BLIND REVERBERATION MITIGATION FOR ROBUST SPEAKER IDENTIFICATION

Seyed Omid Sadjadi and John H.L. Hansen[⋆]

Center for Robust Speech Systems (CRSS),
The University of Texas at Dallas, Richardson, TX 75080–3021, USA
{sadjadi, john.hansen}@utdallas.edu

ABSTRACT

Reverberation poses detrimental effects on performance of automatic speaker identification (SID) systems. This paper proposes a blind spectral weighting technique for combating the late reverberation effect (aka overlap-masking effect) on SID systems. The technique is blind in the sense that prior knowledge of neither the anechoic signal nor the room impulse response is required. Performance of the proposed technique is evaluated in terms of: 1) accuracy obtained from closed-set SID experiments, using speech material from the TIMIT corpus and four different measured room impulse responses from Aachen impulse response (AIR) database, and 2) equal-error rate (EER) obtained from experiments on a new data corpus well suited for speaker verification experiments under actual reverberant mismatched conditions, entitled MultiRoom8. Results prove that incorporating the proposed blind technique into the standard MFCC feature extraction framework yields significant improvement in SID performance.

Index Terms— Blind dereverberation, mismatched condition, overlap-masking effect, speaker identification, speaker verification

1. INTRODUCTION

In a reverberant enclosure, sound waves arrive at the receiver (e.g., ears or microphone) via a direct path, and via multiple paths and directions after reflecting off walls and objects defining the acoustic enclosure. The reflections arriving within 50-80 ms after the direct sound are called early reflections, which tend to build up to a level louder than the direct sound and cause an internal smearing effect known as the “self-masking effect”. The echoes reaching the receiver after the early reflections are called late reflections, which tend to smear the direct sound over time and mask succeeding sounds. This phenomenon is commonly referred to as the “overlap-masking effect”, and has been shown to be the primary cause of degraded speech identification performance for both human and machine listeners [1], [2]. The overlap-masking effect can also mask/obscure the spectral details and acoustic cues essential for automatic speaker identification (SID), resulting in a major drop in performance [3], [4].

From a signal processing perspective, reverberation can be considered a convolutive/channel distortion, nevertheless, in the seminal work of [2] it has been shown that the overlap-masking effect can be modeled as an uncorrelated additive interference. Hence, it can be compensated via spectral subtraction, given that an estimate of the late reverberation spectral variance is available. This has inspired several single and multichannel approaches that have considered spectral subtraction for blind late reverberation suppression [5], [6]. Because a rough estimate of the reverberation time (aka T_{60}) is required to compute the late reverberation spectral variance, performance of these approaches are highly dependent on the accuracy of the T_{60} estimation.

In this paper, following the uncorrelated and additive assumption for late reverberation, we propose a spectral weighting technique to mitigate the reverberation overlap-masking effect on performance of automatic SID systems. The weights are computed using a parametric gain function which is based on *a priori* signal-to-interference ratio (SIR) estimate. A smoothed and shifted version of the reverberant power spectrum is used as an approximation for the late reverberation spectral variance. The technique is entirely blind, meaning that prior knowledge of neither the anechoic signal nor the room impulse response (RIR) is required.

Performance of the proposed technique in mitigating the adverse reverberation impact on SID is evaluated through closed-set SID and speaker verification experiments. For the closed-set SID task, we consider four different reverberant mismatched conditions simulated using TIMIT speech data and measured RIRs, with T_{60} ranging from 0.11 s to 0.83 s, extracted from the Aachen impulse response (AIR) database [7]. For the verification experiment, we consider 7 distinct reverberant mismatched scenarios from the MultiRoom8 corpus made available by AFRL. We employ the proposed spectral weighting solution as a pre-processing step in the standard MFCC feature extraction framework, and evaluate its effectiveness in suppressing the late reverberation effect on SID. For the sake of comparison, we also perform the same experiments with two other blind reverberation compensation strategies, namely long-term log-spectral subtraction (LTLSS) [8], and Gammatone subband based non-negative matrix factorization (NMF) [9].

2. BLIND SPECTRAL WEIGHTING (BSW) ALGORITHM

2.1. Mathematical model of reverberation

In a reverberant environment, the speech signal received at the microphone is a delayed sum of a direct sound and its reflections from walls and objects in the acoustic enclosure, and hence can be modeled as the convolution of the RIR with the speech signal,

$$r(n) = \sum_{j=0}^{L-1} s(n-j)h(j) = s(n) * h(n), \tag{1}$$

where $r(n)$ and $s(n)$ are the reverberant and anechoic signals, respectively, and $h(n)$ is the RIR. The RIR $h(n)$ can be partitioned into two parts $h_e(n)$ and $h_l(n)$ as,

$$h(n) = \begin{cases} 0, & n < 0 \\ h_e(n), & 0 \le n < n_e \\ h_\ell(n), & n_e \le n < L \end{cases} \tag{2}$$

[⋆]This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067. Approved for public release; distribution unlimited.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

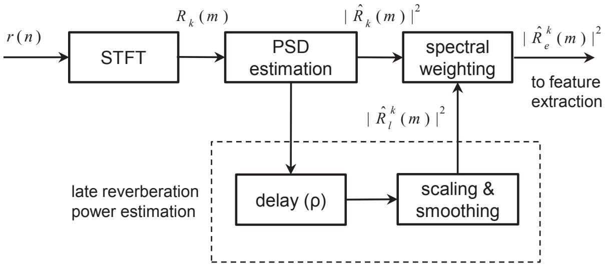

Fig. 1. Block diagram of the proposed spectral weighting technique for suppression of the late reverberation.

where L is the length of $h(n)$, and n_e is a time window threshold chosen such that $h_e(n)$ consists of the direct path signal and a few early reflections, while $h_\ell(n)$ consists of all the late reflections. The time threshold n_e is commonly set to a value within 50-80 ms range. Late reflections that smear the speech spectra and reduce signal quality, are characterized by T_{60}. These have a long-term effect on speech signals and therefore cannot be effectively compensated for using conventional cepstral mean subtraction (CMS) within the short-term speech analysis framework [10]. On the other hand, early reflections that cause coloration distortion and increase the prominence of low-frequency energy, are characterized by the direct-to-reverberant ratio (DRR) which is dependent on the distance between the sound source and microphone.

Taking (2) into account, (1) can be rewritten as,

$$r(n) = \underbrace{\sum_{j=0}^{n_e-1} s(n-j)h_e(j)}_{r_e(n)} + \underbrace{\sum_{j=n_e}^{L-1} s(n-j)h_\ell(j)}_{r_\ell(n)}, \qquad (3)$$

where $r_e(n)$ and $r_\ell(n)$ are referred to as the early speech component and late reverberant speech, respectively. Our objective is to blindly suppress the late reverberant speech using spectral weighting, hence mitigating the detrimental effects of reverberation on the performance of automatic SID systems.

2.2. Parametric gain function

A block diagram illustrating the proposed spectral weighting technique for mitigating the reverberation overlap-masking effect on the performance of automatic SID systems is depicted in Fig. 1. The late reverberant speech is suppressed in the short-time Fourier transform (STFT) domain by applying spectral weights as,

$$\hat{R}_e^k(m) = G_k(m) \cdot R_k(m), \qquad (4)$$

with m and k being the time frame and frequency-bin indices, respectively. The spectral weights are computed using a parametric gain function defined as,

$$G_k(m) = \left(\frac{\xi_k(m)}{\xi_k(m)+\alpha}\right)^\beta, \qquad (5)$$

where $\xi_k(m)$ denotes the *a priori* SIR, and α and β are some constant parameters. The *a priori* SIR is defined as,

$$\xi_k(m) = \frac{\lambda_{r_e}^k(m)}{\lambda_{r_\ell}^k(m)}, \qquad (6)$$

where $\lambda_{r_e}^k(m) = E\left[\left|R_e^k(m)\right|^2\right]$ and $\lambda_{r_\ell}^k(m) = E\left[\left|R_\ell^k(m)\right|^2\right]$ denote spectral variances of the early and late speech components, respectively, both of which are to be estimated.

It is common practice to recursively estimate $\xi_k(m)$ via the decision-directed method [11] as,

$$\hat{\xi}_k(m) = \eta \frac{\left|\hat{R}_e^k(m-1)\right|^2}{\hat{\lambda}_{r_\ell}^k(m-1)} + (1-\eta)\max[\gamma_k(m)-1, 0], \qquad (7)$$

where η $(0 \leq \eta \leq 1)$ is a smoothing constant that controls the trade-off between interference reduction and transient distortion introduced into the signal. The first term $\frac{\left|\hat{R}_e^k(m-1)\right|^2}{\hat{\lambda}_{r_\ell}^k(m-1)}$, represents the estimate of $\xi_k(m)$ from the previous time frame, while the second term $\max[\gamma_k(m)-1, 0]$, is the maximum likelihood (ML) estimator for $\xi_k(m)$ and solely dependent on the current frame. The parameter $\gamma_k(m)$ is called the *a posteriori* SIR and is defined as,

$$\gamma_k(m) = \frac{E\left[\left|R_k(m)\right|^2\right]}{\lambda_{r_\ell}^k}. \qquad (8)$$

The two SIRs are related via $\gamma_k(m) = \xi_k(m) + 1$. The recursive relationship in (7) provides smoothness in the estimate of $\xi_k(m)$ which consequently helps eliminate the musical noise distortion. In practice, to further reduce distortions introduced by the spectral weighting, the gain function $G_k(m)$ is lower bounded by a constant gain floor G_f.

The motivation behind employing a gain function in the form of (5), is twofold. First, the parametric Wiener filtering [12] has been successfully applied to a similar problem in the context of noisy speech enhancement. In addition, it can be easily shown that common speech enhancement algorithms such as spectral subtraction and maximum-likelihood methods, are special cases of the parametric Wiener filtering. Second, the two parameters α and β provide more degrees of freedom and control over the late reverberation suppression and speech distortion reduction. It has been shown in [13] that the speech distortion introduced by speech enhancement algorithms can result in a severe performance degradation for automatic SID systems.

2.3. Late reverberation power estimation

In order to estimate the two SIRs, an estimate of the late reverberation spectral variance must be available. In [2], a simple statistical model for the RIR was considered and an estimator for $\lambda_{r_\ell}^k(m)$ was derived. The estimator is dependent on the T_{60}, which can be estimated directly from the reverberant data, albeit at the cost of a more complex algorithm. This approach was further investigated in [5] and [6] to accommodate for the estimation and reduction of additive noise. In addition, a ML approach for T_{60} estimation was proposed in [6].

Here, an alternative approach for the estimation of the late reverberation spectral variance is taken which obviates any need for direct T_{60} estimation. Considering the smearing effects of the late reverberation on the speech signal, the power spectrum of the late speech component can be assumed to be a smoothed and shifted version of the reverberant speech power spectrum as [14],

$$\hat{\lambda}_{r_\ell}^k = \mu\, w(m-\rho) * \left|R_k(m)\right|^2 \qquad (9)$$

where the symbol $*$ denotes the convolution in the time domain, $w(m)$ is a smoothing function, and $\rho = n_e$ is the time threshold between early and late components of the RIR. As noted earlier,

n_e is commonly set to a value within the 50-80 ms range, and is independent of reverberation characteristics. The parameter μ is a scaling factor that specifies the relative strength of the late speech component.

Since RIRs have a decaying exponential shape, a right skewed smoothing function with a long tail would be a reasonable choice for $w(m)$. Therefore, as in [14], Rayleigh distribution function is adopted,

$$w(m) = \begin{cases} 0, & m \le -b \\ \frac{m+b}{b^2} \exp\left(\frac{-(m+b)^2}{2b^2}\right), & m > -b \end{cases} \tag{10}$$

where b determines the overall spread of the smoothing function, and is set in accordance with the time threshold between early and late components of the RIR, n_e.

3. EXPERIMENTS

Performance of the proposed blind spectral weighting technique for suppression of late reverberation is evaluated in the context of: 1) GMM based closed-set SID experiments, and 2) GMM-UBM speaker verification experiments [15]; SID accuracy and EER are reported as performance measures, respectively. The proposed technique is integrated into the MFCC feature extraction framework as a pre-processing stage, and performance is compared to that of the baseline system with no pre-processing.

For closed-set SID experiments, training and test speech material are obtained from the TIMIT corpus consisting of speech from 630 speakers including 192 female and 438 male speakers. There are 10 sentences per speaker recorded under clean laboratory conditions at a sampling rate of 16 kHz. A total of 8 sentences ($\sim$ 24 s) are used to train speaker models, while the remaining 2 sentences ($\sim$ 6 s) test the models. To simulate different reverberant conditions, RIR samples extracted from the AIR database are convolved with test material. Four RIRs with distinct source-to-microphone distances (d_{SM}) and with T_{60} ranging from 0.11 s to 0.83 s are used including studio booth, meeting, office, and lecture rooms. Further information concerning the RIRs is summarized in Table 1. Here, a 32-mixture GMM SID system is trained on anechoic data for evaluations.

For speaker verification experiments, speech material from MultiRoom8 corpus are utilized. The MultiRoom8 database, which is made available by AFRL, was designed to capture multi-session audio impacted by environmental contamination, i.e., background noise and room reverberation. It contains a development set with a total of 100 speech files which are used for building the UBM, 7 different training-test conditions representing a range of distinct reverberant and noisy mismatched scenarios, and a training-test condition involving different communication channels, which is not exploited in this study. Four rooms were used for data collection including: small (5.3 × 3.6 m^2), medium (11.3 × 3.6 m^2), large (14.6 × 12.9 m^2), and a conference room. The rooms are labeled as Sm, Med, Lg, and Enroll, respectively. Except for the conference room where recordings were collected using only close-talking microphones, for each environment, 6 uni- and omni-directional microphones located at a range of distinct distances from the speaker were used for speech capture. Each session was recorded at least 1 week from the previous session for each speaker. In an interview-like scenario, a total of 52 speakers were recorded, although not each speaker is present for every room. The average length of the recordings is approximately 3 minutes. Here, a 1024-mixture UBM is built on the development set, and individual speaker models are MAP adapted from the UBM.

Table 1. Properties of the four RIRs extracted from the AIR database for experiments. d_{SM} denotes the source-to-microphone distance.

Room Type	Dimension (m^3)	d_{SM} (m)	T_{60} (s)	DRR (dB)
Studio booth	3.0 × 1.8 × 2.2	1.0	0.11	8.78
Meeting	8.0 × 5.0 × 3.1	2.8	0.25	2.89
Office	5.0 × 6.4 × 2.9	3.0	0.48	-0.89
Lecture	10.8 × 10.9 × 3.15	10.2	0.83	-5.62

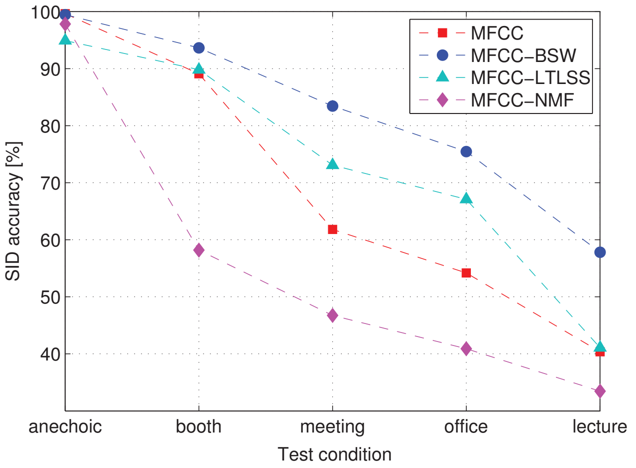

Fig. 2. Performance of blind reverberation compensation front-ends in terms of closed-set accuracy (%), obtained from SID experiments on TIMIT corpus under anechoic and four different reverberant mismatched test conditions.

To perform the spectral weighting, the reverberant signals are transformed into the STFT domain using Hamming windowed frames of 25 ms duration with a 10 ms skip rate. The *a priori* SIR is estimated using the decision-directed approach (7) with a smoothing factor $\eta = 0.6$. The time threshold between early and late components of the RIRs is set to 50 ms which, considering the 10 ms skip rate, corresponds to 5 frames. In order to find the optimum parameters for the parametric gain function (5), speech data from 80 speakers, including 37 females and 43 males, from the TIMIT corpus are used as the development set. It was found that setting $\alpha = 2$ and $\beta = 2.5$, on average yields the best performance across the various reverberant mismatched conditions. The gain floor parameter G_f is fixed to 0.01 which is equivalent to a maximum attenuation of -20 dB. In contrast to the findings reported in [14], tuning the scaling factor μ, that specifies the relative strength of the late speech component, seems to be very important for SID tasks. Here, μ is set to 0.1, since larger values for this parameter will result in speech distortion that is intolerable for the SID system, which in turn can lead to a great drop in performance [13]. Standard MFCC features are extracted from the processed spectra, and normalized to a standard Gaussian distribution over a 3-second sliding window for SID experiments.

4. RESULTS

Fig. 2 shows closed-set speaker identification accuracies obtained by the GMM based system on TIMIT data under anechoic and four distinct reverberant mismatched test conditions, with and without the proposed blind spectral weighting (BSW) algorithm as the pre-processing stage for the MFCC feature extraction. It is clear that incorporating BSW within the feature extraction framework results in significant improvements in performance. An average absolute improvement of 16.71 % is achieved over the baseline system with plain MFCC features. For the studio booth test condition, improvement is the smallest because the self-masking effect is dominant, therefore the spectral weighting will not be very effective in this case. However, under the remaining reverberant mismatched conditions, improvements are significantly higher because the overlap-masking effect has been the major source of performance degradation. The results obtained here indicate that, in line with the findings in psychoacoustic studies (e.g., [1]) and when compared to the self-masking effect, the overlap-masking effect has a greater impact on performance of SID systems. Suppressing this effect can thus alleviate its adverse impact on SID performance.

To compare the performance of our BSW technique with other blind reverberation compensation strategies, we performed the same SID experiments using MFCC features extracted from data pre-processed with the LTLSS [8], and Gammatone subband NMF [9]. Results are presented in Fig. 2. It is evident from the figure that our technique consistently outperforms the other two strategies in suppressing the reverberation effects on SID. Note that the system performance with NMF under reverberant conditions is even worse than the performance with plain MFCCs. This is due to the fact that this method introduces a great amount of processing artifacts intolerable for the SID system (confirmed through informal listening experiments). In addition to the superior performance, there is no need for signal reconstruction with our technique, as required with both the LTLSS and NMF strategies.

Results for speaker verification experiments on the MultiRoom8 corpus are summarized in Table 2. Train-Test labels in the first column denote the room/microphone combinations. For instance, Lg5-Sm4 implies that the speaker models are trained on data recorded in the large room using microphone number 5, while the evaluation is performed on material recorded in the small room using microphone number 4. Consistent with the closed-set SID experiment outcomes, employing the proposed BSW technique as a pre-processing stage in the MFCC extraction has resulted in remarkable performance improvements. Here, an average absolute improvement of 3.16 % is achieved over the baseline system with plain MFCC features. The results also further confirm the superiority of our technique over the other two blind reverberation compensation strategies [8], [9].

5. CONCLUSION

In this paper we proposed a blind spectral weighting (BSW) technique for alleviating the impact of late reverberation on performance of SID systems. The technique is blind in the sense that prior knowledge of neither the anechoic signal nor the room impulse response is required. In addition, the late reverberation spectral variance was estimated without the direct need for T_{60} estimation. It was confirmed that incorporating the proposed BSW technique as a pre-processing stage in the MFCC feature extraction framework results in significant improvements in both automatic SID performance under simulated and actual reverberant mismatched conditions. We believe that this technique can potentially benefit other automatic speech applications, such as automatic speech recognition (ASR), under the same mismatched conditions.

Table 2. Performance of blind reverberation compensation front-ends in terms of EER (%), obtained from speaker verification experiments on MultiRoom8 corpus.

Train-Test	EER [%]			
	MFCC	MFCC-BSW	MFCC-LTLSS	MFCC-NMF
Lg5-Sm4	13.16	**10.53**	13.16	13.87
Sm4-Lg5	11.17	**7.89**	10.53	15.86
Enroll-Sm6	21.15	**16.28**	20.93	20.93
Enroll-Sm4	13.46	**9.30**	11.63	16.28
Med3-Sm3	11.67	**10.26**	14.91	12.82
Lg4-Med5	19.44	**16.67**	22.22	25.00
Med5-Sm5	10.93	**7.96**	10.66	10.66

6. REFERENCES

[1] A. K. Nabelek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Am.*, vol. 86, pp. 1259–1265, Oct. 1989.

[2] K. Lebart, J. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica*, vol. 87, pp. 359–366, 2001.

[3] P. Castellano, S. Sridharan, and D. Cole, "Speaker recognition in reverberant enclosures," in *Proc. IEEE ICASSP*, vol. I, May 1996, pp. 117–120.

[4] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, pp. 2023–2032, Sept. 2007.

[5] E. A. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. IEEE ICASSP*, vol. 4, Mar. 2005, pp. 173–176.

[6] H. Lollmann and P. Vary, "A blind speech enhancement algorithm for the suppression of late reverberation and noise," in *Proc. IEEE ICASSP*, Apr. 2009, pp. 3989–3992.

[7] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. IEEE DSP*, Jul. 2009, pp. 1–5.

[8] D. Gelbart and N. Morgan, "Double the trouble: handling noise and reverberation in far-field automatic speech recognition," in *Proc. ICSLP*, Sept. 2002, pp. 2185–2188.

[9] K. Kumar, R. Singh, B. Raj, and R. Stern, "Gammatone subband magnitude-domain dereverberation for ASR," in *Proc. IEEE ICASSP*, May 2011, pp. 4604–4607.

[10] Y. Pan and A. Waibel, "The effects of room acoustics on MFCC speech parameter," in *Proc. ICSLP*, Oct. 2000, pp. 129–132.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[12] J. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.

[13] S. O. Sadjadi and J. H. L. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions," in *Proc. INTERSPEECH*, Sept. 2010, pp. 2138–3141.

[14] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, pp. 774–784, May 2006.

[15] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, Jan. 2000.