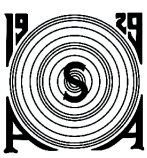

euronoise

An MTF-based blind restoration of temporal power envelopes as a front-end processor for automatic speech recognition systems in reverberant environments

Xugang Lu[a], Masashi Unoki[b] and Masato Akagi[a]

[a]Japan Advanced Institute of Science and Technology, 1-1, Asahidai, Nomi, 923-1292 Sendai, Japan
[b]JAIST, 1-1 Asahidai, 923-1292 Nomi, Japan
unoki@jaist.ac.jp

Abstract

To reduce speech degradation in reverberant environments, we previously proposed a modulation transfer function (MTF) based method of speech restoration. The room impulse response (RIR) in this restoration does not need to be measured at any time since we modeled the power envelope of the RIRs as an exponential decay function. Speech is assumed to be temporal modulated with white noise carrier in all sub-bands. We have tested how effective this method was as a front-end for ASR systems in artificial and real reverberant environments. Reverberant speech signals were created by simply convoluting clean speech signal (AURORA-2J database) with the artificially produced or real RIRs. A method based on the auditory power spectrum was used as a baseline for comparison. Compared with the baseline, the proposed method for artificial reverberant environments produced a 35.67% relative improvement in the error reduction rate (on average, for reverberation times from 0.2 to 2.0 s), and for real reverberant environments (43 RIRs), it produced a 25.78% relative improvement in the error reduction rate. The results demonstrate that our new approach can improve the robustness of ASR systems in reverberant environments, and it performs better than conventional methods.

1 Introduction

It is well known that reverberation smears significant features of speech so that recognition rates (RRs) of automatic speech recognition (ASR) systems are drastically reduced as reverberation time (RT) increases [1, 2, 3]. Achieving robust speech recognition in reverberant environment (RE) is therefore an important issue.

Reverberation can be regarded as convolution processing of speech signal and room acoustics. In a RE, the temporal and spectral structure of speech is distorted by room reverberation characteristics. It is difficult to distinguish clean speech signals in a RE by using the statistical properties of the original and of the reverberant speech signals. Although the traditional methods for reducing additive noise based on the statistical properties such as spectral subtraction have widely used to enhance smeared speech, these do not work well in REs.

Several algorithms for reducing reverberation distortion have been proposed. The two most well known are cepstral mean normalization (CMN) [4] and RASTA filtering [5]. These can effectively reduce the distortions caused by short-term convolution channels. In real room acoustics, the RT is far longer, and the properties of REs are that they are both time and spatially variant.

Several dereverberation algorithms using single or multi microphones have been proposed for solving the room-reverberation problem. The basic principle of dereverberation has been to measure the room impulse response (RIR), and then use inverse filtering to obtain dereverberated speech [6]. However, these methods require the RIRs for each dereverberation to be remeasured if room acoustics have changed. One possible way to utilize blind speech dereverberation is to use speech characteristics. For example, the harmonic structure of speech can be used [7]. This method needs the fundamental frequency from reverberant speech to be accurately estimated, which is difficult [8], and it does not seem to restore the consonant parts in speech.

In this study, we utilized the characteristics of speech and the RIRs for speech dereverberation. Speech signals are highly temporally modulated, and most of their intelligibility information is encoded in temporal modulation envelopes in each sub-band [9]. This means that we need to restore the temporal modulated envelope of clean speech from the reverberant speech for recognition. We previously proposed a sub-band power envelope inverse filtering algorithm based on the modulation transfer function (MTF) [10, 11] for dereverberating speech signals [12, 13]. It was designed to be used as a front-end processor for automatic speech recognition. Correlation and SNR measurements showed that it improves power envelope restoration accuracy [12, 13], and testing showed that it restores speech signals with a high level of speech intelligibility [14]. We have now tested its ability to recognize Japanese digital speech in both artificial REs [15] and real REs [16].

2 MTF-based sub-band power envelope restoration method

2.1 Model concept

The MTF concept was proposed by Houtgast and Steeneken [10] to account for the relationship between the transfer function in an enclosure in terms of input and output signal envelopes and the characteristics of the enclosure such as RE. This concept was introduced as a measure in RIRs for assessing the effect of enclosure on speech intelligibility [10]. The MTF is obtained [10, 11] as

$$M(\omega) = \left| \frac{\int_0^\infty h^2(t) e^{j\omega t} dt}{\int_0^\infty h^2(t) dt} \right| = \left[1 + \left(\omega \frac{T_R}{13.8} \right)^2 \right]^{-\frac{1}{2}}, \quad (1)$$

where ω is the radian frequency and the RIR, $h(t)$, is

$$h(t) = e_h(t)\mathbf{n}_1(t) = a \exp(-6.9t/T_R)\mathbf{n}_1(t), \quad (2)$$

where $e_h(t)$ is the exponential decay temporal envelope, a is a constant amplitude, T_R is the RT defined as T_{60}, and $\mathbf{n}_1(t)$ is a random white noise as a random variable. Eq. (2) is a well-known stochastic approximation of the RIR. For a dominant frequency in the temporal envelope, Eq. (2) can be regarded as the modulation index, i.e., the degree of the relative fluctuation in the normalized amplitude with respect to the modulation frequency. On the basis of this characteristic, T_R can be predicted from a specific frequency by using the MTF.

We model what effect of room acoustics had on speech signals on the MTF concept. The convolution distortion in sub-band representation is written as

$$y_n(t) = x_n(t) * h(t), \qquad n = 1, 2, \cdots, N, \quad (3)$$

where $y_n(t)$ and $x_n(t)$ correspond to the reverberant and clean speech signals in the sub-band, n is the sub-band index, and N is the total number of sub-bands. Using the temporal modulation properties of the speech signal, we model the sub-band speech, $x_n(t)$, as

$$x_n(t) = e_{x,n}(t)\mathbf{n}_2(t). \quad (4)$$

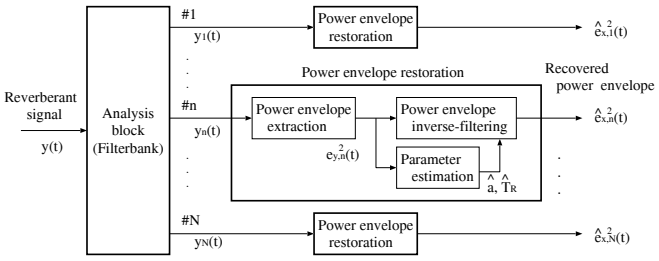

Figure 1: Sub-band power envelope method of inverse filtering based on the MTF concept.

The temporal envelope of sub-band n is $e_{x,n}(t)$. Here, $\mathbf{n}_1(t)$ and $\mathbf{n}_2(t)$ are mutually independent random variables that satisfies $\langle \mathbf{n}_k(t)\mathbf{n}_k(t-\tau)\rangle = \delta(\tau), k = 1, 2$, where $< \cdot >$ is the ensemble average operator. From these, we can calculate the power envelope of $y_n(t)$ as

$$\left\langle y_n^2(t)\right\rangle = e_{y,n}^2(t) = e_{x,n}^2(t) * e_h^2(t). \tag{5}$$

This equation shows that the restoration of $e_{x,n}^2(t)$ can be completed by deconvolution of $e_{y,n}^2(t)$ with $e_h^2(t)$. These signals are transformed a continuous signal into a discrete signal on the basis of sampling theorems, such as $e_{x,n}^2[m]$, $e_{y,n}^2[m]$, and $e_h^2[m]$ (m is the number of samples). $E_{x,n}(z)$, $E_{y,n}(z)$, and $E_h(z)$ are the z-transforms of $e_{x,n}^2[m]$, $e_{y,n}^2[m]$, and $e_h^2[m]$. The input-output relationship for deconvolution can be represented as

$$E_{x,n}(z) = \frac{E_{y,n}(z)}{a^2}\left\{1 - e^{-\frac{13.8}{T_{R,n}\cdot f_s}} z^{-1}\right\}, \tag{6}$$

where f_s is the sampling frequency. $e_{x,n}^2[m]$ can be restored using the inverse z-transform of $E_{x,n}(z)$. In Eq. (6), we only need to estimate parameters $T_{R,n}$ and a. Here, the parameter, $T_{R,n}$, is assumed to be a function of n since it is dependent on the sub-band, and is independently estimated from each sub-band.

2.2 Algorithm implementation

The algorithm for inverse filtering the sub-band power envelope was developed on the basis of the analysis above. The processing scheme for inverse filtering the sub-band power envelope is outlined in Fig. 1. In the processing scheme, observed signal $y(t)$ is decomposed into a series of frequency sub-bands; envelope detectors then extract temporal modulation envelopes $e_{y,n}^2(t)$. Considering the co-modulation characteristics of speech signals in sub-bands [12], we deliberately designed a series of FIR-type band-pass filters with a constant bandwidth (100 Hz was chosen in this study) for the decomposition. Thus, this filterbank is referred to as a constant-bandwidth filterbank (CBFB) in this paper. The extracted envelopes are used for inverse filtering $e_{y,n}^2(t)$, which is controlled by estimated parameters $T_{R,n}$ and a. The final output is the restored or dereverberated power envelope, $\hat{e}_{x,n}^2(t)$, for all sub-bands. The implementation is as follows.

2.2.1 Sub-band power envelope extraction

The power envelopes in the sub-bands are extracted by low-pass filtering the Hilbert transform of the sub-band signals [12, 13]:

$$\hat{e}_{y,n}(t)^2 = \mathrm{LPF}\left[\left|y_n(t) + j\mathrm{Hilb}(y_n(t))\right|^2\right], \tag{7}$$

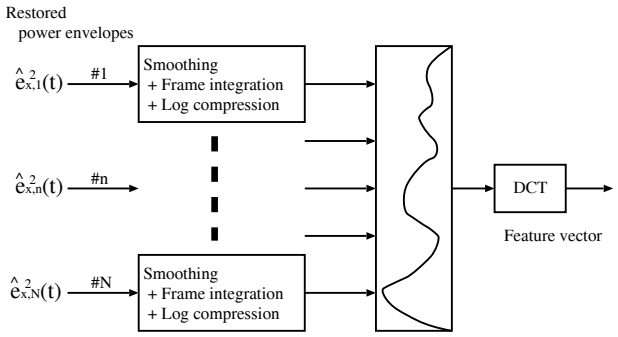

Figure 2: Extraction of speech features based on restored power envelope in all sub-bands.

where $\mathbf{LPF}[\cdot]$ is a low-pass filtering operator, and $\mathbf{Hilb}(\cdot)$ is the Hilbert transform. We set the cut-off frequency of the low pass filtering to 20 Hz to retain most of the important modulation information for speech perception.

2.2.2 Parameter estimation

The $T_{R,n}$ and a (referred to as $\hat{T}_{R,n}$ and $\hat{a}$) in Eq. (6) are estimated using Unoki *et al.*'s formulas [12]

$$\hat{T}_{R,n} = \max\left(\arg\min_{T_{R,n}} \int_0^T \left|\min\left(\hat{e}_{x,n,T_{R,n}}^2(t), 0\right)\right| dt\right), \tag{8}$$

$$\hat{a} = \sqrt{1/\int_0^T e^{-\frac{13.8t}{\hat{T}_{R,n}}} dt}, \tag{9}$$

where T is signal duration and $\hat{e}_{x,n,T_{R,n}}^2(t)$ represents the candidates of the restored power envelope as a function of $T_{R,n}$. The RT is constrained as $T_{R,\min} < T_{R,n} < T_{R,\max}$. These are the lower and upper bounds of $T_{R,n}$.

2.2.3 Power envelope inverse filtering

After the power envelopes ($e_{y,n}^2(t)$) and the parameters of the RIR ($\hat{T}_R$ and $\hat{a}$) are obtained, the power envelopes are inverse filtered using Eq. (6) to restore the power envelopes of the dereverberated speech in the sub-bands ($e_{x,n}^2(t)$). Here, the restored power envelope in a sub-band is denoted as $\hat{e}_{x,n}^2(t)$.

3 ASR for reverberant speech

3.1 Feature extraction

We tested the effectiveness of the proposed algorithm for dereverberation as a front-end processor for ASR of reverberant speech. We used clean speech from the AURORA-2J database as speech material [17], and used 8,840 clean speech sentences to train the acoustic models. We used 1,001 clean speech sentences to produce reverberant speech to test recognition in REs by convolving the speech signals with the RIRs. We used the $f_s = 8$ kHz and 40 sub-band channels ($N = 40$) to cover the frequency region from 0 to 4 kHz.

In Fig. 2, the first block for a smoothing and it is comprised of frame integration and log compression. Because the inverse filtering of power envelope is a high-pass, low-pass filtering with a forgotten parameter, λ

was used to smooth the envelope dips:

$$\bar{e}_{x,n}[m] = \lambda \bar{e}_{x,n}[m-1] + (1-\lambda)\hat{e}_{x,n}[m], \qquad (10)$$

where $\hat{e}_{x,n}[m]$ is the original restored sub-band power envelope, and $\bar{e}_{x,n}[m]$ is the smoothed output. In this paper, we set λ to 0.98. To integrate the frames, we used a 32-ms frame length with a Hamming window and a frame rate of 16-ms. After the integrated spectrum was obtained, log compression was carried out. The DCT was used for dimensional decorrelation. The first 12 dimensions of the decorrelated log power spectrum were used. Combining the log power energy, we obtained 13-dimensional static feature sets. Together with their first and second order delta dynamic values, 39-dimensional feature vectors were formed. HTK [18] was used for training the HMM acoustic models. The acoustic models were configured the same as in the AURORA-2J experiments [17].

For comparison, we also tested the performance of conventional methods of feature extraction under the same conditions. One is to use the standard feature representation, i.e., the Mel Frequency Cepstral Coefficient (MFCC) representation. Another is to use the auditory cepstral feature vector on the basis of the sub-band power envelopes. In this representation, a gammatone auditory filterbank an equivalent rectangular bandwidth (ERB) was used. This can be regarded as a constant-Q filterbank (CQFB) so that this feature vector is referred to as CQFB in this paper. Two conventional post-processing methods, i.e., CMN [4] and RASTA filtering [5], were also used in this paper to deal with convolution distortion. Consequently, the features extracted are denoted here as Fea_CMN, Fea_RASTA, and Fea_IMTF (where "Fea" is either CQFB or CBFB).

3.2 Recognition experiments in artificial REs

We tested the recognition of our proposed method in artificial REs by using 1,001 clean speech sentences to produce reverberant speech. The speech signals were convolved with the artificial RIRs (produced using Eq. (2)) with RT of 0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8 or 2.0 s. In total, we used 1,001 clean speech signals and 10,010 (= 1,001 × 10) reverberant speech signals.

We simulated speech recognition using many types of feature: CQFB, CBFB, CBFB_CMN, CBFB_RASTA, and CBFB_IMTF. The recognition for short RTs ($T_R <$ 0.2 s) was best as can seen from the magnified plot in Fig. 3b. The recognition rate (RR) decreased as the RT increased; the rate of decrease was especially high when the RT was longer ($T_R > 0.2$ s). When $T_R < 0.2$ s, all the features performed well. The CBFB-based feature, performed better for $T_R > 0.15$ s, and a little worse for $T_R < 0.15$ s than the CQFB. Also as shown in Fig. 3, CBFB_RASTA performed even worse than CBFB alone. The CBFB_CMN performed a little worse or almost the same as the CBFB alone (except for $T_R < 0.2$ s). However, the CBFB_IMTF consistently improved the performance of CBFB alone. We also tested how well CQFB-based feature performed. They performed worse than using CBFB alone. Consequently, we did not use CQFB-based feature for comparison in later experiments. In addition, we found that adding CMN or RASTA processing to either CQFB or CBFB did not improve the RR in our experiments, and sometimes even decreased performance. Therefore, in this paper, the performance of CQFB is used as a baseline.

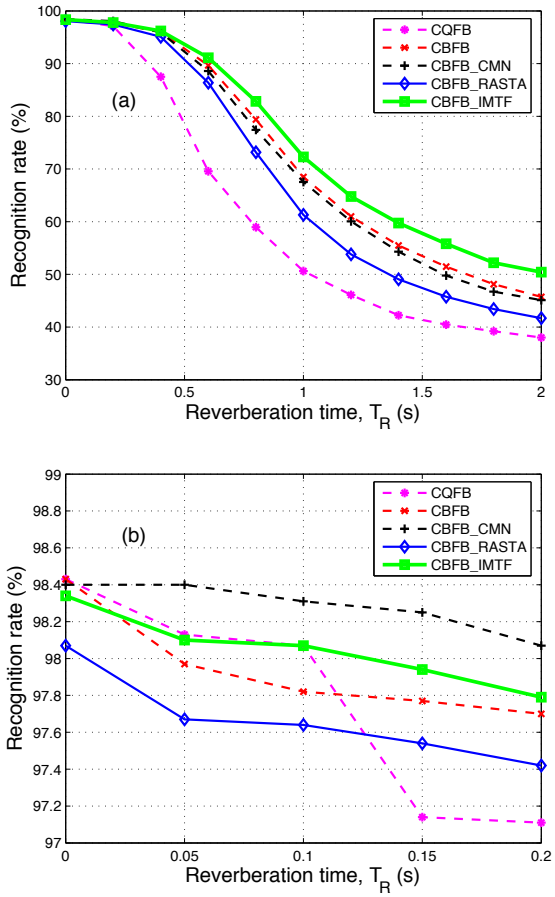

Figure 3: Comparative evaluations of reverberant speech recognition rates: (a) whole evaluation and (b) zoom-up plot in range from 0.0 to 0.2 s.

A relative improvement (RI) in the RR was adopted to show the improvement in recognition in different REs from different baselines, which is defined as

$$\mathrm{RI} = \frac{(\mathrm{TRR} - \mathrm{BRR})}{(1 - \mathrm{BRR})} \times 100 \qquad (\%), \qquad (11)$$

where "TRR" and "BRR" denote the testing RR and baseline RR. Based on this definition, the proposed CBFB and CBFB_IMTF derived RIs of 28.64% and 35.67% on average (for 0.2 s $< T_R <$ 2 s) in the error reduction rate compared with CQFB.

3.3 Recognition experiments in real REs

We then tested our method on speech recognition in many real REs (e.g. 43 halls, rooms, and theaters [16]) under various conditions. The reverberant speech signals were obtained from the convolutions between clean speech signals and the RIRs (the sampling rate of the impulse responses of the environments were sub-sampled to 8 kHz to adapt to the sampling rate of the speech database). The speech corpus, features, and acoustic models were the same as those used for the artificial

REs described in previous subsection. The characteristics of the REs and the RRs are listed in Table 1.

Rooms and halls constructed of different materials and with different configurations have vastly different reverberant characteristics. As listed in Table 1, the RTs for rooms and halls ranged widely from 0.36 to 3.62 s. The middle-six columns list RRs with the highest rates for each marked in bold. For comparison, the tenth and eleventh column also list RRs of CBFB_IMTF with CMN and RASTA. The rightmost column indicates the relative improvements in the error reduction rate of the CBFB_IMTF feature compared with that of the CQFB feature. The CQFB feature performs almost the same as MFCC or slightly better (on average).

The CBFB-based features outperformed the CQFB-based features. The CBFB_IMTF based feature, had the highest recognition in almost every case. On average, CBFB and CBFB_IMTF had relative improvements of 15.74% and 25.78% compared with that of CQFB. The table also lists how the differences in acoustic characteristics of the various environments affected the RR. Although CBFB_CMN and CBFB_RASTA had no improvements in performance compared with that of CBFB alone, a combination of IMTF with CMN and RASTA had good improvements. In most cases, it was found that CMN interacts to make an improvement with IMTF. There was almost no degradation in speech recognition for the meeting room, wooden house, living room, or movie theater (RR $> 90\%$) because these REs had been designed to minimize reverberant properties.

4 Conclusion

Our analysis and experiments demonstrated that our MTF-based sub-band power envelope extraction and inverse filtering algorithm improves the robustness of speech recognition for reverberant speech. The results revealed that: (1) Constant-Q band-pass processing or MFCC had no advantages for improving ASR in REs; (2) Considering the exponential decay properties of a RIR in a RE and the temporal modulation properties of speech, we can estimate the sub-band temporal power envelope of speech to some degree without having to measure the RIRs, thereby improving the ASR of reverberant speech; (3) in real REs, the proposed estimates of the sub-band temporal envelope with inverse filtering based on dereverberation consistently improves ASR; and (4) although CBFB_CMN had no improvement compared with that of CBFB alone, a combination of CBFB_IMTF with CMN can improve the RRs of CBFB_IMTF.

Comparing the RR of CBFB_IMTF and CBFB in Table 1, we find that adding inverse filtering to CBFB does not greatly improve the RR. The RR still low for many reverberant conditions. This suggests that we need to reconsider how some things are handled from both model and implementation aspects. In our experiments, reverberant speech was obtained by manual convolution between the speech and RIRs in artificial and real REs. However, we need to consider real reverberant speech, which should be recorded in a RE. In addition, apart from convolution-distortion, additive noise in real REs may cause speech to degrade. In the future, we will extend our method to deal with these problems.

ACKNOWLEDGEMENTS

This work was supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (Nos. 18680017 and 18700172). It was also partially supported by the SCOPE (071705001) of the Ministry of Internal Affairs and Communications, Japan. We would like to thank ATR Spoken Language Translation Research Laboratories for permitting us to use the AURORA-2J data.

References

[1] S. Furui and M. M. Sondhi, *Advances in Speech Signal Processing*, New York Marcel Dekker, Inc., 1991.

[2] T. Takiguchi, S. Nakamura, and K. Shikano, "Hands-Free Speech Recognition by HMM composition in Noisy Reverberant Environments," *IEICE Trans. D-II*, J79-D-II(12), 2047–2053, 1996.

[3] S. Nakagawa, "A Survey on Automatic Speech Recognition," *IEICE Trans. D-II*, J83-D-II(2), 433–457, 2000.

[4] F. Liu, R. Stern, X. Huang, and A. Acero, "Efficient Cepstral Normalization for Robust Speech Recognition," *Proceedings of ARPA Human Language Technology Workshop*, 1993.

[5] H. Hermansky, N. Morgan, and H. G. Hirsch, "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," *ICASSP'93*, 83–86, 1993.

[6] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. on Acoustics, speech, and signal processing*, ASSP (36), 145–152, 1988.

[7] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," *Proc. ICASSP'03*, 1, 92–95, 2003.

[8] M. Unoki, T. Hosorogiya, and Y. Ishimoto, "Comparative evaluations of robust and accurate F0 estimates in reverberant environments," *Proc. ICASSP2008*, 2008.

[9] R. V. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech Recognition with Primarily Temporal Cues," *Science*, 270, 303–304, 1995.

[10] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acustica*, 28, 66-73, 1973.

[11] M. R. Schroeder, "Modulation transfer function: definition and measurement," *Acustica*, 49, 179–182, 1981.

[12] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, "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.

[13] M. Unoki, K. Sakata, M. Furukawa, and M. Akagi, "A speech dereverberation method based on the MTF concept in power envelope restoration," *Acoust. Sci. & Tech.*, 25(4), 243–254, 2004.

[14] M. Unoki, M. Toi, and M. Akagi, "Development of the MTF based speech dereverberation method using adaptive time-frequency division," *Proc. Forum Acusticum 2007*, 51–56, 2005.

[15] X. Lu, M. Unoki, and M. Akagi, "A robust feature extraction based on the MTF concept for speech recognition in reverberant environment," *Proc. ICSLP'06*, 2546–2549, 2006.

[16] SMILE2004, Sound Material in Living Environment, Architectural Institute of Japan and GIHODO SHUPPAN Co., Ltd., 2004.

[17] http://sp.shinshu-u.ac.jp/CENSREC/, AURORA-2J database.

[18] The HTK Book (version 3.2), Cambridge University Engineering Department, 2002.

Table 1: Comparison of reverberant speech recognition rates (%) in actual reverberant environments. IRdata No. indicates File No. in SMILE2004 [16]. The reverberation time, T_R, was determined as an average from all T_Rs on the transfer function at 125 Hz to 8 kHz in octave frequencies. The "RI_CQFB" and "RI_CBFB" mean the relative improvement in the error reduction rate of the CBFB_IMTF feature compared with those of CQFB and CBFB features. MPH: Multi-purpose hall; CCH: Classic concert hall; GSH: General speech hall, RB: Reflex board, AB: Absorptive board, AC: Absorptive curtain.

Room condition (RIRs)	RIR No.	T_R (s)	MFCC	CQFB	CBFB	CBFB_ CMN	CBFB_ RASTA	CBFB_ IMTF	CBFB_ IMTF_ CMN	CBFB_ IMTF_ RASTA	RI_ CQFB	RI_ CBFB
MPH 1 [1]	301	1.09	42.55	45.56	52.44	57.63	48.51	**60.30**	66.53	59.53	*27.08*	*16.53*
MPH 1 [2]	302	0.80	55.39	54.31	68.52	71.85	66.17	**74.33**	81.95	76.36	*43.82*	*18.46*
MPH 2 [3]	303	1.44	32.88	36.60	40.62	39.70	32.64	**45.41**	47.47	40.31	*13.90*	*8.07*
MPH 2 [4]	304	1.04	39.70	43.51	47.56	45.49	36.20	**52.38**	54.93	46.55	*15.70*	*9.19*
MPH 3 [5]	305	1.93	30.70	33.40	33.80	35.31	31.26	**39.31**	42.46	36.54	*8.87*	*8.32*
MPH 3 [6]	306	1.35	42.12	43.48	46.52	53.42	47.50	**54.19**	61.87	55.08	*18.95*	*14.34*
MPH 4 [7]	307	1.42	55.70	55.07	69.63	74.24	71.05	**75.87**	80.87	76.39	*46.29*	*20.55*
MPH 4 [8]	308	1.54	52.44	53.42	67.02	71.08	66.78	**73.10**	77.31	72.06	*42.25*	*18.44*
MPH 5 [9]	319	1.47	46.55	47.28	61.38	59.84	54.71	**64.04**	68.28	62.02	*31.79*	*6.89*
MPH 6 [10]	321	2.16	40.13	42.83	49.95	49.43	47.99	**54.49**	58.43	53.05	*20.40*	*9.07*
CCH 1 [11]	309	2.35	27.72	34.20	35.19	33.50	28.92	**35.92**	42.09	33.65	*2.61*	*1.13*
CCH 1 [12]	310	2.34	30.09	35.65	39.88	37.03	33.22	**42.74**	45.16	39.61	*11.02*	*4.76*
CCH 1 [13]	311	2.35	30.40	35.22	37.67	35.34	33.19	**43.17**	43.29	40.31	*12.27*	*8.82*
CCH 1 [14]	312	2.39	30.58	35.37	39.73	38.44	35.55	**45.47**	46.95	42.00	*15.63*	*9.52*
CCH 1 [15]	313	2.38	27.82	33.93	36.17	34.30	32.36	**40.56**	42.09	37.58	*10.03*	*6.88*
CCH 2 [16]	314	1.14	40.34	44.34	50.60	58.12	49.59	**59.84**	67.21	59.50	*27.85*	*18.17*
CCH 3 [17]	315	1.96	35.00	36.81	37.73	42.80	39.12	**46.33**	52.72	45.41	*15.07*	*13.81*
CCH 4 [18]	316	1.92	41.23	41.42	50.02	49.95	46.15	**54.38**	58.46	51.24	*22.12*	*8.72*
CCH 4 [19]	317	2.55	34.33	36.72	41.97	41.14	37.15	**44.43**	47.62	43.44	*12.18*	*4.24*
CCH 5 [20]	323	2.32	31.78	37.70	38.29	34.85	32.58	**44.09**	44.67	39.91	*10.19*	*9.40*
CCH 6 [21]	324	1.77	37.73	41.42	43.57	42.55	38.38	**53.45**	54.50	49.06	*20.54*	*17.51*
CCH 6 [22]	325	1.74	40.13	44.18	47.87	46.27	42.25	**55.14**	57.23	51.24	*19.63*	*13.95*
CCH 6 [23]	326	1.69	34.73	38.23	44.34	43.11	41.42	**52.69**	52.07	46.67	*23.41*	*15.00*
Lecture room [24]	201	1.36	46.76	45.72	60.85	**70.31**	67.58	68.53	77.00	73.07	*42.02*	*19.62*
Theater hall [25]	318	0.85	46.24	48.82	60.55	60.39	53.39	**63.68**	73.88	66.23	*29.03*	*7.93*
Meeting room [26]	401	0.62	77.43	72.24	89.10	91.25	89.16	**91.62**	94.04	91.83	*69.81*	*25.87*
Lecture room [27]	402	1.12	55.85	53.18	70.83	**81.12**	78.75	80.32	86.98	84.71	*57.97*	*32.53*
Lecture room [28]	403	1.09	57.48	51.30	68.35	**83.97**	80.75	78.85	87.90	86.06	*56.57*	*33.18*
GSH [29]	404	1.54	40.44	44.89	51.58	46.58	44.55	**54.34**	55.63	50.05	*17.15*	*5.70*
Church 1 [30]	405	0.71	57.35	56.95	70.34	76.60	72.43	**77.56**	85.91	82.28	*47.87*	*24.34*
Church 2 [31]	406	1.30	33.71	37.21	41.42	40.87	30.52	**42.49**	49.77	42.92	*8.41*	*1.83*
Event hall 1 [32]	407	3.03	27.51	31.19	33.40	33.40	30.80	**36.87**	42.22	31.07	*8.25*	*5.21*
Event hall 2 [33]	408	3.62	28.77	32.98	35.62	37.27	34.88	**41.63**	44.58	36.48	*12.91*	*9.34*
Gym 1 [34]	409	2.82	21.61	26.59	29.08	27.88	25.39	**30.09**	34.82	29.14	*4.77*	*1.42*
Gym 2 [35]	410	1.70	32.51	37.33	39.98	41.60	36.29	**48.23**	51.70	47.62	*17.39*	*13.90*
Living room [36]	411	0.36	89.81	86.40	**98.31**	96.75	95.30	96.90	97.33	93.61	*77.21*	*-83.93*
Movie theater [37]	412	0.38	88.36	84.22	93.49	**95.95**	92.85	93.18	97.21	93.61	*56.78*	*-4.76*
Antrum [38]	413	1.57	35.19	36.91	39.70	43.97	36.08	**48.60**	52.96	46.61	*18.53*	*14.76*
Tunnel [39]	414	2.72	28.52	25.05	25.33	26.76	**35.06**	33.87	46.46	28.46	*11.77*	*11.44*
Concourse [40]	415	1.95	36.66	39.64	44.06	**46.18**	34.48	45.93	55.14	47.10	*10.42*	*3.34*
GSH 2 [41]	416	1.53	38.26	41.45	48.33	46.88	42.80	**56.13**	58.00	50.11	*25.07*	*21.10*
GSH 2 [42]	417	1.49	34.26	37.67	45.13	44.98	41.26	**51.77**	52.63	46.64	*22.62*	*12.10*
GSH 2 [43]	418	1.40	39.73	39.05	54.41	59.81	56.19	**65.18**	67.79	62.51	*42.87*	*23.62*

Note: [1] (with RB; capacity: $2,000$ m^3), [2] (without RB), [3] (with RB; capacity: $5,700$ m^3), [4] (without RB), [5] (with RB; capacity: $7,200$ m^3), [6] (without RB), [7] (with AB; capacity: $12,000$ m^3), [8] (without AB), [9] (capacity: $14,000$ m^3), [10] (capacity: $19,000$ m^3), [11] (capacity: $5,600$ m^3), [12] ($d = 6$ m), [13] ($d = 11$ m), [14] ($d = 15$ m), [15] ($d = 19$ m), [16] (capacity: $6,100$ m^3), [17] (capacity: $20,000$ m^3), [18] (with AC; capacity: $7,100$ m^3), [19] (without AC), [20] (capacity: $17,000$ m^3), [21] (1F front; capacity: $17,000$ m^3), [22] (2F side), [23] (3F), [24] (with flatter echo), [25] (capacity: $3,900$ m^3), [26] (capacity: 130 m^3), [27] (capacity: 400 m^3), [28] (capacity: $2,400$ m^3), [29] (capacity: $11,000$ m^3), [30] (capacity: $1,200$ m^3), [31] (capacity: $3,200$ m^3), [32] (capacity: $28,000$ m^3), [33] (capacity: $41,000$ m^3), [34] (capacity: $12,000$ m^3), [35] (capacity: $29,000$ m^3), [36] (wooden, capacity: 110 m^3), [37] (capacity: 560 m^3), [38](capacity: $4,000$ m^3), [39] (capacity: $5,900$ m^3, length: 120 m), [40] (train station), [41] (1F front), [42] (1F central), [43] (balcony).