Dereverberation with an Iterative Least-Squares Technique and Minimum Mean-Square Error Estimation for Automatic Speech Recognition

Florian Müller and Alfred Mertins

Institute for Signal Processing, University of Lübeck, 23562 Lübeck
Email: {mueller,mertins}@isip.uni-luebeck.de
Web: www.isip.uni-luebeck.de

Abstract

This work is about dereverberation for automatic speech recognition. The use of a linear minimum mean-square error estimator for enhancing a recently proposed dereverberation method is investigated. The conducted phoneme recognition experiments show that the resynthesis step, which was done in the original work of the dereverberation method, can be omitted. Furthermore, it is shown that the recognition performance can be increased with the proposed estimator approach under certain reverberant conditions.

1 Introduction

Noise and reverberation have a major impact on the performance of *automatic speech recognition* (ASR) systems. While methods for increasing the robustness against noise have been investigated over several decades, dereverberation has only quite recently become the focus of ASR research. Generally, dereverberation methods can be grouped into three categories. The first category comprises techniques that try to enhance the time signal of an utterance prior to feature extraction, for example [1–3]. The second group of methods tries to adapt the parameters of the acoustic models to the characteristics of the reverberant speech. Besides the training of acoustic models on reverberant speech [4] another commonly used approach is the *maximum-likelihood linear regression* (MLLR) [5, 6], which uses linear transforms to adapt the means and covariances of the acoustic models. The third group of methods tries to increase the robustness to reverberation during the feature extraction, with the RASTA methodology [7] and *cepstral mean normalization* (CMN) [8] being two prominent approaches. Another approach from this group of methods determines *minimum mean-square error* (MMSE) estimates of the clean-speech features within a Bayesian framework [9].

Recently, an *iterative deconvolution technique* (ITD) was proposed [10], which was shown to lead to superior accuracies in comparison to other state-of-the-art dereverberation approaches. This method relies on a *non-negative matrix factorization* (NMF) framework [11], and an advantage of this approach is the low computational cost. In this work we investigate how the ITD method could be combined with a linear MMSE estimator in order improve the ASR performance under reverberant conditions. While the ITD method tries to find the clean-speech components of a reverberant speech signal blindly, the parameters of an MMSE estimator are determined with the help of training data, which adds prior knowledge of the true clean-speech spectral values to the ITD approach.

The paper is structured in the following way: The next section gives an overview of the ITD method and describes the enhancement approach proposed in this work. Section 3 describes the experimental setup and the recognition results. Conclusions and an outlook are given in the final section.

2 Iterative Least-Squares Deconvolution and MMSE Refinement

In the first part of this section, we give a brief overview of the iterative deconvolution method. The proposed enhancement method is described in the second part.

2.1 Review of the Iterative Deconvolution Method

Under time invariant conditions, a reverberated speech signal $y[n]$ can be described as the output of a linear time-invariant system,

$$y[n] = \sum_m x[m]\, h[n-m], \tag{1}$$

where $x[n]$ is the clean speech signal, $h[n]$ is a *room impulse response* (RIR), and n is the discrete time index. Note that the signal model in Eq. (1) does not take additive noise into account. This somewhat unrealistic assumption is handled in [12] by combining the ITD dereverberation method with a noise compensation technique that is supposed to compensate for additive noise and estimation artifacts. The dereverberation problem is to compute $x[n]$ from $y[n]$. However, neither $x[n]$ nor $h[n]$ are known and one has to bring in some a priori information on the nature of the problem to obtain a solution. The method of [10] does this by describing the convolution (1) as a convolution of *short-time Fourier transform* (STFT) magnitude spectra and imposing nonnegativity on the estimated terms. In the following, let $X[n,k]$, $H[n,k]$, and $Y[n,k]$ denote the respective STFT magnitudes of $x[n]$, $h[n]$, and $y[n]$ and let

$$\widetilde{Y}[n,k] = \sum_m X[m,k]\, H[n-m,k]. \tag{2}$$

Here, k is the channel index with $1 \leq k \leq K$. Then, the STFT magnitude spectrum $Y[n,k]$ of a reverberated speech signal $y[n]$ can be approximated with

$$Y[n,k] \approx \widetilde{Y}[n,k]. \tag{3}$$

An iterative least-squares approach that minimizes the errors

$$E_k = \sum_i \left(Y[i,k] - \sum_m \left(X[m,k]\, H[i-m,k] \right) \right)^2, \tag{4}$$

initialized by NMF, is used to find $X[m,k]$ and $H[n,k]$ from $Y[n,k]$. Initialized with the NMF solution from [11] Kumar et al. [10] showed that their proposed ITD method

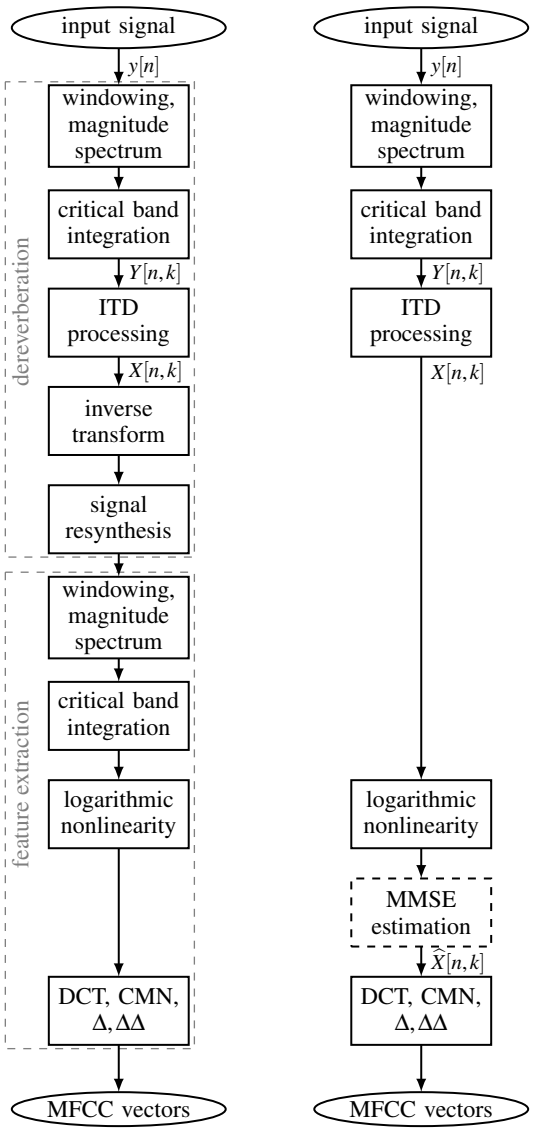

Figure 1: Illustration of the individual feature extraction stages of (left) the original ITD approach [10], and (right) the proposed enhanced approach of this work.

converges to a global optimum with respect to the error criterion (4). An overview of the processing stages of ITD is shown in Figure 1 on the left. Basically, the original processing scheme of the ITD method consists of a dereverberation part, which generates an enhanced time-signal, followed by a feature extraction part, which extracts MFCC features in this work.

2.2 Enhancing ITD

The method proposed in this work is motivated by several aspects. In the originally proposed ITD method an analysis window with a length of 64 ms is used for the time-frequency analysis and dereverberation, and shorter lengths reduce the effectiveness of the approach. The length of 64 ms is in contrast to the analysis window length used for feature extraction in ASR systems, which commonly is 20-25 ms long. While the resynthesis of the time signal, as carried out in [10] and shown in Figure 1 on the left, and subsequent ASR feature extraction circumvents

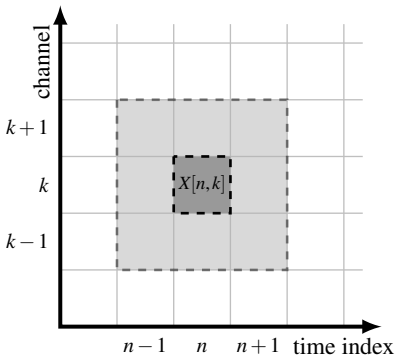

Figure 2: Illustration of considered neighboring spectral values $\mathscr{R}_{n,k}^{S}$ (light gray, here with a first-order neighborhood, i.e. $S=1$ for an enhanced estimation of the clean speech spectral value of $X[n,k]$ (dark gray).

this discrepancy, it increases the computational load of the front-end of the ASR system. Thus, by avoiding the resynthesis the computational costs of the ITD dereverberating are decreased. Here, we present a method that makes use of a channel-dependent linear regression model, which is known as "linear MMSE estimator". As is described in more detail in the following, this approach can be used for compensating the different lengths of the analysis windows as well as the artifacts that are introduced by the ITD dereverberation.

In the following, the MMSE estimation used within this work is explained in detail: Given a vector of observed parameters $\mathbf{r}_{n,k} \in \mathbb{R}^M$, which is assigned to the observed spectral value $X[n,k]$ at time instance n and channel k, a channel-dependent linear estimator $\mathbf{a}_k \in \mathbb{R}^M$ tries to estimate the true parameter as $\widehat{X}[n,k]$, where

$$\widehat{X}[n,k] = \mathbf{a}_k^{\top}\,\mathbf{r}_{n,k}. \tag{5}$$

Recall that $X[n,k]$ denotes the dereverberated spectrogram estimates at the output of the ITD stage. Given a spectral value $X[n,k]$, the set $\mathscr{R}_{n,k}^{S}$ of spectral values of the S-th order neighborhood is formally given by

$$\mathscr{R}_{n,k}^{S} = \{\, X[k+\kappa, n+\nu] \mid \kappa,\nu = -S,\ldots,+S \,\}. \tag{6}$$

Figure 2 illustrates an exemplary first-order neighborhood. As can be seen, the number of elements M of $\mathscr{R}_{n,k}^{S}$ is given by

$$M = 1 + 8\sum_{j=1}^{S} j. \tag{7}$$

Now, the vector of observed parameters $\mathbf{r}_{n,k} \in \mathbb{R}^M$, which is assigned to the spectral value $X[n,k]$ at time instance n and channel k, is defined as the linearized set $\mathscr{R}_{n,k}^{S}$.

The parameters of the k estimators are determined with the help of stereo training data that consists of the ITD-processed spectral values $X[n,k]$ and of the corresponding true clean-speech spectral values $X^{\text{true}}[n,k]$. The two TF representations can either use the same analysis window lengths or use different lengths. In this work, we consider the case in which both X and X^{true} are based on a window length of 20 or 64 ms, as well as the case in which

X is based on a 64 ms analysis window and X^{true} is based on a 20 ms analysis window. The latter approach investigates, whether the estimation of spectral values based on an ASR-typical window length is beneficial for a subsequent recognition stage in this case. With N denoting the number of all available training frames, let $\mathbf{R}_k \in \mathbb{R}^{M \times N}$,

$$\mathbf{R}_k = \begin{bmatrix} \mathbf{r}_{1,k} & \mathbf{r}_{2,k} & \dots & \mathbf{r}_{N,k} \end{bmatrix}^\top, \tag{8}$$

denote all available observed parameters of channel k and let $\mathbf{x}_k^{\text{true}} \in \mathbb{R}^N$ denote all corresponding clean-speech spectral values of channel k. Now, the MMSE estimator $\mathbf{a}_k$ for channel k is chosen such that it minimizes the diagonal components of the estimated correlation matrix of error,

$$\frac{1}{N}\left(\mathbf{x}_k^{\text{true}} - \mathbf{a}_k^\top \mathbf{R}_k\right)\left(\mathbf{x}_k^{\text{true}} - \mathbf{a}_k^\top \mathbf{R}_k\right)^\top. \tag{9}$$

The solution that leads to minimal diagonal components of (9) can be obtained as

$$\mathbf{a}_k = \left(\mathbf{R}_k \mathbf{R}_k^\top\right)^{-1} \mathbf{R}_k \mathbf{x}_k^{\text{true}}. \tag{10}$$

Because it turned out to be beneficial in preliminary experiments, we used the logarithmized magnitudes of the spectral values. To ensure non-negativity of the spectral components, we floored the potentially negative components to a small positive constant. The overall processing scheme with the MMSE estimator as described above is shown in Figure 1 on the right. Compared to the originally proposed ITD approach, the resynthesis step as well as the second time-frequency analysis step are omitted, which decreases the computational costs in comparison to the original formulation of ITD.

3 Experiments

3.1 Data and Experimental Setup

For the performance evaluation of the proposed method we conducted phoneme recognition experiments on the TIMIT database with a sampling rate of 16 kHz and without the SA sentences. The training and test sets consist of 3969 and 1344 utterances, respectively, and are spoken by 630 female and male adults. With a frame shift of 10 ms about $1.1 \cdot 10^6$ frames were available for the training of the MMSE estimators and the acoustic model parameters. We used artificially generated RIRs with different reverberation times T_{60}, which refers to the time it takes the RIR energy to decay by 60 dB. To generate reverberant speech signals in this work, values for T_{60} of 150 ms, 300 ms, and 600 ms were chosen. For the simulation the implementation from [13] of the image method [14] was used. We used a simulated room of dimension $5 \times 4 \times 3$ m. The microphone was located in the center of the room and the source was located 1 m away from the microphone. The original clean speech training set was used for the training of the acoustic models in all experiments. The *hidden Markov model toolkit* (HTK) [15] was used throughout the experiments. The acoustic models were three-state, left-to-right monophone HMMs without state skips. Following the standard procedure on TIMIT [16] the initial phoneme set of 61 phonemes was folded to 48 phonemes and further reduced to 39 phonemes for the evaluation of the phoneme recognition rate. The output distributions were modeled with eight Gaussians and diagonal covariances. The used language model was a bigram model derived from the training data of TIMIT. For the TF analysis a gammatone filter bank with 40 channels was used. The final feature vectors consisted of 39 components, comprising 13 cepstral coefficients together with the corresponding delta and delta-delta features. CMN was applied in all cases.

For each considered reverberation condition, an individual set of estimators was trained. In these experiments we used a first-order neighborhood, i.e., $S = 1$. In practice, an estimate of the reverberation time would be needed, which could be achieved with, e.g., [17]. ITD is initialized with the solution of the mentioned NMF method. The number of iterations for both methods were empirically determined in preliminary experiments.

Table 1: Baseline phoneme recognition rates

	Reverberation time [ms]		
Enhancement	150	300	600
-	**60.7**	38.6	25.6
NMF	58.6	46.7	27.7
ITD	58.6	48.8	32.0

3.2 Baselines

Table 1 shows baseline phoneme recognition rates for the three different reverberation conditions and different enhancement methods. The rows from top to bottom show the accuracies for the cases in which no further enhancement is performed, when only NMF is applied, and when ITD is applied, respectively. In case of $T_{60} = 300$ ms and $T_{60} = 600$ ms it can be seen that the NMF enhancement leads to higher accuracies compared to the case in which no enhancement is done. Surprisingly, the accuracy decreases slightly for $T_{60} = 150$ ms. This might indicate a nonoptimal choice for the parameters or even the optimized error function and has to be further investigated. The application of ITD, which is initialized with the NMF solution, further increases the accuracies for $T_{60} = 300$ ms and $T_{60} = 600$ ms, while keeping the same accuracy as NMF for $T_{60} = 150$ ms.

4 Results for Enhanced Method and Discussion

The first two rows of Table 2 show the recognition rates for different analysis window lengths and the case in which the original ITD method is used without the resynthesis and without MMSE estimation (see also Figure 1 on the right without the MMSE estimation). The analysis window lengths of 20 and 64 ms were considered. It can be observed that the use of a window length of 64 ms is beneficial for reverberation times of 300 and 600 ms in comparison to a window length of 20 ms. By comparing the accuracies of the ITD method with a 64 ms window and without resynthesis with the ITD accuracies from Table 1 one can see that the performance decreases for reverberation times of 150 and 300 ms and increases for a reverberation time of 600 ms.

The last three rows of the table show the accuracies

Table 2: Phoneme recognition rates without resynthesis and with the proposed enhancement method

window length [ms]	Enhancement	Reverberation time [ms] 150	300	600
20	ITD	55.7	46.3	31.4
64	ITD	55.1	48.5	34.3
$20 \rightarrow 20$	ITD + MMSE est.	56.2	47.3	32.8
$64 \rightarrow 64$	ITD + MMSE est.	56.9	**50.6**	**35.6**
$64 \rightarrow 20$	ITD + MMSE est.	55.6	49.7	35.5

that are obtained when the MMSE estimation as described above is used. For these cases, the notation "$A \rightarrow B$" in the left column of the table refers to the analysis window length A of the ITD-processed spectral values X and to the analysis window length B of the true spectral values X^{true}, which was used for determining the parameters of the estimator. Compared to the accuracies as results of ITD without resynthesis as shown in the first two rows, it can be seen that the application of the linear MMSE estimator leads to performance improvements. In average, the increase in accuracy is larger with analysis window lengths of 64 ms than with lengths of 20 ms. As can be seen in the bottom row of Table 1, a conversion of the windows lengths from 64 ms to 20 ms my means of the MMSE estimator does not show any benefits.

5 Conclusions

In this work we investigated a possible way for enhancing a recently proposed dereverberation technique for ASR referred to as ITD. This technique originally generated a dereverberated speech signal, which is subsequently passed to the front-end of an ASR system. In this work, we investigated, whether a linear MMSE estimator can be used such that the resynthesis is omitted and the overall recognition performance of the ASR system is enhanced.

The results of the experiments showed that the combination of ITD and MMSE estimator can lead to increased recognition rates under certain reverberant conditions. Furthermore, it was shown that the use of an analysis window length of 64 ms is preferable for this approach compared to a window length of 20 ms. Using a target window length for the MMSE estimator (20 ms) that is different from the length used by the ITD method (64 ms) did not lead to higher recognition rates than using a fixed analysis window length of 64 ms.

The results of this work can only be seen as preliminary for several reasons. First, the ASR system of this work used monophones as acoustic models to make the experiments more feasible and a triphone systems might give more competitive accuracies. Second, the use of nonlinear estimation techniques might prove beneficial in comparison to the linear MMSE estimator of this work. As pointed out by [10] the initialization and also the number of iterations of the ITD method are crucial for a dereverberation in ASR and might still give room for improvements. Furthermore, a formulation of an MMSE estimator that adapts to the reverberation could avoid the necessity for several individually trained MMSE estimators.

6 Acknowledgments

This work has been supported by the German Research Foundation under Grant No. ME1170/4-1.

References

[1] K. Lebart, J. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acust. United with Acust.*, vol. 87, no. 8, pp. 359–366, 2001.

[2] E. A. P. Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*. PhD thesis, Technische Universiteit Eindhoven, Netherlands, Jun. 2007.

[3] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 534–545, 2009.

[4] L. Couvreur and C. Couvreur, "Blind model selection for automatic speech recognition in reverberant environments," *J. VLSI Signal Proc. Syst.*, vol. 36, no. 2/3, pp. 189–203, 2004.

[5] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, Apr. 1995.

[6] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, Apr. 1998.

[7] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.

[8] A. E. Rosenberg, C.-H. Lee, and F. K. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification," in *Proc. Int. Conf. Spoken Language Processing*, (Yokohama, Japan), pp. 1835–1838, Sept. 1994.

[9] A. Krueger and R. Haeb-Umbach, "Model-based feature enhancement for reverberant speech recognition," *Audio, Speech, and Language Processing, IEEE Trans.*, vol. 18, pp. 1692 –1707, Sept. 2010.

[10] K. Kumar, B. Raj, R. Singh, and R. Stern, "An iterative least-squares technique for dereverberation," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, (Prague, Czech Republic), pp. 5488 –5491, May 2011.

[11] K. Kumar, R. Singh, B. Raj, and R. Stern, "Gammatone sub-band magnitude-domain dereverberation for ASR," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, (Prague, Czech Republic), pp. 4604–4607, May 2011.

[12] K. Kumar, *A Spectro-Temporal Framework for Compensation of Reverberation for Speech Recognition*. PhD thesis, Carnegie Mellon University, Pittsburgh, USA, May 2011.

[13] E. A. P. Habets, "Room impulse response generator for matlab." web resource: http://home.tiscali.nl/ehabets/rir_generator.html, May 2012.

[14] J. B. Allen, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[15] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4.1)*. Cambridge University Engineering Department, Cambridge, UK, 2009.

[16] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37, pp. 1641–1648, Nov. 1989.

[17] R. Ratnam, D. L. Jones, B. C. Wheeler, J. William D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *J. Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.