



Methods for Robust Speech Recognition in Reverberant Environments: A Comparison

Rico Petrick , Thomas Fehér , Masashi Unoki , Rüdiger Hoffmann

Laboratory of Acoustics and Speech Communication, Dresden University of Technology, Germany
School of Information Science, Japan Advanced Institute of Science and Technology, Japan

[Rico.Petrick,Thomas.Feher,Ruediger.Hoffmann]@ias.et.tu-dresden.de unoki@jaist.ac.jp

Abstract

In this article the authors continue previous studies regarding the investigation of methods that aim to improve the decreased recognition rate (RR) in reverberant environments of automatic speech recognition (ASR) systems. Previously three robust front-end methods are tested, the harmonicity based feature analysis (HFA), the temporal power envelope feature analysis (TPEFA) and their combination (HFA+TPEFA). This paper additionally introduces two well-known methods into the comparison. These are the dereverberation method using the inverse modulation transfer function (IMTF) and the delay-and-sum beamformer (DSB). Recognition experiments are accomplished for command word recognition, the reverberant environments are comprehensive chosen as functions of the reverberation time and the speaker to microphone distance (SMD) as the most important parameters to describe reverberant distortions. The results of this first comparison of such methods prove experimentally some drawn assumptions, e. g. the IMTF method achieves robustness only in the far field, the DSB improves the RR slightly but is outperformed by the HFA due to its indirectivity at low frequencies.

Index Terms: reverberation, harmonicity, robust ASR

1. Introduction

This article compares front-end processing methods that are developed to increase the robustness of automatic speech recognition (ASR) systems in unknown reverberant environments. It describes the progress in research which is based on previous work of the authors, e. g. [1, 2, 3]. The background of this research is the design of practical command and control applications in rooms. Therefore several restricting working conditions for the recognizer but also for methods achieving robustness have to be met [2]. These are (a) a robust ASR with acceptable recognition rate (RR) (% [4]) under typical varying indoor conditions (reverberation time s and speaker to microphone distance (SMD) m), (b) no or real-time adaptation (s, adaptation only on command words), (c) robustness against changes of the room impulse response (RIR) due to movements of speakers/objects, and (d) feasible numerical complexity for implementation on an embedded processor.

In the last 10 years many researchers have faced the problem of robustness in ASR against room reverberation. The developed methods can be classified in the same manner as the approaches against additive noises, which is the classification in signal, feature and model domain approaches. Signal domain approaches are basically blind dereverberation methods, such as [5, 6]. Although large progress has been achieved in this research field, most of these methods are still not suitable

for the above mentioned practical conditions because of high adaptation times, low robustness or high numerical complexity. Feature domain approaches that rely on speech characteristics, such as Harmonicity based Feature Analysis (HFA) [2] or Temporal Power Envelope Feature Analysis (TPEFA), have been successfully tested in reverberant environments [3, 7]. Model based approaches are according to [8] the most successful way of increasing environmental robustness if an HMM can be derived that already includes the environmental characteristics. These methods basically subdivide into model adaptation techniques and reverberant training, where for the former only a limited number of approaches against reverberation is developed so far ([9, 10, 11]). When compared with other methods, reverberant training, however, seems to work best. Consequently the authors propose to combine it with front-end methods as investigated in [3]. In the experiments in this paper, the reverberant training is avoided in order to obtain the true improvement caused by the front-end methods. Previous work of the authors evaluates the methods HFA, TPEFA and the combination HFA+TPEFA in reverberant environments [3]. Continuing this investigation the present paper adds two further enhancing methods into the evaluation, the dereverberation method based on the inverse modulation transfer function (IMTF) and the delay-and-sum beamformer (DSB). It is shown that both methods can enhance ASR performance. Beside the comparison with the other methods, the main contributions of this paper are to show shortcomings of these two methods and their limited appropriateness for enhancing the ASR robustness in reverberant environments. Explanations and experimental proves are given. All tested enhancing methods (HFA, TPEFA, HFA+TPEFA, IMTF, DSB, DSB+HFA) are chosen because they fulfill the above mentioned requirements of practical command and control applications.

2. Previous Evaluations

To carry on the previous research of the authors, this work uses the same evaluation system as described in [1, 2, 3]. It consists of the UASR recognizer [12], a subset of the APOLLO corpus [4] (1020 command phrases of 17 classes, each 2 s speech). The RR is measured without rejection. Room acoustic test and training conditions (varying and SMDs) can be simulated by convolving the applied corpus with the dedicated measured room impulse response (RIR). The task of these evaluations is to find analysis methods which can enhance the bad ASR performance in reverberant conditions. Following the methods evaluated in the previous work of the authors [3] are briefly described and certain aspects of their advantages concerning for the present comparison are outlined. For detailed description of the methods refer to the given reference.

10.21437/Interspeech.2010-228

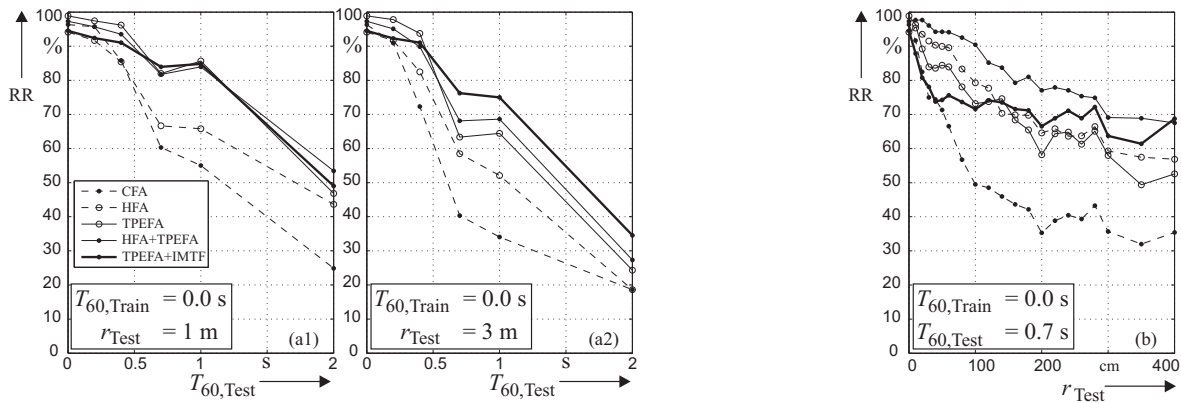


Figure 1: Measured recognition rates dependent on $T_{60,Test}$ ((a1) and (a2)) and on the SMD ((b)) in the SMART-Room environment [13]. The SMD for (a1) is $r_{Test} = 1$ m (near field) and for (a2) is $r_{Test} = 3$ m (far field). For the four front-ends CFA, HFA, TPEFA and HFA+TPEFA the results correspond to those in [3] for clean training. The new contributions are the results for the IMTF method in comparison to the others.

CFA – Conventional Feature Analysis. The term is chosen to distinguish from the below introduced methods. It is the base line feature analyses in the present evaluations and consists of a 30 channel mel filterbank (MFB, described in [12]). The authors also tested MFCCs, but they did not achieve better results than MFB features.

HFA – Harmonicity based Feature Analysis is proposed in [2] as a method for increasing ASR robustness against reverberation. It emphasizes reliable (harmonic spectral components in voiced sections) and suppresses unreliable speech features (nonharmonic spectral components in voiced sections and low frequencies in unvoiced sections). It has been shown that HFA can improve the performance of an ASR system in reverberant environments [2, 3]. Combined with reverberant training an advantage of HFA is that it can achieve high and stable RRs in reverberant environments.

TPEFA – Temporal Power Envelope Feature Analysis is proposed in [7] (in [7] named as constant-bandwidth filterbank (CBFB) with a subsequent MFCC feature analysis) and later in [3]. It extracts the temporal power envelopes (TPEs) of subbands (channel index k) of the incoming speech $s(k, t)$. The TPEs are meant to contain most speech intelligibility information [14]. TPEFA further performs low pass filtering of the TPEs (f_{LP} Hz), which is basically the application of the ideas of RASTA [15]. As also the TPEs are reverberated in rooms. Mathematically the reverberation of clean TPEs can be described by the multiplication

$$TPE_{rev}(k, t) = TPE_{clean}(k, t) * MTF(k, t) \quad (1)$$

where $TPE_{clean}(k, t)$ is a subband TPE of the RIR $h(k, t)$, $M(k, t)$ is the so called modulation spectrum and $MTF(k, t)$ is the (sub-band) modulation transfer function (MTF) [16]. Despite the reverberation in $h(k, t)$ it is shown that its large-scale structure still contains reliable speech features, which makes TPEFA a candidate for robust front-end processing in reverberant environments (proven in [3]). Combined with reverberant training TPEFA achieves little better results than HFA for a specific trained condition. However, it fails for other test conditions [3] where HFA performs stably.

HFA + TPEFA in Combination is the series connection of HFA and TPEFA. HFA additionally needs to resynthesize the

enhanced spectra to a speech signal. The way how to achieve this resynthesis is described in [3]. It is further shown that HFA+TPEFA combines the advantages of both methods, stable plus high RR.

3. IMTF based Dereverberation

The first addition to our evaluation is the dereverberation method based on the inverse modulation transfer function (IMTF). It is proposed in [17] and already tested in [7] for ASR (only in the far field). Near field and far field SMDs are separated by the critical distance r_c of room acoustics ($r_{near} < r_c < r_{far}$). Due to the preferences of the acoustic sound field in rooms the authors assume that the IMTF method has a well working far field behavior, but a poor working behavior in the near field. This assumption is due to the theory of the IMTF method which is defined for the far field condition and is here mentioned and experimentally proven for the first time.

3.1. Algorithmic Aspects

The above mentioned introduction of reverberation on the TPEs by the multiplication with the MTF $M(k, t)$ is aimed to be compensated by the IMTF method. This is achieved by multiplying the inverse MTF $M^{-1}(k, t)$ on the reverberant TPEs. Since the IMTF is unknown it is blindly estimated by the simple equation

$$M^{-1}(k, t) = \frac{TPE_{clean}(k, t)}{TPE_{rev}(k, t)} \quad (2)$$

that relies on the far field assumption ($r > r_c$). With this assumption only the subband k is needed as a parameter in (2). A method for blindly estimating $M^{-1}(k, t)$ out of the reverberant TPEs is proposed in [17]. However, since this method relies on the far field assumption, it may not work for near field conditions ($r < r_c$). In fact the near field assumption results in a far more complex equation for the MTF as derived in [18], but this is seldom mentioned and cited. Because of this complex equation the near field MTF is difficult to estimate blindly and is therefore also difficult to invert.

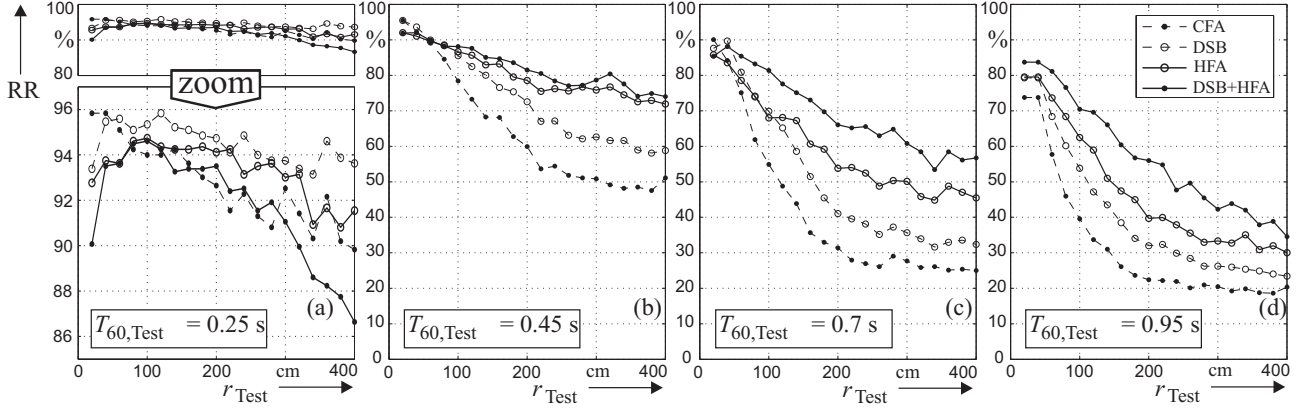


Figure 2: Measured recognition rates for the four front-ends CFA, DSB, HFA and DSB+HFA dependent on the SMD r_{Test} . (a)...(d) show the results for the different rooms in order of increasing $T_{60,\text{Test}} = (0.25; 0.45; 0.7; 0.95)$ s. (a) is zoomed for convenience.

3.2. Experiments

As mentioned above, the experiments in Fig. 1 are accomplished on exactly the same data as in [3]. The results prove that the IMTF concept following equation (2) only works well in the far field condition, but not in the near field. Since the IMTF is an enhancing method of TPEFA, it is to be compared with the results of TPEFA. This is shown in Fig. 1 (b) where the RR dependent on the SMD is measured. It shows that the IMTF-RR strongly decreases for clean data (small SMDs). The RR is much worse than for TPEFA. While increasing the SMD at some point ($r_{\text{Test}} = 100$ m) the graphs IMTF and TPEFA cross. In the very far field IMTF performs best compared to all other methods. This behavior can also be observed in Fig. 1 (a1) and (a2): For the far field condition $r_{\text{Test}} = 400$ m IMTF performs best, where for the near field $r_{\text{Test}} = 100$ m no improvement can be observed (this distance may correspond to the crossing of the graphs in Fig. 1 (b)). The decreasing of the RR for nonreverberant data can also be observed in Fig. 1 (a1) and (a2) for results at $r_{\text{Test}} = 100$ m.

4. DSB – Delay-and-Sum Beamformer

Although it is mostly mentioned as a method that improves performance in noisy or reverberant environments, for two reasons the authors assume only a small ASR improvement in reverberant environments. First, DSBs can attenuate side noises (and side reverberation) only in a light manner due to the limited number of microphones (doubling the number of microphones results in an SNR improvement of only 3 dB). Second, DSBs have a strong frequency dependent directivity increasing from 0 dB for lower frequencies up to several dB for higher frequencies (Fig. 3). But this effect may possibly not be useful in this context, since the authors have already identified in [1] that high frequency reverberation is almost harmless for ASR whereas low frequency reverberation, which are virtually not directed by DSBs, is most harmful for ASR. Even if small ASR improvement can be expected, this paper contributes that the DSB is not the most appropriate enhancing method for ASR in reverberant environments (theoretical assumption and experimentally proven in section 4.2).

4.1. Algorithmic Aspects

The principle of the DSB with microphone arrays (MA) is well known [19]. It is simply adding multiple microphone signals.

The resulting signal will be either amplified or attenuated depending on the position of the source and the microphones and the signal frequency. Consequently beamformers have a very complex frequency dependent directivity pattern. The steering direction can be changed by delaying the signals in time-domain DSB. For the present work three microphone arrays according to Fig. 3 (a)...(c) are considered, where 3 (d) shows the directivity index $\gamma(f)$. It clearly shows that DSBs are ineffective for signals with high wavelength in comparison to the distance between the microphones. Adding microphones without increasing the overall array size even reduces directivity for low frequencies although it increases directivity at high frequencies [19]. To improve low frequency directivity only arrays with larger dimensions are a solution, but not useful in practical situations.

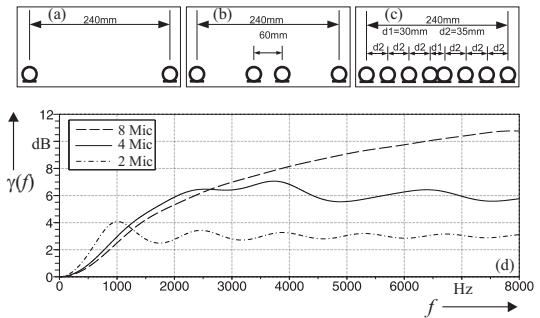


Figure 3: (a)...(c) geometrical structure of the considered microphone arrays (2, 4 and 8 channels). (d) Frequency dependent directivity index $\gamma(f)$ of three microphone arrays. For definition of $\gamma(f)$ refer to [19]. For the ASR experiments the geometry (b) was chosen as a compromise between handiness and performance, which is the DSB composed of 4 nonequally spaced cardioid microphones. The different distances between the microphones are an optimization to prevent large side lobes and to smooth the frequency dependent directivity index [20].

4.2. Experiments

For comparison of the performance of the DSB to other approaches it is not possible to use the old acoustic environment as applied in the previous experiments in Fig. 1, since these

are single channel recordings. For that reason the authors measured new RIRs in four rooms (sound studio ms, living room ms, office room ms, computer laboratory ms). For each room the RIRs are recorded at different SMDs increasing in steps of 20 cm (m). The RIRs of the four channels of the microphone array are simultaneously recorded for each SMD position. Subsequently the RIRs are convolved with the data of the test corpus. The experiments compare four different ASR front-ends: CFA, HFA (both as used in previous experiments [3]), DSB and the series connection of DSB and HFA. The results are displayed in Fig. 2. As assumed the ASR performance decreases with increasing (from (a) to (d)). The same can be stated for increasing SMDs within one room. Attenuating side reflections in the far field the DSB achieves some improvement (10 %). At very close SMDs (20 cm) the RR decreases slightly due to geometrical near field effects of the DSB. HFA is able to attenuate the harmful low frequency reverberation and performs much better than the DSB (Fig. 3 (d)). However, HFA decreases the RR slightly for clean speech (Fig. 2 (a) . . (c), Test cm), since it deletes some useful information from the signal (already mentioned in [3]). The series connection DSB+HFA (DSB output is HFA input) combines the positive effects of both methods and leads to the best results in these experiments.

5. Conclusion

The authors have compared a number of front-end processors for robust ASR in reverberant environments which are chosen because they meet practical limitations, such as very low (or no) adaptation time and feasible processing requirements. For the first time the experiments have proved the assumption that the IMTF method works well for the far field but has no or even a disturbing effect for nonreverberant speech, e. g. at near field SMDs. Consequently, the authors recommend to detect the level of reverberation and switch the IMTF based dereverberation off in case of low reverberation. There the remaining TPEFA performs already very well, as can be seen in Fig. 2 (short SMDs). The DSB experiments show that using DSBs without additional techniques is questionable, since it is the only method that needs more than one microphone, while gaining rather small improvement compared to other approaches. Further the performance of the DSB is dependent on the speaker position, which is a limitation for practical considerations, e. g. moving speakers. If these limitations are accepted the DSB can be used as a support for other approaches since it adds some gain which the experiments in Fig. 2 have proven. It applies for future work to generate experimental results for TPEFA, HFA+TPEFA and IMTF for the new recorded environments in Fig. 2. However, comparing the Figs. 1 and 2 it could be carefully assumed that HFA+DSB perform similar to HFA+TPEFA. Further it is also future work to test all methods with reverberant training as already started in [3].

6. References

- [1] Petrick, R., Lohde, K., Wolff, M. and Hoffmann, R., "The harming part of room acoustics for automatic speech recognition," *Proc. INTERSPEECH 2007*, Antwerp, 2007. pp. 1094 – 1097.
- [2] Petrick, R., Lohde, K., Lorenz, M., and Hoffmann, R., "A new feature analysis method for robust ASR in reverberant environments based on the harmonic structure of speech," *Proc. EU-SIPCO 2008*, Lausanne, 2008. CD-ROM.
- [3] Petrick, R., Lu, X., Unoki, M., Akagi, M. and Hoffmann, R., "Robust Front End Processing for Speech Recognition in Reverberant

Environments: Utilization of Speech Characteristics," *Proc. INTERSPEECH 2008*, Brisbane, Australia, 2008. pp. 658 – 661.

- [4] Maase, J., Hirschfeld, D., Koloska, U., Westfeld, T., and Helbig, J., "Towards an evaluation standard for speech control concepts in real-world scenarios," *Proc. EUROSPEECH 2003*, Geneva, 2003. pp. 1553 – 1556.
- [5] Gillespie, B. W., Malvar, H. S., Florencio, D. A., "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Proc. ICASSP 2001*, Salt Lake City, Utah, USA, 2001. pp. 3701 – 3704.
- [6] Kinoshita, K., Delcroix, M., Nakatani, T., Miyoshi, M., "Multi-step linear prediction based speech dereverberation in noisy reverberant environment," *Proc. INTERSPEECH 2007*, Antwerp, Belgium, 2007. pp. 3 – 15.
- [7] Lu, X., Unoki, M., and Akagi, M., "Comparative evaluation of modulation-transfer-function-based blind restoration of sub-band power envelopes of speech as a front-end processor for automatic speech recognition systems," *Acoust. Sci. & Tech.*, Vol. 29, No. 6, 2008. pp. 351 – 361.
- [8] Droppo, J. and Acero, A., "Environmental Robustness," In: *Springer Handbook of Speech Processing*. Benesty, J; Sondhi, M. M.; Huang, Y. (Eds.), XXXVI, Springer New York, 2008, ISBN: 978-3-540-49125-5. pp. 653 – 679.
- [9] Raut, C. K., Nishimoto, T. and Sagayama, S., "Model Adaptation for Long Convolutional Distortion by Maximum Likelihood State Filtering Approach," *Proc. ICASSP 2006*, Toulouse, France, 2006. pp. 1133 – 1137.
- [10] Hirsch, H.-G. and Finster, H., "A New HMM Adaptation Approach for the Case of a Hands-free Speech Input in Reverberant Rooms," *Proc. INTERSPEECH 2006*, Pittsburgh, USA, 2006. pp. 781 – 784.
- [11] Sehr, A. und Kellermann, W., "Towards Robust Distant-Talking Automatic Speech Recognition in Reverberant Environments," In: Hnsler, E. und Schmidt, G. (Eds.): *Speech and Audio Processing in Adverse Environments*. Springer Berlin Heidelberg, 2008. ISBN.: 978-3-540-70601-4. pp. 679 – 728.
- [12] Hoffmann, R., Eichner, M. and Wolff, M., "Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system," In Esposito, A., et al. (eds.): *Verbal and Nonverbal Communication Behaviours*, Berlin etc.: Springer, LNAI 4775, 2007. pp. 200 – 218.
- [13] Neumann, J., Gasas, J. R., Macho, D., Hidalgo, J. R., "Integration of audio-visual sensors and technologies in a smart room," *Personal and Ubiquitous Computing*, Vol. 13, Springer London, ISSN: 1617-4909 (print), 1617-4917 (online), 2007. pp. 15 – 23.
- [14] Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M., "Speech recognition with primarily temporal cues," *Science*, 270, 1995. pp. 303 – 304.
- [15] Hermansky, H., Morgan, N., and Hirsch, H. G., "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," *Proc. ICASSP 1993*, 1993. pp. 83 – 86.
- [16] Houtgast, T. and Steeneken, H. J. M., "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acustica*, Vol. 28, 1973. pp. 66 – 73.
- [17] Unoki, M., Sakata, K., Furukawa, M., and Akagi, M., "A speech dereverberation method based on the MTF concept in power envelope restoration," *Acoust. Sci. & Tech.*, 25(4), 2004. pp. 243 – 254.
- [18] Houtgast, T., Steeneken, H. J. M. and Plomb, R., "Predicting speech intelligibility in rooms from the modulation transfer function," I. General room acoustics. *Acustica*, Vol. 46, 1980. pp. 60 – 72.
- [19] Brandstein, M. and Ward, D., "Microphone Arrays: Signal Processing Techniques and Applications," Springer Berlin, 2001. ISBN-10: 3540419535.
- [20] Fehér, T., "Design and optimization of a directional microphone consisting of several single microphones in connection with signal processing," Diploma Thesis, Laboratory of Acoustics and Speech Communication, TU Dresden, 2006. In German.