GAMMATONE SUB-BAND MAGNITUDE-DOMAIN DEREVERBERATION FOR ASR

Kshitiz Kumar[1]*, Rita Singh*[1]*, Bhiksha Raj*[2]*, Richard Stern*[1]

Carnegie Mellon University, Pittsburgh, PA, USA[1]
Disney Research, Pittsburgh, PA, USA[2]
Email: {kshitizk, rsingh, bhiksha, rms}@cs.cmu.edu

ABSTRACT

We present an algorithm for dereverberation of speech signals for automatic speech recognition (ASR) applications. Often ASR systems are presented with speech that has been recorded in environments that include noise and reverberation. The performance of ASR systems degrades with increasing levels of noise and reverberation. While many algorithms have been proposed for robust ASR in noisy environments, reverberation is still a challenging problem. In this paper, we present an approach for dereverberation that models reverberation as a convolution operation in the speech spectral domain. Using a least-squares error criterion we decompose reverberated spectra into clean spectra convolved with a filter. We incorporate non-negativity and sparsity of the speech spectra as constraints within an NMF framework to achieve the decomposition. In ASR experiments where the system is trained with unreverberated and reverberated speech, we show that the proposed approach can provide upto 40% and 19% relative reduction respectively in performance.

***Index Terms*—** Dereverberation, Spectral modeling, Spectral decomposition, NMF, Speech recognition

1. INTRODUCTION

Current state-of-the-art ASR systems work quite well in controlled environments where the speech recorded is clean. However, the presence of noise and reverberation effects in real environments can severely degrade the performance of these systems. While a number of algorithms have been proposed for robust ASR in noisy environments [1][2], reverberation remains a challenging problem for ASR. The objective of the current work is to develop an approach for dereverberation that, in addition to improving the quality of the signal, also improves ASR performance.

Reverberation is an acoustic phenomenon that happens when a sound wave traveling in an enclosure is repeatedly reflected by the different surfaces in the enclosure. The multiple reflections cause the sound to persist even after original sound is switched off, causing interference with the current sound. Reverberation for an enclosure is measured in terms of reverberation time (RT), which is the time taken for the signal power to decay by 60-dB from the instant the signal source is switched off. Thus, environments with higher RTs imply greater signal self-interference.

While the human auditory system is surprisingly robust to the effects of reverberation, ASR systems perform poorly even when the reverberation effects in speech are small. There have been many approaches to mitigate reverberation effects in speech. Cepstral Mean Normalization (CMN) was initially proposed for compensating an unknown linear filtering operation whose impulse response had a very short duration, shorter than the analysis window of 25 ms for speech feature. CMN works well to mitigate the channel effects of microphones etc. but does not work in reverberation, as the typical RT for a room extends up to 200-500 ms, which is significantly larger than the typical speech ASR feature analysis window size (25 ms). Consequently, other approaches such as long-term log-spectral subtraction (LTLSS) [3] have been proposed as a direct extension of CMN processing. In LTSS, CMN processing is applied to longer analysis windows (1-2 s). ASR features are then obtained from the reconstructed speech. Recently, sparse-NMF [4], likelihood-maximizing filtering [5] *etc.* have also been proposed for dereverberation.

In this paper we present an approach for dereverberation that works in the Gammatone magnitude spectral domain. In our convolutional model in the spectral domain, reverberated speech spectra are assumed to be the resultant of convolution of clean speech spectra and room response spectra. The framework we present for estimating these is based on a constrained non-negative matrix factorization that uses a least-squares error criterion to decompose reverberated spectra into its convolutive constituents. The constrains that we use are of non-negativity [6] [7] and sparsity (optional) [4] of speech spectra.

The rest of the paper is arranged as follows. In Sec. 2 we describe our model for dereverberation in the Gammatone spectral domain. We also distinguish our model from previously presented work on dereverberation using sparse-NMF [4] and highlight the key differences. In Sec. 3 we describe our technical approach where we present a constrained NMF formulation with the incorporation of the various constraints into its solution. In Sec. 5 we present our experimental results and finally in Sec. 6 we present our conclusions.

2. A MODEL FOR DEREVERBERATION IN THE SPECTRAL DOMAIN

Reverberation is mathematically modeled as a linear system to represent the delayed and attenuated components of the sound in

$$\tilde{s}[n] = s[n] * h[n] \tag{1}$$

where, $s[n]$ is a discrete-time speech signal, $h[n]$ is the impulse response of a linear system (also called the room impulse response (RIR)), $\tilde{s}[n]$ is the reverberated signal and n is the time index. The parameters of the filter $h[n]$ change with changes in the environmental parameters such as size of the room, room configuration, position of objects etc. It is typically assumed that compared to the rate at which the spectral characteristics of speech change, the rate of change in room-response spectral characteristics is slow. As a result, for a short duration (2-3 s) we can assume that $h[n]$ is time-invariant and thus the entire system in (1) becomes a linear time-invariant (LTI) system.

The time-domain model in (1) is a useful abstraction that has been utilized in approaches like [3][8]. Although useful, its direct application (1) in dereverberation for ASR is not consistent with the domain in which ASR is performed. While the LTI system in (1) is in time-domain, ASR systems work (on features derived) in the spectral domain. Our hypothesis is that a more useful model for reverberation, especially for speech recognition, would be one that models reverberation directly in the spectral domain. Such a model would map clean spectra to reverberated spectra. Within the model framework we can then develop dereverberation algorithms that infer the underlying clean spectra from the observed spectra of the reverberated signal under some assumptions of the mapping.

Following this hypothesis we use the model proposed in [9] that represents reverberation in the signal spectral domain as a linear-filtering operation as follows:

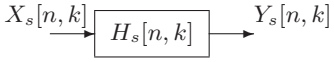


Fig. 1. Modeling reverberation in spectral feature domain

In Fig. 1, the symbols X_s and Y_s denote the spectra of clean speech and reverberated speech respectively. H_s denotes the spectrum of the RIR. n is a frame index and k is a frequency index.

Note that the basic model presented in Fig. 1 has recently been used in other work such as [4][5][7]. However our model differs from these approaches in several key aspects. Firstly, while traditional approaches model reverberation in the Fourier spectral domain, our model works in the Gammatone spectral domain. Secondly, approaches such as [4] present solutions that are based on power spectra, while our model incorporates magnitude spectra. This has special implications for ASR that work in favor of ASR performance. These implications are further discussion in Sec. 4.1 and Sec. 4.2. Also, as further discussed in the next section, the sparsity constraint does not fit well with the Gammatone spectra and accordingly, while our model can incorporate the sparsity constraint, we do not explicitly do so. Further details of the model represented by (1), including its derivation, are given in [9].

3. MATHEMATICAL FORMULATION OF NMF

Our approach for dereverberation using the spectral domain model presented in Sec. 2 is to try to estimate the spectrum of clean speech X_s through a decomposition of the reverberated speech spectrum Y_s into its convolutive components X_s and H_s. In this section we formulate a least-squares error criterion to achieve this decomposition.

In general, reverberation compensation algorithms should not require *a priori* knowledge of nature of the reverberation. This is the case for our algorithm also – we do not require any knowledge of X_s and H_s. Our model of reverberation represents the reverberation effects as the filter H_s, the H_s filter parameters are not observed directly. Rather we attempt to infer the filter parameters through the reverberated spectra Y_s. This problem is however highly unconstrained. According to the model in Fig. 1, there exist infinitely many decompositions of Y_s into X_s and H_s. To constrain the solution space, it becomes necessary to assume some knowledge about either X_s or H_s that we can use as constraints. In our work, we choose two such constraints. One is that the spectral components are non-negative *i.e.* all the elements in X_s and H_s are ≥ 0. This is apparent since the magnitude spectra are inherently non-negative. The second assumption is an optional one, wherein we assume that the clean spectra X_s are sparse. Later in this paper we discuss these constraints in greater detail.

To solve the problem of decomposition we use a non-negative matrix factorization (NMF) framework. NMF was initially proposed for data clustering application in [6]. It was further developed and applied in audio applications in [7], and for speech signal dereverberation in [4]. We use the NMF paradigm in [4][7] to build our framework for dereverberation for ASR.

Next we consider the mathematical formulation of NMF. We first assume that our actual observation sequence is $Z_s[n,k]$, which is approximately Y_s[n,k]

$$Z_s[n,k] \approx Y_s[n,k] = X_s[n,k] * H_s[n,k] \tag{2}$$

The difference between Z_s and Y_s can result from observation noise or from the error in decomposing Z_s into the convolutional components X_s and H_s. Using (2), we define our objective to be the minimization of the mean-squared error between Z_s and Y_s. This objective function is minimized by a gradient descent process that guarantees *at least* a locally optimal solution. We further impose the non-negativity and sparsity constraints [4] as defined below:

$$\begin{aligned} \text{Min. } E = &\sum_i \Big(Z_s[i,k] - \sum_m \big(X_s[m,k]\, H_s[i-m,k] \big) \Big)^2 \\ &+ \lambda \sum_i X_s[i,k]^p \\ \text{Where } X_s[n,k] \geq 0,\ &H_s[n,k] \geq 0,\ \sum_n H_s[n,k] = 1 \end{aligned} \tag{3}$$

where we also constrain the $H_s[n]$ to sum to 1 to avoid scaling problems. Note that sparsity implies that while a small number of spectral components in X_s are expected to exhibit high values, most other components have very small (negligible) values. Note also that of the many ways that exist to include sparsity constraints in an NMF framework, we choose to use the L_1-norm. The first term in the objective function (3) minimizes the mean-squared error and the second term imposes sparsity on X_s. The optimization is solved subject to the stated non-negativity constraints on X_s and H_s. Corresponding to the L_1 norm, we choose $p = 1$ in (3).

3.1. Minimization of the Objective Function in an NMF Framework

We minimize the objective function in (3) by a variant of the gradient descent approach that ensures that the spectral components at the end of each iteration of the gradient descent are non-negative. Noting that $p = 1$ in (3), the derivative of the objective function with respect to X_s is

$$\frac{\partial E}{\partial X_s[n,k]} = -2 \sum_i \big(Z_s[i,k] - Y_s[i,k] \big)\, H_s[i-n,k] + \lambda \tag{4}$$

with the X_s update equation being $\bar{X}_s[n,k] = X_s[n,k] - \eta_s \frac{\partial E}{\partial X_s[n,k]}$, where η_s is the learning-rate parameter. Note that in general there is no guarantee that the updated $\bar{X}_s$ is non-negative. However, we can select a special value of η_s to impose non-negativity. We choose $\eta_s = \frac{X_s[n,k]}{2 \sum_i Y_s[i,k] H_s[i-n,k] + \lambda}$. Incorporating the above value of η_s in (4), the updates become:

$$\begin{aligned} \bar{X}_s[n,k] &\leftarrow X_s[n,k]. \frac{\sum_i Z_s[i,k] H_s[i-n,k]}{\sum_i Y_s[i,k] H_s[i-n,k] + \lambda/2} \\ \bar{H}_s[n,k] &\leftarrow H_s[n,k]. \frac{\sum_i Z_s[i,k] X_s[i-n,k]}{\sum_i Y_s[i,k] X_s[i-n,k]} \end{aligned} \tag{5}$$

The updates for H_s can be derived in parallel to the X_s updates in (5). The iterative update is done for a specified number of iterations. Further, given a non-negative initialization, the updates are guaranteed to be non-negative. Eq. 5 provides iterative updates for the output of a particular sub-band indexed by k. Similar processing will also be applied individually to each of the sub-bands. The NMF optimization will at least reach a locally optimal solution. While the estimated X_s may not be identically equal to the actual clean spectra, it is expected that the processing will result in a solution for X_s that will be largely dereverberated.

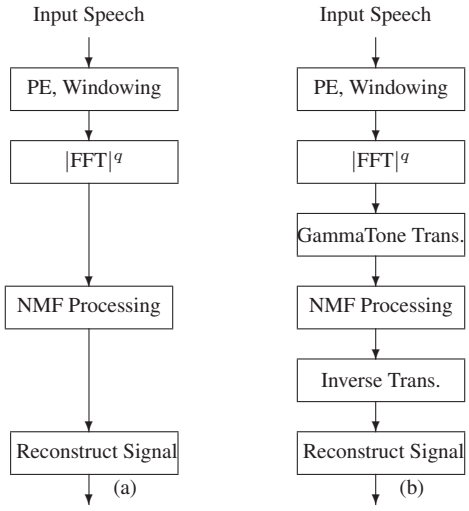


Fig. 2. (a) NMF processing in frequency domain, (b) NMF processing in Gammatone frequency domain.

Fig. 2 presents both, the general procedure for frequency domain NMF processing for dereverberation and our specific approach using Gammatone spectra. In both, the speech signal is first pre-emphasized (PE) with a causal filter with a single pole at $z = 0.97$. It is then windowed and FFT analysis is done on the windowed signal. In Fig. 2(a) which represents NMF processing in the Fourier frequency domain, the NMF decomposition is directly applied individually to each of the FFT channels. In contrast, in Fig. 2(b) which represents our method, NMF processing is applied to each individual channel of the Gammatone filtered spectra, and this is followed by an inverse transformation. Gammatone sub-bands are obtained from the Fourier frequency sub-bands via the Gammatone matrix $G_s[k\prime, k]$, that stores Gammatone frequency response for the $k\prime$ Gammatone sub-band against each of the k Fourier frequency sub-bands. NMF processing is applied on the product $Y_s[n,k].G_s[k,k\prime]$. The NMF processed spectra in the Gammatone domain is multiplied with pseudo-inverse of G_s to obtain the processed Fourier frequency components, from which the signal is reconstructed. Since our processing is done on individual channels in the Gammatone filtered magnitude spectral domain, we call our approach *Gammatone sub-band magnitude-domain dereverberation*. Finally in both the cases the signal is optionally reconstructed or feature vectors for speech recognition may be derived from the resultant dereverberated spectra.

4. KEY FEATURES OF GAMMATONE SUB-BAND NMF

In this section we highlight some key aspects of our proposed approach as shown in Fig. 2(b).

4.1. Advantage of using Magnitude spectra over Power spectra

The model in (2) is an approximation and will in general incur an approximation error E_s as follows:

$$Y_s[n,k] = \hat{Y}_s[n,k] + E_s[n,k] = X_s[n,k] * H_s[n,k] + E_s[n,k] \quad (6)$$

We have empirically observed that the approximation error E_s is lower in the magnitude spectral domain than in the power spectral domain. Experimentally we found that the power of E_s is usually about 13-dB below the power of $\hat{Y}_s$ in the power spectral domain. In contrast, in the magnitude spectral domain, we observed an approximation error attenuation of 17-dB. Thus, working in the magnitude-spectral domain incurs lower error. In our approach shown in Fig. 2(b) setting the parameter $q = 1$ results in magnitude domain NMF processing, and $q = 2$ results in power domain processing. We will refer to magnitude domain processing as "M-NMF" and power domain processing as "P-NMF".

4.2. Advantage of using Gammatone Sub-bands

Processing in the Gammatone domain provides two key benefits. Firstly, the Gammatone sub-bands apply a perceptual weighting to the signal and emphasize the frequency regions where the speech signal is supposed to be dominant for better perception. This directly benefits the quality of the clean signal obtained through the decomposition. Secondly, working in Gammatone sub-bands offers significant saving in computation: there are about 257 sub-bands for a 512 points FFT in Fourier frequency NMF as against only about 40-80 Gammatone sub-bands for the same processing. We also get a significant practical advantage since we estimate fewer parameters from the same overall data. We refer to Gammatone based NMF processing as "GNMF".

4.3. Using Different H_s for Different Sub-bands

In general, we expect the $H_s[n,k]$ in (2) to be different for each of the different sub-bands indexed by k. This is expected to result in a more effective solution for X_s than can be obtained by using the same H_s for all the sub-bands. We verify this empirically in the experimental section. To use the same H_s across all sub-bands, the updates in (5) can be adapted as follows:

$$\bar{H}_s[n,.] \leftarrow H_s[n,.].\frac{\sum_k \sum_i Z_s[i,k] X_s[i-n,k]}{\sum_k \sum_i Y_s[i,k] X_s[i-n,k]} \quad (7)$$

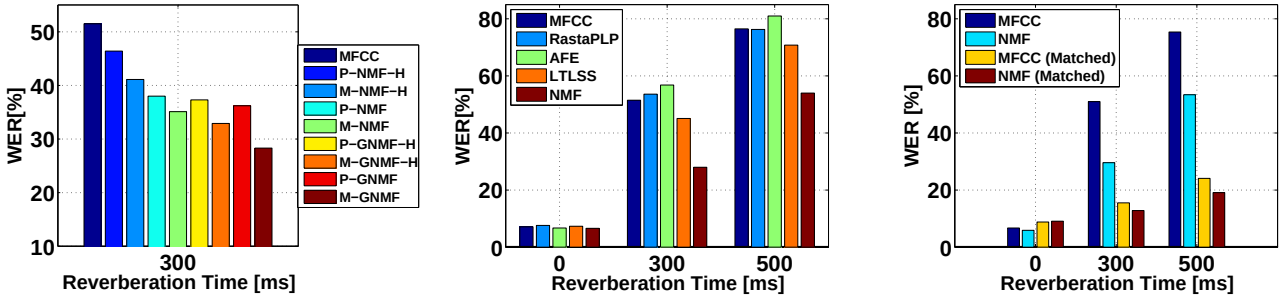
We refer to NMF with same H_s across all sub-bands as "NMF-H".

5. EXPERIMENTAL RESULTS

We applied the NMF [1] formulation in Sec. 3 to the problem of dereverberation for ASR. In our experiments, we simulated reverberation effects to various degrees in the the DARPA Resource Management (RM) database, dereverberated the signals, and then measured the recognition accuracy on dereverberated signals using matched and mismatched recognizers. The ASR system we used for training and decoding speech was the CMU Sphinx-3 [2] opensource system. We used 13-dimensional Mel-frequency cepstra (MFC) as features for ASR. For actual recognition, these were augmented with

[1] NMF software will be available at `http://www.cs.cmu.edu/~robust/archive/algorithms/NMF_ICASSP2010/`.

[2] Available online at `http://cmusphinx.sourceforge.net/html/cmusphinx.php`



(a) (b) (c)

Fig. 3. (a) WER comparisons for different flavors of NMF (b) NMF WER comparisons for clean-training (c) NMF WER comparisons for matched-training.

13-dimensional delta and double-delta cepstra. CMN was done in all cases. The acoustic models consisted of 3-state Bakis topology HMMs with a mixture of 8 Gaussians per state, and a total of 1000 tied states. A bigram language model was used in all experiments.

Utterances in the RM database were artificially reverberated with different RTs [9], as shown in Fig. 3(b). In the first experiment, NMF processing methods as shown in Fig. 2 were applied to dereverberate the utterances. We used 15-20 iterations of NMF processing with a window size of 64ms for the NMF processing, reconstructed the speech and extracted conventional MFC features for ASR from the reconstructed speech. These use a window size of 25ms. In Fig. 3(a), we present our experimental results with the different flavors of the NMF processing in Secs. 4.1, 4.2, and 4.3. Note that the bar titled "P-NMF" shows ASR results for conventional sparsity constrained power domain NMF [4], while the bar titled "M-GNMF" shows the performance obtained with our approach. Experimentally, we found that sparsity was not helpful for the Gammatone sub-bands and hence not applied. A small sparsity [4] factor was applied in Fourier frequency domain.

Overall, we note that NMF processing in gammatone bands provides 20-25% relative reduction in word error rate (WER) over the same processing in Fourier frequency domain. NMF processing in magnitude domain provides 10-18% relative improvement over power domain processing. Further, different H_s for different sub-bands provide 10-20% relative improvement over same H_s for all sub-bands. We thus finalize NMF processing in magnitude domain with Gammatone sub-bands and different H_s for different sub-bands as our baseline NMF processing and now ownards refer to the experiment "M-GNMF" as "NMF".

In Fig. 3(b), we plot word error rate percent (WER%) results for the case where the system is trained with clean (unreverberated speech) and tested on dereverberated speech. We see that here the relative reduction in WER is limited to 15-20% for the baseline dereverberation algorithms. The NMF processing provides 45% relative reduction WER at RT-300 ms which is substantially better than any of the baseline algorithms.

In Fig. 3(c), we plot WER results for the case when the ASR system is trained on the same kind of speech as it is tested on (matched-condition training). We note from these experiments that algorithm is able to improve over matched-condition testing, and to substantially improve over clean-condition testing. Note that it is usually very difficult to improve over matched-condition testing in ASR. It is usually used as a gold-standard in many instances. Matched-NMF provides an additional 19% relative reduction in WER over simple matched training and testing with MFCs from reverberated speech.

6. CONCLUSION

We have presented an NMF-based approach for dereverberation of speech signals in the Gammatone sub-band domain. This has specific advantages for ASR that we have experimentally shown to be valid. The algorithm presented is able to improve WER performance by about 15% relative to matched-condition training, which has been generally observed to be a performance threshold that is hard to exceed. The algorithm results in 30-40% WER reduction under mismatched conditions, where the system is trained on clean speech but attempts to recognize reverberated speech, or dereverberated speech as the case may be.

7. ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation (Grant IIS-10916918) and the C. Stark Draper Laboratory.

8. REFERENCES

[1] R. Singh, R. M. Stern, and B. Raj, "Signal and feature compensation methods for robust speech recognition," *Noise Reduction in Speech Applications, Ed. G. Davis, Chapter 9*, pp. 221–246, 2002.

[2] R. Singh, B. Raj, and R. M. Stern, "Model compensation and matched condition methods for robust speech recognition," *Noise Reduction in Speech Applications, Ed. G. Davis, Chapter 10*, pp. 247–278, 2002.

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

[4] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," *Proc. IEEE ICASSP*, pp. 45–48, 2009.

[5] K. Kumar and R. M. Stern, "Maximum-likelihood-based cepstral inverse filtering for blind speech dereverberation," *in Proc. IEEE ICASSP*, 2010.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, 1997.

[7] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," *Proc. ICA*, 2004.

[8] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," *Proc. ICASSP*, 1996.

[9] K. Kumar, "A spectro-temporal framework for compensation of reverberation for speech recognition," 2010, Ph.D. Proposal, Dept. of ECE, Carnegie Mellon University.