst VAD results, respectively. In most cases, the proposed method had FARs with the best results, and it had FRRs that were better than those of the conventional methods with the exception of G.729. Although G.729 had the best FRRs, it also had the worst FARs. It is suggested that G.729 with the proposed methods may have a potential of the most robust VAD in realistic reverberant environments. The results of two evaluations demonstrate that conventional VAD with the MTF-based restoration can significantly improve robustness in reverberant environments.

5. Conclusion

In this paper, we reported that comparative evaluations of conventional VAD methods in reverberant environments. Results showed that FAR is significantly increased due to FRR being moderately increased by the reverberation. We then proposed a method based on MTF-based power envelope restoration that improves the robustness of VAD in reverberant environments. The proposed method consists of an MTF-based restoration

Table 1: Comparison of VAD performance, FAR/FRR (%), in actual reverberant environments. The reverberation time T_R was determined as an average from each T_R on the transfer function at 125 to 8 kHz in octave frequency. Bold and italic fonts indicate the best and worst results, respectively. Threshold of conventional methods was set as -30 dB because this was an optimal value in Fig. 4.

Room (Impulse response)	IRdata	T_R (s)	G.729	Conventional	Power Env.	CMN	Proposed
Lecture room with flatter echo	201	1.36	*55.4*/1.57	33.8/4.21	33.6/3.13	**13.7**/*4.64*	15.8/**1.32**
Multipurpose hall 1 with reflex board	301	1.09	*74.0*/**3.35**	20.4/*8.63*	20.1/7.48	15.4/8.16	**10.4**/5.32
Multipurpose hall 1 without reflex board	302	0.80	*68.3*/**3.40**	19.0/7.44	19.3/5.97	11.3/*7.65*	**8.72**/3.70
Multipurpose hall 2 with reflex board	303	1.44	*77.6*/**3.61**	29.9/*8.78*	30.0/7.21	14.8/6.91	**14.7**/3.62
Multipurpose hall 2 without reflex board	304	1.04	*71.7*/3.12	23.6/*6.57*	23.6/4.73	13.6/6.41	**11.0/2.66**
Multipurpose hall 3 with reflex board	305	1.93	*72.1*/**4.78**	29.9/*11.4*	30.1/9.53	19.0/9.35	**17.5**/5.96
Multipurpose hall 3 without reflex board	306	1.35	*77.4*/**4.55**	23.0/*9.96*	22.6/7.96	15.1/9.03	**12.5**/5.40
Multi purpose hall 4 with reflex board	307	1.42	*51.6*/2.90	22.9/6.67	22.7/4.88	9.96/*7.12*	**9.94/2.60**
Multi purpose hall 4 without reflex board	308	1.54	*52.6*/2.98	22.54/6.67	23.0/4.91	10.8/*7.05*	**10.5/2.74**
Classic concert hall 1	309	2.35	*64.4*/**3.26**	32.1/*8.33*	32.1/6.62	**19.1**/7.25	19.4/4.17
Classic concert hall 1 ($d = 6$ m)	310	2.34	*64.6*/**1.91**	31.3/*6.78*	31.5/4.91	18.2/5.71	**17.7**/2.30
Classic concert hall 1 ($d = 11$ m)	311	2.35	*64.0*/**3.26**	32.0/*8.32*	32.0/6.62	**19.0**/7.25	19.3/4.17
Classic concert hall 1 ($d = 15$ m)	312	2.39	*63.7*/**4.75**	32.3/*10.6*	32.3/9.11	19.6/9.43	**19.4**/5.43
Classic concert hall 1 ($d = 19$ m)	313	2.38	*62.6*/**6.50**	34.2/*13.0*	34.2/12.1	21.5/11.4	**21.5**/8.45
Classic concert hall 2	314	1.14	*50.9*/**4.28**	13.8/*10.0*	13.7/8.59	4.97/8.48	**4.50**/5.27
Classic concert hall 3	315	1.96	*76.5*/**2.16**	27.0/*8.39*	26.3/6.23	26.6/5.96	**18.7**/4.13
Classic concert hall 4 with	316	1.92	*59.1*/3.33	30.7/*7.62*	30.8/6.15	**11.7**/7.43	12.9/**2.54**
Classic concert hall 4 without	317	2.55	*69.1*/3.39	37.2/*7.73*	37.2/6.29	**12.2**/7.43	16.1/**2.34**
Theater hall	318	0.85	*70.5*/**3.30**	17.6/*7.58*	18.0/6.10	14.1/7.17	**9.88**/4.05
Multipurpose hall 5	319	1.47	*52.8*/**2.45**	25.9/*7.14*	25.8/5.24	12.4/6.50	**11.8**/3.12
Multipurpose hall 6	321	2.16	*42.2*/**2.44**	27.7/*7.01*	27.6/5.65	16.0/6.49	**15.8**/2.77
Classic concert hall 5	323	2.32	*40.8*/3.28	28.5/*7.89*	28.8/5.97	**16.9**/6.74	18.0/**3.01**
Classic concert hall 6 (1F front)	324	1.77	*47.2*/**3.29**	24.7/*8.21*	24.9/6.05	15.2/7.07	**12.8**/4.01
Classic concert hall 6 (2F side)	325	1.74	*48.0*/**2.46**	22.1/*8.39*	22.0/6.29	16.3/6.59	**13.5**/3.06
Classic concert hall 6 (3F)	326	1.69	*48.5*/**9.41**	29.8/*16.5*	30.2/15.2	20.2/14.7	**18.4**/10.6
Meeting room	401	0.62	*66.5*/**0.84**	10.9/2.50	11.1/1.73	7.27/*5.01*	**6.47**/2.82
Lecture room (capacity: 400 m^3)	402	1.12	*75.2*/1.22	24.3/3.57	23.4/2.82	**7.58**/*4.27*	11.8/**1.01**
Lecture room (capacity: 2, 400 m^3)	403	1.09	*74.8*/3.17	33.5/*7.97*	33.6/6.19	**7.71**/7.41	17.5/**2.63**
General speech hall (capacity: 11, 000 m^3)	404	1.54	*79.4*/**3.94**	28.9/*9.19*	29.0/7.08	18.0/8.23	**15.2**/4.39
Church 1 (capacity: 1, 200 m^3)	405	0.71	*67.3*/**3.08**	13.3/*8.36*	13.4/6.43	11.3/7.62	**7.31**/4.89
Church 2 (capacity: 3, 200 m^3)	406	1.30	*76.5*/**4.29**	25.3/*9.13*	25.7/7.75	18.9/8.55	**15.3**/5.14
Event hall 1 (capacity: 28, 000 m^3)	407	3.03	*70.2*/**6.39**	46.3/*14.4*	45.6/13.2	**19.4**/13.2	20.5/9.71
Event hall 2 (capacity: 41, 000 m^3)	408	3.62	*61.3*/**6.63**	48.1/*15.0*	48.2/13.5	**18.9**/14.0	23.7/10.3
Gym 1 (capacity: 12, 000 m^3)	409	2.82	*52.6*/**3.46**	34.2/*15.2*	34.2/12.5	25.2/11.1	**22.5**/9.67
Gym 2 (capacity: 29, 000 m^3)	410	1.70	*79.6*/**6.47**	30.0/*13.4*	29.4/11.5	20.3/12.5	**17.4**/7.63
Living room in wooden house	411	0.36	*52.1*/**0.42**	7.90/1.94	9.51/1.59	**4.35**/*7.84*	5.30/7.68
Movie theater	412	0.38	*57.7*/**2.25**	8.34/9.92	8.61/7.54	5.41/*10.2*	**4.10**/7.55
antrium	413	1.57	*81.1*/**1.72**	28.1/*10.6*	26.9/9.40	17.0/8.32	**14.7**/4.63
Tunnel	414	2.72	40.2/**4.72**	35.5/*16.2*	35.7/13.5	28.9/10.2	**29.1**/7.78
Concourse in train station	415	1.95	*82.2*/**9.62**	51.9/*14.3*	48.3/12.6	12.6/*14.3*	**11.9**/10.1
General speech hall 2 (1F front)	416	1.53	*53.1*/**2.98**	28.8/*8.10*	29.0/6.33	11.7/6.49	**11.0**/3.24
General speech hall 2 (1F central)	417	1.49	*53.5*/**4.43**	29.6/*11.5*	30.0/9.32	13.1/6.49	**12.7**/4.95
General speech hall 2 (1F balcony)	418	1.40	*54.8*/**7.69**	29.4/*16.6*	29.0/14.6	13.9/14.4	**13.0**/10.3

method as the front end and a conventional VAD method as the final decision. Comparative evaluations demonstrated that the proposed method was much better than conventional ones in terms of robustness and providing accurate VAD (reducing both FAR and FRR) in reverberant environments. In future work, we plan to further develop our method to ensure even more robust VAD in noisy reverberant environments [5].

6. Acknowledgements

The work was partially supported by the Research Foundation for the Electrotechnology of Chubu (REFEC).

7. References

[1] Ramirez, J., Gorriz, J. M., Segura, J. C. "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness," Robust Speech Recognition and Understanding, 1–22, 2007.

[2] Ishizuka, K., Fujimoto, M., and Nakatani, T. "Advances in voice activity detection," *J. Acoust. Soc. Jpn.*, **65**(10), 537–543, 2009.

[3] Unoki, M., Hosorogiya, T., and Ishimoto, Y. "Comparative evaluations of robust and accurate F0 estimation in reverberant environments," *Proc. ICASSP 2008*, pp. 4569–4572, 2008.

[4] Unoki, M., Furukawa, M., Sakata, K., and Akagi, M. "An improved method based on the MTF concept for restoring the power envelope from a reverberant signal," *Acoust. Sci. & Tech.,* **25**(4), 232–242, 2004.

[5] Unoki, M., Yamasaki, Y., and Akagi, M. "MTF-based power envelope restoration in noisy reverberant environments," *Proc. EUSIPCO2009*, 228–232, 2009.

[6] http://sp.shinshu-u.ac.jp/CENSREC/en/CENSREC/CENSREC-1-C/.

[7] Sound Material in Living Environment, Architectual Institute of Japan and GIHODO SHUPPAN Co., Ltd., Tokyo, 2004.

[8] Takeda, K. *et al.*, Speech Database User's Manual, ATR Technical Report TR-I-0028, 1988.

[9] ITU-T, A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70, Recommendation G.729 Annex B, 1996.