AN ITERATIVE LEAST-SQUARES TECHNIQUE FOR DEREVERBERATION

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

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

ABSTRACT

Some recent dereverberation approaches that have been effective for ASR applications, model reverberation as a linear convolution operation in the spectral domain, and derive a factorization to decompose spectra of reverberated speech in to those of clean speech and of the room-response filter. Typically, a general NMF framework is employed for this. In this work[1] we present an alternative to NMF and propose an iterative least-squares deconvolution technique for spectral factorization. We propose an efficient algorithm for this and experimentally demonstrate it's effectiveness in improving ASR performance. The new method results in 40-50% relative reduction in word error rates over standard baselines on artificially reverberated speech.

Index Terms— Dereverberation, Spectral Decomposition, Iterative Least-Squares, ASR, NMF

1. INTRODUCTION

The presence of noise and reverberation in speech signals causes a performance of automatic speech recognition (ASR) systems to degrade significantly. While effective algorithms for the mitigation of various types of noise have been proposed and used effectively in real applications over the past decade, we have seen the emergence of good dereverberation techniques for speech only relatively recently [1][2][3]. These algorithms different models of reverberation and operate under different assumptions. In this work we focus on those techniques that use spectral domain convolutional models for dereveberation. So far the related approaches reported in the literature used non-negative matrix factorization (NMF) to decompose the spectra of dereverberated speech into clean speech and of a room-response filter.

A convolution operation as we know, takes 2 arguments as its operands, say $x[n]$ and $h[n]$, and results in an output $y[n]$ which represents the signal obtained by passing x through an linear time-invariant (LTI) filter with filter parameters h. We mathematically represent convolution as $y[n] \leftarrow x[n] * h[n]$, where the symbol $*$ represents the convolution operation. Convolution jointly maps x and h to y and this mapping is unique. The deconvolution operation is the inverse of convolution and can be written as $y[n] \rightarrow x[n] * h[n]$. Thus it takes a single argument y as its input and factorizes y into x and h.

While convolution operation results in a unique solution, deconvolution typically results in infinitely many solutions. This is because deconvolution suffers from scaling ambiguity *i.e.* if $x[n]$ and $h[n]$ in $y[n] \rightarrow x[n] * h[n]$ constitute a valid solution, then $x_c[n] = cx[n]$ and $h_c[n] = h[n]/c, \forall c \neq 0$ also constitute an equally valid solution. Aside from this scaling ambiguity there is also an issue of variable assignment *i.e.* $y[n] \rightarrow h[n] * x[n]$ is also a valid solution, so essentially the variables $x[n]$ and $h[n]$ can be reversed in order in the deconvolved output. If either $x[n]$ or $h[n]$ is known, the deconvolution operation may have unique solutions. However, if both $x[n]$ and $h[n]$ are unknown and the problem is unconstrained, there may be infinity of solutions. In such situations, we use some knowledge from the physics of the problem to find an acceptable solution. Some of the successful deconvolution techniques that have been applied for dereverberation are ICA based [4] deconvolution, which assume that the operands are non-Gaussians and Wiener-filter deconvolution which minimizes L_2-norm which implicitly assumes a Gaussian-distribution on the error. Recently, as mentioned earlier in this section, NMF based approaches [5] have based applied effectively for deconvolution [6] especially for audio and speech related problems. NMF is further discussed in Sec. 2.4. In this paper we propose an alternate method for deconvolution. Our method is an iterative least-squares error minimization technique that works in the magnitude spectral domain.

The rest of the paper is arranged as follows. In Sec. 2.2 we describe our ITD approach in detail. In Sec. 3 we discuss issues related to its computational efficiency. In Sec. 4 we present our experimental results using ITD and finally in Sec. 5 we present out conclusions.

2. ITERTIVE DECONVOLUTION (ITD) FOR SPECTRAL FACTORIZATION

In this section, we present the details of our proposed approach for spectral deconvolution based dereverberation for ASR.

We define $X_s[n,k]$ and $Y_s[n,k]$ as respectively the clean and reverberated spectra with $H_s[n,k]$ being the filter spectra representing the spectra of room-response filter. n is a frame index and k is an index to a particular narrow-band or a sub-band frequency channel. The dereverberation problem for ASR is to infer the spectrum of the clean signal *i.e.* X_s from the observed Y_s. We thus seek the following decomposition:

$$Y_s[n,k] \rightarrow X_s[n,k] * H_s[n,k] \tag{1}$$

and formulate a least-squares solution to it as

$$E = \sum_i \Big(Y_s[i,k] - \sum_m \big(X_s[m,k] - H_s[i-m,k] \big) \Big)^2 \tag{2}$$

We now discuss how we can derive a least-squares formulation for the solution of (2).

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

2.1. Interpretation of Convolution in terms of Matrix Multiplication

The convolution operation in (2) can be expressed as a matrix operation in order to derive a least-squares formulation for the iterative minimization of the error. Note that initialization is an important aspect of this solution. Starting from initial estimates of either X_s or H_s, we iteratively obtain updated values of these variables, until a convergence criterion is achieved. We illustrate the above with the following example. Consider a simple convolution operation:

$$Y = X * H \tag{3}$$

where, $X = [x_0\, x_1\, x_2]^T$, $H = [h_0\, h_1]^T$, and $Y = [y_0\, y_1\, y_2\, y_3]$. The convolution operation in (3) can be equivalently expressed in terms of the following two matrix operations

$$Y = \underbrace{\begin{bmatrix} x_0 & 0 \\ x_1 & x_0 \\ x_2 & x_1 \\ 0 & x_2 \end{bmatrix}}_{T_X} H = \underbrace{\begin{bmatrix} h_0 & 0 & 0 \\ h_1 & h_0 & 0 \\ 0 & h_1 & h_0 \\ 0 & 0 & h_1 \end{bmatrix}}_{T_H} X \tag{4}$$

Note that in in (4) T_H is a Toeplitz matrix of size $Dim(Y) \times Dim(X)$, similarly H_X is a Toeplitz matrix of size $Dim(Y) \times Dim(H)$, where $Dim(Y)$ indicates the dimensionality of the Y vector.

2.2. Iterative Least Square Deconvolution

We now build on our matrix perspective of convolution as in (4) to arrive at an iterative solution for estimating X_s and H_s jointly given an initial estimate of either H_s or X_s.

In order to do this, we rewrite the convolution in (1) as a matrix operation as follows

$$\underbrace{Y_s[.,k]}_{Y_k} \approx \underbrace{T_{X_s[.,k]}}_{T_{X_k}} \underbrace{H_s[.,k]}_{H_k}, \quad \underbrace{Y_s[.,k]}_{Y_k} \approx \underbrace{T_{H_s[.,k]}}_{T_{H_k}} \underbrace{X_s[.,k]}_{X_k} \tag{5}$$

where, $T_{X_s[.,k]}$ is a Toeplitz matrix consisting of elements from $X_s[.,k]$. For brevity, in the rest of this paper we will refer to $X_[.,k]$ as X_k. Note that the T_{X_k} is a matrix of size $Dim(Y_k) \times Dim(H_k)$. Similarly $H_s[.,k]$ is equivalently written as H_k. T_{H_k} is a Toeplitz matrix, consisting of elements from H_k, and is of size $Dim(Y_k) \times Dim(X_k)$. Our goal is to use (5) to obtain a factorization for Y_k in terms of X_k and H_k.

Given an initial estimate of either H_k or X_k, the spectral factorization in (1) can be solved by least-squares error optimization using the framework in (5). Starting with an initial estimate for say, H_k we can obtain an updated $\bar{X}_k$ in the following manner:

$$\begin{aligned} \bar{X}_k &= \arg\min_{X_k} ((Y_k - T_{H_k} X_k)^T (Y_k - T_{H_k} X_k)) \\ &= \underbrace{(T_{H_k}^T T_{H_k})^{-1} T_{H_k}^T}_{T_{H_k}^+} Y_k \\ \bar{H}_k &= \underbrace{(T_{\bar{X}_k}^T T_{\bar{X}_k})^{-1} T_{\bar{X}_k}^T}_{T_{\bar{X}_k}^+} Y_k \end{aligned} \tag{6}$$

where, $T_{H_k}^+$ is the pseudo-inverse of T_{H_k}. In a manner similar to the $\bar{X}_k$ updates, we can obtain updated values of $\bar{H}_k = T_{\bar{X}_k}^+ Y_k$. Note that for updating H_k, we use $\bar{X}_k$. In practice, the algorithm may start with an initial estimate of either H_k or X_k. We will discuss the initialization of H_k and/or X_k in more detail later in this section. If say the algorithm starts with an estimate for H_k, then we first evaluate the Toeplitz matrix T_{H_k}, and using (6), we obtain $\bar{X}_k$. The algorithm then reiterates using the $\bar{X}_k$ as an initial estimate to obtain $\bar{H}_k$, and repeats.

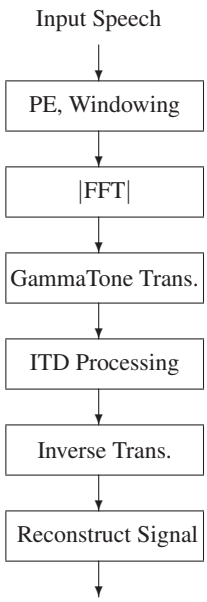

Fig. 1. Iterative Deconvolution (ITD) in sub-band frequency domain.

Note that thus far the methodology proposed above applies to any generic deconvolution problem. Since our goal is use this algorithm specifically for dereverberation of speech signal through a representation in the magnitude/power spectral domain, we enforce the constraint that the constituent spectra being non-negative. Thus, our goal is to arrive at a non-negative decomposition and to this end, we floor the potential negative values in $\bar{X}_k$ and $\bar{H}_k$ to a small positive constant.

2.3. Proof of convergence

We now show that the above method is guaranteed to converge. Note that since the error criterion represented by (2) is quadratic this also implies that the method will converge to a globally optimal solution. Assuming that we begin iteration of the algorithm with an initialization for say H_k and consequently iteratively obtain $\bar{X}_k$ and $\bar{H}_k$ to complete one iteration of the algorithm, it is trivial to show that

$$L_2(Y_k - X_k * H_k) \geq L_2(Y_k - \bar{X}_k * H_k) \geq L_2(Y_k - \bar{X}_k * \bar{H}_k) \tag{7}$$

where, L_2 is the standard mean squared-error metric. The inequality in (7) holds since the updates $\bar{X}_k$ and $\bar{H}_k$ are obtaining by minimizing the squared-error metric in (6). Thus, the iterative algorithm guarantees error minimization at each iteration. Further, since the error function is convex in its argument, each of the individual updates are guaranteed to be their global best solution.

Note that while the above proves that the algorithm is guaranteed to reach a global optimum, the optimum itself is only for the error and not for the individual values of X_k and H_k. The estimated values for these depend on the initialization used. Initialization is therefore an important aspect of the proposed algorithm and we will next consider this in detail.

2.4. Non-Negative Matrix Factorization based Initialization

As discussed above, our deconvolution approach requires a good initialization. This is especially important for speech dereverberation in ASR because we want to get realistic solution for the spectra of the clean signal X_k. In order to obtain realistic values of X_k it is obviously necessary to initialize the algorithm with realistic values of H_k. We therefore use an estimate of H_k that we obtain from an initial NMF decomposition procedure. Note that this procedure is only used for initialization (which could be replaced by any other procedure that could give a realistic initial value for H_k). Our ITD algorithm itself does not incorporate NMF in its solution.

NMF solves the deconvolution problem by imposing non-negativity of the spectra as a constraint to guide the optimization. Optionally a sparsity constraint can also applied on the resulting X_k. We will only list the NMF updates here and refer the reader to [7] for further details. The NMF update equations for H_k and X_k are as follows:

$$\begin{aligned} \bar{X_k}[n] &\leftarrow X_k[n].\frac{\sum_i Y_k[i]\, H_k[i-n]}{\sum_i Y_k[i]\, H_k[i-n] + \lambda} \\ \bar{H_k}[n] &\leftarrow H_k[n].\frac{\sum_i Y_k[i] X_k[i-n]}{\sum_i Y_k[i] X_k[i-n]} \end{aligned} \tag{8}$$

where, λ is an optional sparsity parameter.

2.5. Overall ITD Approach

We present the overall ITD based factorization in Fig. 1. We obtain the spectra of the reverberated signal in the conventional manner. We first preemphasize the time domain signal and window it. We then perform the FFT transform on the windowed signal. The ITD decomposition can be applied on these signal spectra. However in our prior experiments [7], we found that the decomposition algorithms give better result for ASR when we apply the ITD to sub-bands in the Gammatone frequency domain. Working in the Gammatone spectral doain also reduces the computational and logistic requirements.

Accordingly, in our work we applied ITD in the Gammatone spectral domain. After ITD processing is done, the signal is reconstructed after an inverse Gammatone transformation. If ASR is to be performed on the resulting speech signal, we compute mel-frequency cepstra (MFC) from the reconstructed signal. In the next section, we discuss computational issues relating to ITD in detail.

3. COMPUTATIONAL OPTIMIZATION

In this section we propose two strategies for computational cost and memory saving in the implementation of the algorithm proposed in Sec. 2.2. In one strategy we achieve substantial saving in computation (thereby achieve greater computational efficiency) by efficiently evaluating the two matrix products $T_{H_k}^T T_{H_k}$, and $T_{H_k} Y_k$ in (6). In the second strategy a significant improvement in speech is obtained by utilizing the Toeplitz property of the matrix in (6) which allows us to use the Levinson recursion.

3.1. Avoiding Direct Matrix Inversion and using Correlations

A naive approach for obtaining the updates in (5) would be to start with a given H_k, build a Toeplitz matrix T_{H_k}, and obtain the pseudo-inverse $T_{H_k}^+$ explicitly. After this an updated $\bar{X_k}$ would be obtained. The cost of this explicit matrix inversion is $O(N^3)$ for a square matrix of size $N \times N$. Since the objective in (5) is only to obtain $\bar{X_k}$, we can avoid the explicit matrix inversion by reformulating the solution in (5) as follows:

$$T_{H_k}^T T_{H_k} \bar{X_k} = T_{H_k}^T Y_k, \quad T_{X_k}^T T_{X_k} \bar{H_k} = T_{X_k}^T Y_k \tag{9}$$

which can be solved using Cholesky decomposition in $O(N^3/3)$. This represents a significant saving over the naive pseudo-inverse based approach in (6).

The approach in (6) still requires building T_{H_k}, $T_{H_k}^T T_{H_k}$ and $T_{H_k} Y_k$ matrices sequentially. This incurs substantial computational cost, which we can optimize by directly obtaining the products $T_{H_k}^T T_{H_k}$ and $T_{H_k} Y_k$ without having to first build the Toeplitz matrix T_{H_k}. We can do so by observing that the product $T_{H_k}^T T_{H_k}$ is the standard autocorrelation matrix of H_k, and that $T_{H_k} Y_k$ is be a vector of the coefficients of the cross-correlation of H_k and Y_k. For later use, we rewrite (10) as follows:

$$\Phi_{H_k} \bar{X_k} = P_{H_k Y_k}, \quad \Phi_{\bar{X_k}} \bar{H_k} = P_{\bar{X_k} Y_k} \tag{10}$$

where the matrices Φ_{H_k}, $\Phi_{\bar{X_k}}$ respectively comprise the correlation coefficients of H_k and $\bar{X_k}$; and the matrices $P_{H_k Y_k}$ and $P_{\bar{X_k} Y_k}$ respectively comprise the cross-correlation coefficients. Specifically, $P_{H_k Y_k}[m] = \sum_i (Y_k[i] H_k[i-m])$. Thus, using correlation and cross-correlation terms, we can avoid building the giant T_{H_k} matrix consisting of $Dim(Y_k) \times Dim(X_k)$ number of elements. Instead we can work with just $Dim(X_k)$ correlation coefficients for H_k and $Dim(Y_k)$ cross-correlation coefficients between H_k and Y_k.

3.2. Levinson Recursion

A naive solution for (10) can be obtained through Gaussian-Elimination or Cholesky decomposition, but Φ_{H_k} is a Toeplitz matrix consisting of only $Dim(X_k)$ unique elements. This makes it suitable for the use of Levinson recursion for solution of the least-squares problem in (10). We refer the reader to [8] for specific details of this method and simply give an outline here. The computational complexity of Levinson recurion is $\mathcal{O}(N^2)$ which represents a significant saving over $\mathcal{O}(N^3/3)$ computations required by the Cholesky decomposition.

The Levinson recursion approach is as follows. For a given $(N \times N)$ Toeplitz matrix T^N, the aim of the recursion is to build the following forward and backward vectors

$$T^N f^N = e_f^N, \quad T^N b^N = e_b^N \tag{11}$$

where f^N and b^N are vectors of length N and are respectively referred to as forward and backward vectors. e_f^N and e_b^N are also vectors of length N and $e_f^N = [1 \underbrace{0 \cdots 0}_{N-1}]^T$ and $e_b^N = [\underbrace{0 \cdots 0}_{N-1} 1]^T$. The algorithm is initialized with a sub-Toeplitz matrix T^1, which is in fact a single element $T^N(1,1)$, for which $f^1 = b^1 = 1/T^1$, trivially. From this initialization the algorithm evaluates f^2 and b^2 and upto f^N and b^N. The algorithm then recursively solves the linear least-squares solution for (10) using the backward vectors [8].

4. EXPERIMENTS AND RESULTS

We conducted several experimets to test the effectiveness of the proposed ITD algorithm [2]. in improving the performance of a speech recognition system. In all experiments, we artificially reverberated

[2] Software will be available at http://www.cs.cmu.edu/ ~robust/archive/algorithms/ITD_ICASSP2010/

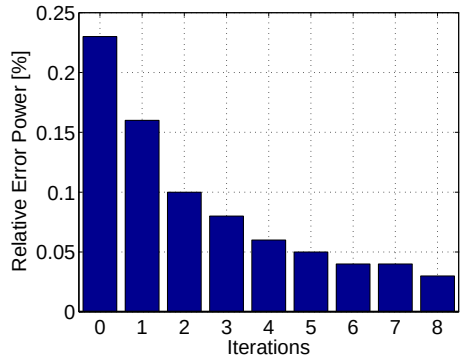

Fig. 2. Residual reconstruction error in ITD factorization.

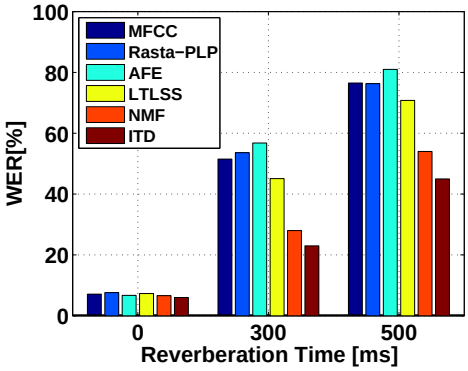

Fig. 3. WER comparisons for ITD. All of the systems include CMN.

data from a standard clean speech database – the Darpa Resource Management (RM) database available from the Linguistic Data Consortium. We used several RT values which are shown in Fig. 3. The reverberated speech signals were then dereverberated using the ITD algorithm, and other available algorithms for comparison. Features for ASR were then extracted from them.

Details of the ASR setup are as follows: the ASR system we used was the CMU Sphinx-3 opensource speech recognition system. We trained the system on clean speech and tested its performance on dereverberated speech. The test data were different from the training data in all the cases, and were the designated training and test sets in the RM database. The acoustic models were all 3-state left-to-right Bakis topology HMMs with no skips permitted between states. Each state output distribution was modeled by a mixture of 8 Gaussians. The total number of tied states used was 1000. The language model used was a standard bigram model for the RM task, built in-house using the CMU Language modeling toolkit. The features used were conventional MFC features augmented by delta and double-delta cepstra. Each full feature vector was 39-dimensional. Cepstral mean normalization (CMN) was applied in all cases.

ITD processing for dereverberation was done exactly as shown in Fig. 1. Fig. 2 presents the improvement in the residual reconstruction error achieved with the proposed ITD technique. The ITD process was initialized with the NMF algorithm in [7] and run for 5 iterations. Fig. 2 shows that applying ITD updates on NMF-initialized estimates for the filter spectra reduces the residual reconstruction error in (1) by 78% in 5 iterations.

In Fig. 3 we plot word error rates (WER%) results across a number of competitive feature extraction systems and dereverberation algorithms. This figure shows results for clean-condition training, where the ASR system is trained with clean speech and tested on dereverberated speech. Results show that reverberation is difficult to compensate for, even using advanced baseline systems, in a manner that is conducive to speech recognition. Even the Advanced Front-End [9] system does not lead to improvements in performance for ASR on reverberated speech. Long-term log-spectral subtraction (LTLSS) [10] algorithm is a direct extension of CMN processing, where CMN processing is applied to longer analysis windows (1-2 s), and speech is reconstructed in an analysis-by-synthesis framework. LTLSS provides 15% reduction in WER over MFCs. The ITD algorithm provides 54% relative reduction WER at RT-300 ms, which is substantially better than any of the baseline algorithms. We also note that the ITD approach builds on the NMF approach, and improves it by 16% relative.

5. CONCLUSION

The least-squares decomposition technique proposed in this paper works well for ASR. Deconvolution is generally difficult to solve, especially if none of the constituent convolutive components are given, and we only have the observed convolved signal. In such situations, realistic constraints that relate to the physics of the signal of interest must be provided. Our use of NMF for initializing the room-response filter spectra is motivated by this. Also, the use of this algorithm in the Gammatone filtered magnitude spectral domain is driven by our current understanding of speech and ASR. If the constraints are reasonable, the ITD approach can apply to *any* generic deconvolution problem. In our ASR experiments it resulted in 40-54% relative reduction in WER over other baseline processing in RT ranges of 300-500ms. This is a significant improvement in performance.

6. REFERENCES

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

[2] A. Sehr and W. Kellermann, "A new concept for feature-domain dereverberation for robust distant-talking asr," *Proc. IEEE ICASSP*, pp. IV–369–IV–372, 2007.

[3] A. Krueger and R. Haeb-Umbach, "Model based feature enhancement for automatic speech recognition in reverberant environments," *Proc. InterSpeech*, pp. 1231–1234, 2009.

[4] Bell A.J. Bell and T.J. Sejnowski, "Learning the higher-order structure of a natural sound," *Network: Computation in Neural Systems*, 1996.

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

[6] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," *Proc. ICA*, pp. 494–499, 2004.

[7] K. Kumar, "A spectro-temporal framework for compensation of reverberation for speech recognition," 2010, Ph.D. Proposal, Dept. of ECE, Carnegie Mellon University, http://www.ece.cmu.edu/ kshitizk/Thesis/Proposal.pdf.

[8] N. Levinson, "The wiener rms error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, pp. 261–278, 1947.

[9] ETSI: Advanced Front-end, ETSI Doc. No. ES 202 050.

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