Joint Speech Dereverberation and Denoising Using Constrained NMF

Nikhil Mohanan[1]

[1]Indian Institute of Technology Bombay

nikhilm@ee.iitb.ac.in

Abstract

Reverberation and background noise corrupt the speech recorded using distant microphones. This affects the performance of Automatic Speech Recognition (ASR) systems. The main objective of the thesis is to implement a framework to improve the ASR results for a distant, multi-channel recording corrupted by reverberation and background noise. The work focuses on improving the front-end of the ASR system to obtain an enhanced speech signal for the corresponding degraded utterances. The enhancement consists of multi-channel processing (source localization and beamforming) and single channel enhancement. Beamforming utilizes spatial information available in multi-channel recording to provide a single-channel enhanced output. The residual reverberation and noise present in the enhanced data are suppressed by the Non-negative Matrix Factorization (NMF) based enhancement. In the initial work, many meaningful constraints on Room Impulse Response (RIR) spectrogram are used to improve the dereverberation performance in a NMF context. The later work uses a separability approximation on RIR spectrogram to obtain an NMF model for reverberation as opposed to convolutive NMF (C-NMF) used in literature. This model is extended to have an NMF model for reverberation in presence of noise. The proposed method performs better than existing C-NMF based methods in objective measures, such as cepstral distance (CD) and speech-to-reverberation modulation energy ratio (SRMR). The ASR performance of the complete framework remains to be evaluated.

Index Terms: NMF, distant speech recognition, reverberation, noise

1. Introduction

The accuracy of automatic speech recognition (ASR) systems has undergone tremendous improvements. ASR systems have evolved from the first ASR system Audrey, which could recognize digits, to the present large vocabulary continuous speech recognition systems. However, such systems are designed to recognize speech recorded using a single close-talking microphone. In many real-world applications like smart homes, robots, conference meetings and voice-controlled personal assistants, speech recording is done using single or multiple microphones placed a few meters away from the source. The ASR performance of distant speech recordings (DSR) is severely degraded by background noise, reverberation, interfering speakers and microphone failures [1]. In order to improve the performance, modifications are required for both the front-end and the back-end of the ASR system. Front-end processing includes enhancing the degraded speech and compensations for the degradations performed in the feature domain. Back-end processing includes improving the acoustic and language model. This work focuses on improving the front-end of the ASR system to jointly handle reverberation and noise present in a multi-channel recording.

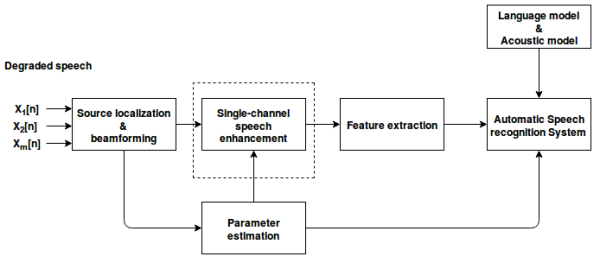

Figure 1: *Block diagram of the proposed framework.*

The effects of reverberation depend on the properties of speech and room impulse response (RIR). Beamforming is one of the common method used to enhance a multi-channel recording [2]. Dereverberation methods proposed in the literature include reverberation cancellation methods, blind deconvolution based methods, and reverberation suppression methods such as spectral subtraction, linear prediction (LP) and non-negative matrix factorization (NMF) based methods [1]. The earliest work on NMF based dereverberation [3] uses a convolutive NMF (referred as C-NMF) model for the reverb spectrogram. Since then many modifications to this have been proposed both in single-channel [4, 5, 6, 7, 8] and multi-channel scenario [9]. The C-NMF model for speech dereverberation was improved by additionally incorporating a NMF model for clean speech [5, 6]. This method is referred to as C-NMF+NMF. Various supervised approaches to handle reverberation in noisy environments have also been proposed [7, 10, 11]. Different regularization on RIRs in single-channel [11, 12] and multi-channel [13] scenario have been proposed leading to better speech enhancement.

2. Proposal Description

The main objective of this thesis is to build a framework to enhance multi-channel degraded recording to get better ASR performance. Figure 1 shows the block diagram of the proposed framework. Beamforming is used to obtain a single-channel enhanced speech from the multi-channel degraded recordings. The performance of beamformer depends on proper estimates of source localization and noise statistics. This is difficult to obtain in the presence of degradations. The residual noise and reverberation present in the beamformer output are further suppressed using single-channel enhancement methods. One objective of this work is in understanding the interaction of multi-channel and single-channel with the ASR system, so as to improve the performance of the ASR system.

The next objective of this work is in improving NMF based single channel enhancement methods. Such methods in literature utilize properties of clean speech spectrogram and noise spectrogram to enhance the degraded speech. These methods ignore the structure present in the RIR spectrogram. In my ini-

Table 1: *Comparison of objective measures for the reverberant and noisy speech, with non-stationary noise (Factory) at* 10 *dB SNR*

	CD				SRMR			
	RIR1	RIR2	RIR3	RIR4	RIR1	RIR2	RIR3	RIR4
Degraded speech	5.28	5.63	5.36	5.68	3.53	2.22	3.25	2.05
C-NMF+NMF [5, 6]	5.60	5.65	5.75	5.82	4.81	3.83	4.83	3.99
R-NMF	5.16	5.43	5.28	5.55	4.50	3.53	4.30	3.84
R-NMF+NMF	**4.77**	**5.12**	**4.80**	**5.17**	**5.50**	**4.32**	**5.15**	**4.31**

tial work, different regularizations on the RIR are imposed for a improved speech dereverberation [12]. These regularizations are motivated by spectral domain representation of the RIR. To obtain better estimates of the RIR and clean speech, three modifications were proposed (i) to obtain a sparse RIR (ii) a frequency envelop constrained RIR and (iii) to include the early part of the RIR. The proposed regularizations helped in improving the speech enhancement measures when compared to other NMF based dereverberation methods. However, one of the limitations for frequency envelop constrained RIR was that it requires prior knowledge of RIR spectrogram.

The latter part of the work was on estimating the frequency envelope of the RIR spectrogram without knowledge of the RIR [14]. Using the separability approximation on RIR spectrogram, a NMF model for reverberation was proposed in contrast to the C-NMF model used in literature [5]. The model uses magnitude spectrogram of the reverb speech and learned clean speech bases to estimate the enhanced speech. Such an approach will allow us to incorporate meaningful constraints in the frequency- and time-domain. This leads to a better speech enhancement, as it has direct control over the estimates of clean speech activations and better RIR estimates. This method is referred to as R-NMF. Figure 2 compares the dereverberated spectrogram obtained using R-NMF with the reference method. The spectrogram obtained using the proposed approach performs better in many regions as indicated in the figure by red boxes. Another advantage of such a model is that it can be easily extended to handle additive noise making it suitable for a noisy reverberant scenario. This method is referred to as R-NMF+NMF.

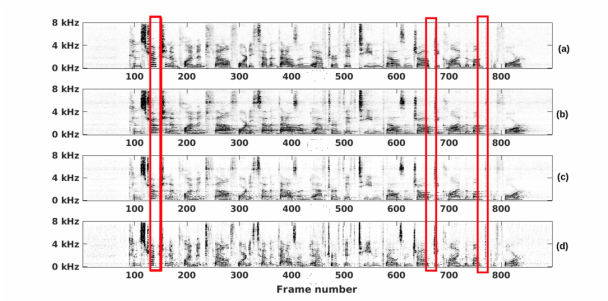

Figure 2: *Spectrogram of (a) Clean speech, (b) Reverb speech, (c) Enhanced speech using C-NMF+NMF, and (d) Enhanced speech using the proposed R-NMF. The regions where R-NMF performs better is shown using red boxes.*

3. Results

Table 1 compares the improvement in objective measures obtained for the proposed single channel enhancement method. The performance is compared with other NMF based reference methods. Degraded data was generated using a subset of 16 TIMIT utterances [15]. Each utterance was convolved with four different measured RIRs and added with a non-stationary (factory) noise at 10dB signal-to-noise (SNR) ratio. 100 speaker specific bases were learned from clean utterances of the speaker. 100 noise bases are learned from noise recordings. The proposed algorithms have better enhancement measures as compared with the reference method.

4. Future Plan

In the proposed single-channel enhancement method, speaker specific clean speech bases are learned from the clean recordings of the same speaker. In many real-world applications, clean utterances of the specific speaker are not available. Further, learning bases for each speaker is a tedious task. One approach used in literature is to learn a large number of clean speech bases vectors, which can be used by a large number of speakers. However, this is computationally expensive. One of the approach to avoid the above limitations is to use of unsupervised bases, where clean speech and noise bases are learned from the degraded data. Noise bases can be learned from the silence regions of the utterances. The performance of the algorithms depends on accuracy in estimation of silence regions. Secondly, the separability approximation of RIR spectrogram ignores the frequency dependence of T_{60}. The approximation needs to be relaxed to take into account the frequency dependency of T_{60}. The separability assumption can be viewed as obtaining a rank-1 NMF approximation for the RIR spectrogram. This can be handled by having a rank-r approximation of the RIR spectrogram instead of the original rank-1 approximation. The final aspect of the work is to improve the ASR performance of the algorithms. Initial experiments indicate that the ASR results do not improve for ASR system trained using clean data, though the enhancement measures show significant improvements. The training-testing mismatch can be one of the reasons for the poor performance. The enhancement algorithms introduce speech distortions to the enhanced data. Due to this, feature vectors estimated for enhanced data differ significantly from the corresponding clean data. The effects of multi-channel and single-channel enhancement algorithms on ASR performance needs to be understood to improve the ASR results. With such a system, we also intend to participate in the 5-th CHiME speech separation and recognition challenge to be held in September 2018.

5. Acknowledgements

I acknowledge the support of my supervisors Prof. Rajbaabu Velmurugan and Prof. Preeti Rao. Part of the work was supported by Bharti Centre for Communication in IIT Bombay, Council of Scientific and Industrial Research (CSIR), India and Tata Consultancy Services (TCS), India.

6. References

[1] N. Patrick and G. Nikolay, *Speech Dereverberation*. New York: Springer, 2010.

[2] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[3] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 45–48.

[4] K. Kumar, R. Singh, B. Raj, and R. Stern, "Gammatone sub-band magnitude-domain dereverberation for ASR," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[5] N. Mohammadiha and S. Doclo, "Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 2, pp. 276–289, 2016.

[6] N. Mohammadiha, P. Smaragdis, and S. Doclo, "Joint acoustic and spectral modeling for speech dereverberation using non-negative representations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[7] D. Baby, T. Virtanen, and J. F. Gemmeke, "Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 11, pp. 1788–1799, 2015.

[8] H. Kallasjoki, J. F. Gemmeke, K. J. Palomaki, A. V. Beeston, and G. J. Brown, "Recognition of reverberant speech by missing data imputation and NMF feature enhancement," in *Proc. REVERB Workshop*, May 2014.

[9] S. Mirsamadi and J. H. L. Hansen, "Multichannel speech dereverberation based on convolutive nonnegative tensor factorization for ASR applications," in *Proc. Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014.

[10] D. Baby, "Non-negative sparse representations for speech enhancement and recognition," Ph.D. dissertation, University of Leuven, 2016.

[11] D. Baby and V. H. Hugo, "Joint denoising and dereverberation using exemplar-based sparse representations and decaying norm constraint," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 2024–2035, 2017.

[12] N. Mohanan, R. Velmurugan, and P. Rao, "Speech dereverberation using NMF with regularized room impulse response," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4955–4959.

[13] M. Yu and F. K. Soong, "Speech dereverberation by constrained and regularized multi-channel spectral decomposition: evaluated on REVERB challenge," in *Proc. REVERB Workshop*, May 2014.

[14] N. Mohanan, R. Velmurugan, and P. Rao, "A non-convolutive nmf model for speech dereverberation," in *Interspeech*, 2018.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic data consortium, Philadelphia*, vol. 33, 1993.