A CONVOLUTIVE SPECTRAL DECOMPOSITION APPROACH TO THE SEPARATION OF FEEDBACK FROM TARGET SPEECH

Gautham J. Mysore

Advanced Technology Labs
Adobe Systems Inc.

Paris Smaragdis

University of Illinois at Urbana-Champaign
Adobe Systems Inc.

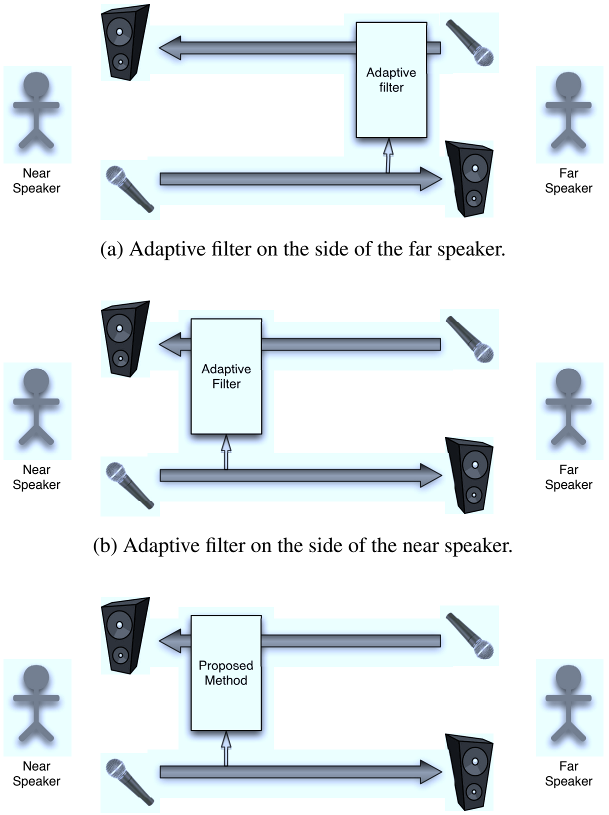
ABSTRACT

Feedback is a common problem in teleconferencing systems. Typical usage of an adaptive filter can be effective for feedback reduction but it relies on the presence of such a filter on the side of the far speaker in order to reduce feedback on the side of the near speaker. In order to avoid this reliance on the far speaker's setup, we can use an adaptive filter on the side of the near speaker. Unfortunately, due to non-linear speech coding typically used during speech transmission, these filters perform poorly in this situation. In this paper, we present a novel probabilistic method, using a non-negative convolutive decomposition of spectrogram data to perform feedback reduction by posing the problem as a source separation problem. Our method is robust to non-linear speech coding as well as continuous double-talk, which often presents a challenge to adaptive filters. We compare our method to the use of an adaptive filter and show superior results with respect to standard source separation metrics.

***Index Terms*—** Non-Negative Spectrogram Factorization, Source Separation, Feedback Reduction

1. INTRODUCTION

The problem of feedback reduction is essentially that of echo cancellation. This has been an ongoing research topic for more than half a century and has resulted in a variety of robust and specialized algorithms which we now use daily (for an overview see [1]). The objective of these techniques is to eliminate a known source after it has undergone a transformation and is then mixed with another new source. This is a situation that often takes place in a teleconferencing scenario when a transmitted signal is accidentally re-recorded in the far side (usually due to speaker-microphone coupling, but also due to network complexities) and is then sent back to the sender thus resulting in audible feedback. Most of the approaches to this problem are centered around the idea of a canceler that assumes that the sent signal will undergo a linear transform before being re-recorded. This linear transform accounts for propagation delay, the speaker/microphone frequency response, and the room in which the recording takes place. It is widespread practice to use an adaptive filter to model these effects and then subtract its output from the observed mixture at the far side to suppress the echo from the return signal (Fig. 1a). Unfortunately, such adaptive filters are not always present at the far side.



(a) Adaptive filter on the side of the far speaker.

(b) Adaptive filter on the side of the near speaker.

(c) Proposed method on the side of the near speaker.

Fig. 1: Illustration of different scenarios for feedback reduction.

In order to avoid this reliance on an adaptive filter on the far side, it could be useful to develop a solution for the near side. The user of such a system can be assured of feedback reduction regardless of the system used by the other speaker. An obvious choice would be to use an adaptive filter on the near side (Fig. 1b). However, there are a few issues with this scenario. Firstly, the filter cannot adapt to good results when both speakers are simultaneously speaking (double-talk). Therefore, double-talk detection is commonly used in this scenario and the filter stops adapting in periods in which both speakers are simultaneously speaking. This is however reliant on a robust double-talk detector. The other issue is a more fundamental problem with adaptive filters in this scenario. These filters model the transmission channel as a linear filter. However, in practical scenarios, some form of non-linear speech coding is generally used in speech transmission. Adaptive filter therefore face difficulties in this situation.

In this paper, we present an alternative model for feedback reduction (Fig. 1c) that circumvents the above problems. We use a source separation approach and thus can operate on constantly overlapping sources without requiring double-talk detection. Moreover, our approach does not produce artifacts that are commonly associated with adaptive filters in the presence of non-linearities. The model we present is operating purely in the magnitude spectral domain and is related to the non-negative models in [2, 3, 4, 5], as well as the probabilistic latent variable model in [6]. In the remainder of this paper, we present the feedback signal path, present our model for mixing, and then demonstrate how it compares against a standard frequency domain NLMS filter [1] for the problem of feedback reduction. We further show its tolerance to permanent double-talk and non-linear processing (speech coding) which are cases where adaptive filters encounter difficulties.

2. FEEDBACK SIGNAL PATH

The signal path of feedback in a typical teleconferencing scenario is follows (Fig. 1c):

1. Speech from the near speaker as well as reverberation from the room of the near speaker are fed into the microphone.

2. Speech coding takes place and the coded signal is transmitted to the far speaker.

3. The speech is decoded at the far speaker's side and is played by a loudspeaker.

4. This decoded speech as well as reverberation from the room of the far speaker are fed into the microphone. Speech from the far speaker (with reverberation) is also fed into the microphone.

5. Speech coding takes place and the coded signal is transmitted to the near speaker.

6. The speech is decoded at the near speaker's side and is played by a loudspeaker. This is the feedback that we wish to suppress. This is of course mixed with the speech of the far speaker, which we wish to retain.

If the above feedback is not suppressed, it is likely be an annoyance to the near speaker. Furthermore, it will be fed back into the microphone and go through the above steps indefinitely.

3. PROPOSED MODEL

3.1. Model definition

We develop a model of the magnitude spectrogram of a sound mixture. Since we only model the magnitude part of the spectrogram, we treat the data as *count data [7]* and represent it as a distribution of acoustic mass along the time–frequency axes. In this representation, a spectrogram is denoted as a distribution $P(f,t)$ over frequency f and time t. We will assume that the observed recording will be comprised of two elements, one being the unwanted feedback and the other being the target speech to extract. We will use a source separation strategy and assume that the observed spectrogram $P(f,t)$ is a superposition of the feedback and target spectrograms $P(f,t|feedback)$ and $P(f,t|target)$. For the sake of generality we will treat this problem as having an arbitrary amount of sources, which we denote by the latent variable z. This results in the mixture model:

$$P(f,t) = \sum_z P(z)P(f,t|z)$$

To model the convolutive effects, which we observe in a feedback scenario, we will further refine this model by independently modeling its sub-bands. We will model each sub-band signal $P(t|f)$ as a weighted summation of individual sub-band signals $P(t|f,z)$ emanating from each source. The mixture weights in the summation are given by $P(z|f)$ and the relative magnitudes of the sub-band signals are given by $P(f)$. The model of the spectrogram is therefore given by:

$$P(f,t) = P(f)\sum_z P(z|f)P(t|f,z)$$

In order to deal with convolutive effects on individual sources such as reverberation, we further model each individual sub-band signal as a convolution. Each sub-band signal, $P(t|f,z)$ of source z is modeled as a convolution of a magnitude sub-band source signal $P_S(\tau|f,z)$ and an imposed channel filter $P_F(t'|f,z)$ operating on that sub-band, where $t' = t - \tau$. Therefore, each sub-band signal is given by:

$$P(t|f,z) = \sum_\tau P_S(\tau|f,z)P_F(t-\tau|f,z)$$

Note that the convolutions in this model are being modeled in the magnitude spectrum domain and are thus capable of modeling only echoes and coloration effects. Although at first this model might seem too coarse and heavily constrained, this design choice allows it to be very tolerant to dynamic filter changes while still being able to model mixtures well enough

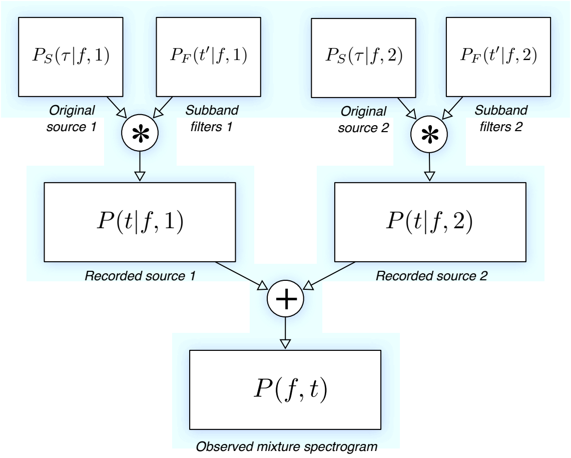


Fig. 2: Construction of the mixture spectrogram distribution. The input spectrogram is assumed to be composed by the sum of spectrograms representing each of the two sources, and each source itself is split into a collection of convolutions between all its sub-bands and a frequency specific impulse function. Note that all convolutions are operating on the horizontal dimension, and that all quantities are magnitude spectrograms.

to extract the desired sources. For the sake of symmetry and notation, we apply filters on both sources even though in a feedback reduction operation we only have to assume that one source (near source) undergoes this transformation. Combining all of the above, the complete model of the spectrogram (Fig.2) is given by:

$$P(f,t) = P(f)\sum_z P(z|f)\sum_\tau P_S(\tau|f,z)P_F(t-\tau|f,z) \quad (1)$$

All of the distributions in the above model are multinomial. It is therefore a convolutive multinomial mixture model. This is a modification of the model introduced in [6], where convolutions now only appear in the left-right dimension and each sub-band is independently scaled. This particular model can also be seen as a probabilistic multi-channel generalization of the non-negative dereverberation model described in [5], which in turn combines multiple sub-band convolutive NMF estimators which have been used in the past to model time-invariance in sound mixtures [2, 3].

3.2. Parameter Estimation

Given a convolutive sound mixture, the first step is to obtain the magnitude spectrogram, V_{ft}. This gives us the number of counts in each time–frequency bin f, t . We use V_{ft} to estimate the parameters of all of the distributions in the right hand side of Eq. 1. The parameters that are of the most interest are the parameters of the individual source distributions, $P_S(\tau|f,z)$ as they correspond to the clean individual sources. There are two latent variables in this model, z and τ. z represents the individual sound source. τ represents an instant of time in the source distribution. t' represents an instant of time in the filter distribution. Given an instant of time t of the spectrogram, the other two time variables are related as $t' = t - \tau$. Therefore, given a specific t, the second latent variable can be either τ or t'. Since this is a latent variable model, we use the Expectation–Maximization (EM) algorithm for the estimation of the model parameters. The E-step, in terms of τ, is given by:

$$P(\tau,z|f,t) \quad = \quad \frac{P(z|f)P_S(\tau|f,z)P_F(t-\tau|f,z)}{\sum_z P(z|f)\sum_\tau P_S(\tau|f,z)P_F(t-\tau|f,z)}$$

The E-step, in terms of t', is given by:

$$P(t',z|f,t) \quad = \quad \frac{P(z|f)P_S(t-t'|f,z)P_F(t'|f,z)}{\sum_z P(z|f)\sum_{t'} P_S(t-t'|f,z)P_F(t'|f,z)}$$

The M-step is given by

$$\begin{aligned}
P_S(\tau|f,z) &= \frac{\sum_t V_{ft}P(\tau,z|f,t)}{\sum_\tau \sum_t V_{ft}P(\tau,z|f,t)} \\
P_F(t'|f,z) &= \frac{\sum_t V_{ft}P(t',z|f,t)}{\sum_{t'} \sum_t V_{ft}P(t',z|f,t)} \\
P(z|f) &= \frac{\sum_t V_{ft}\sum_\tau P(\tau,z|f,t)}{\sum_z \sum_t V_{ft}\sum_\tau P(\tau,z|f,t)} \\
P(f) &= \frac{\sum_t V_{ft}}{\sum_f \sum_t V_{ft}}
\end{aligned}$$

Iterating over the above steps, we usually obtain a satisfactory solution after about 30 to 50 iterations. These steps themselves can be efficiently realized by performing them in the Fourier domain which dramatically accelerates all of the implied convolution and cross-correlations in the E- and M-steps.

3.3. Application to Feedback Reduction

We use a two source case of the above model for feedback reduction. The first source is the interfering source (near speaker) and we wish to suppress it. The magnitude spectrogram of this source will be the first source distribution $P_S(\tau|f,1)$. The channel characteristics (transmission delay, channel filter, and reverberation filter) of the source is modeled as the first filter distribution $P_F(t'|f,1)$. The second source is the target source (far speaker) that we wish to obtain. The magnitude spectrogram of this source is modeled as the second source distribution $P_S(\tau|f,2)$. Similarly the second filter distribution $P_F(t'|f,2)$ is used to model a filter for that source, which of course can be arbitrary depending on the assumed form of $P_S(\tau|f,2)$. If we estimate all of these quantities from an observed mixture, we can reconstruct the

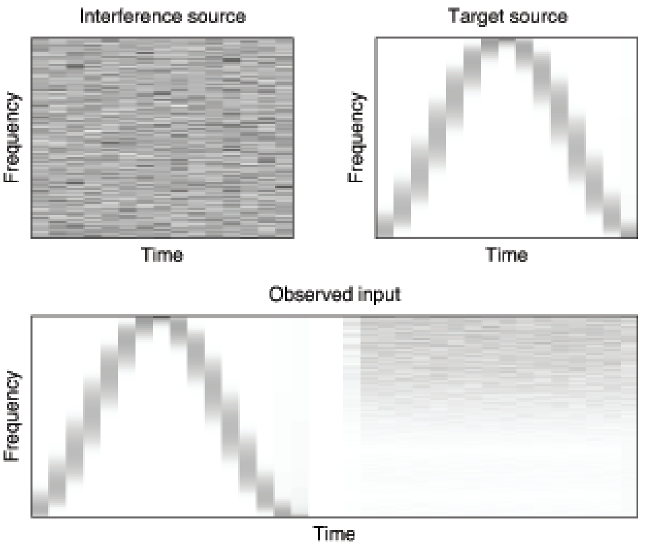


Fig. 3: An artificial example. The two sources are shown at the top and the observed input which consists of delaying and filtering the interference source is shown in the bottom.

magnitude spectrogram of the target source as $P(t|f,2)$ multiplied by the appropriate weights. This is given by:

$$P(f)P(z|f)\sum_{\tau} P_S(\tau|f,z)P_F(t-\tau|f,z)$$

We invert this reconstruction back to the time domain using the original phase of the mixture.

When performing feedback reduction, a clean recording of the interference source before it undergoes filtering is already available to us (Fig. 1c). The feedback that we are trying to suppress is a delayed and filtered version of this source. So we can assume that we know $P_S(\tau|f,1)$ and set it to the magnitude spectrogram of the original interference source. Likewise, the filters that are imposed on the target source are irrelevant since they do not pose a distortion we wish to (or can) remove. This means that we can fix $P_F(t'|f,2)$ to be a collection of constant delays with unit gain for all frequencies. This simplifies our model significantly and makes estimation of the remaining components easier and unambiguous.

Note that our model is linear. However, we stated that it is tolerant to non-linear speech coding. The reason is that our model is linear in the magnitude spectrum domain rather than the time domain as we are not modeling phase (giving us only approximate additivity of the sources). Although typical speech coding is non-linear, it aims to yield speech that is perceptually similar to the original speech. This corresponds to yielding results that are similar in the magnitude spectrum domain. Our model is therefore able to tolerate non-linear speech coding quite well. The NLMS filter on the other hand is a linear model in the time domain. It is therefore quite sensitive to non-linear speech coding.

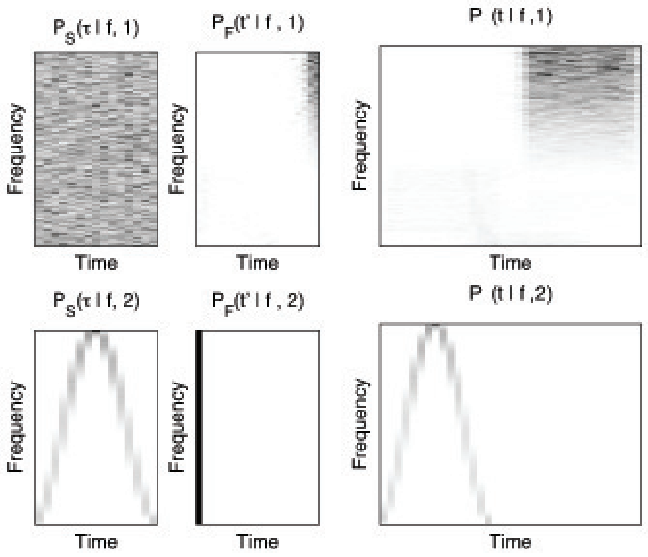


Fig. 4: The results of analyzing the example in Fig. 3. All parameters have been estimated and have converged to good approximations of the original inputs.

Now let us examine a simple case to illustrate how this algorithm works. Fig. 3 shows two sources and the observed recording. The interference source is white noise and the target source is a modulated sinusoid. The echo of the interference is delayed significantly and in addition to that it has undergone high-pass filtering which has changed its spectral characteristics. Fig. 4 shows what we extract by analyzing the observation and knowing how the interference source looks before the delay and filtering. We can see that the discovered source $P_S(\tau|f,2)$ corresponds to the target input and that the filters of the interference source form an appropriate delay with a high-pass filter in $P_F(t'|f,1)$. The resulting reconstruction of the target $P(t|f,2)$ has successfully attenuated the interference signal and produced the output we desired.

4. EXPERIMENTS

In this section, we present experiments that demonstrate the performance of the proposed method and compare it to that of the NLMS filter. We simulate each of the components of the feedback signal path in the same sequence as in Sec. 2 . We simulate room reverberation on both ends be convolving with synthetic room impulse responses (RIRs). The RIRs are exponentially decaying heavy-tailed noise. We simulate speech coding using a G.723.1 encoder and decoder.

The output of the loudspeaker on the near side is the unwanted feedback from the near speaker (interference source) as well as the speech from the far speaker (target source). The goal is to separate these two sources. Therefore we evaluate the quality of the separation using standard source separation metrics [8]. This includes three metrics:

1. Source to Interference Ratio (SIR) – This is a measure of how well we are able to suppress the unwanted source.

2. Source to Artifact Ratio (SAR) – This is a measure of the artifacts introduced by the separation process.

3. Source to Distortion Ratio (SDR) – This is an overall measure of separation performance that takes the above two criteria into account.

In a given run of the experiment, speech files are randomly chosen from the TIMIT database to represent the interference source and the target source. The interference source and target source are then mixed with some amount of overlap between the speakers (to simulate double-talk). Given this data, we evaluate the performance of the proposed method as well as that of the NLMS echo canceler. We use prior information from the construction of the mixtures to provide double-talk detection for the NLMS canceler, which results in optimal double-talk detection. For a given amount of overlap, we run the experiment ten times and report the mean results (Table 1). Since the input mixture already has an inherent SIR and SDR, we report the SIR gain and the SDR gain with respect to the target source.

As shown in Table 1, the proposed method is able to obtain superior results to the NLMS filter with respect to all metrics and in all amounts of overlap. As shown, the performance of the proposed method gradually decreases with increasing amounts of overlap. On the other hand, when using the NLMS filter, the SIR gain and SDR gain have a sudden drop in performance with high amounts of overlap. The reason for this is that there is continuous double–talk for a significant portion of the mixture. The NLMS filter does not adapt at these times. It is therefore not able to effectively suppress the interference source. The SAR actually increases with large amounts of overlap. This is likely to be due to the fact that the output is quite similar to the input in these instances, thereby not introducing much additional artifacts. Of course, in the process, it is not performing feedback reduction as reflected by the SIR.

The proposed method is an offline system. However, the parameter estimation equations can be derived to be updated recursively in an online manner, which allows a real-time deployment of the algorithm. We plan to address the recursive updates in a future publication. In order to be fair in our comparison to the NLMS filter, we use an offline version of the NLMS filter. Particularly, we use thirty passes over the data rather than a single pass (as in the case of a real-time system). Multiple passes over the data allows the filter to better adapt to the data. We found that the performance of this offline NLMS filter is significantly better than the real-time single pass version.

Overlap %	0	25	50	75	100
NLMS	18.89	18.58	15.12	3.83	0.15
Proposed Method	**24.47**	**22.39**	**19.82**	**17.87**	**14.83**

(a) Source to Interference Ratio (SIR) Gain

Overlap %	0	25	50	75	100
NLMS	6.23	5.65	4.12	0.49	-1.11
Proposed Method	**19.08**	**17.19**	**13.58**	**11.89**	**9.48**

(b) Source to Distortion Ratio (SDR) Gain

Overlap %	0	25	50	75	100
NLMS	8.09	7.74	6.09	8.17	10.67
Proposed Method	**22.19**	**20.59**	**16.37**	**14.78**	**13.48**

(c) Source to Artifacts Ratio (SAR)

Table 1: Experimental results

5. CONCLUSIONS

In this paper, we presented a novel formulation of feedback reduction based on source separation ideas. By formulating the feedback reduction problem as a constrained convolutive mixture model, we were able to bypass some of the problems that are inherent in a more traditional approach. We showed that the resulting approach is tolerant to non-linear speech coding) We additionally demonstrated that it is not dependent on double-talk detection and can operate even during constant overlap between sources. These features make it a good candidate even for challenging problems such as feedback reduction for networked collaborative music making, which is filled with double-talk of musical instruments.

6. REFERENCES

[1] M.M. Sondhi, "Adaptive echo cancellation for voice signals," *Handbook of Speech Processing*, 2008.

[2] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 1–12, January 2007.

[3] T. Virtanen, "Separation of sound sources by convolutive sparse coding," in *Proceedings of the Workshop on Statistical and Perceptual Audition*, 2004.

[4] N. Cahill and R. Lawlor, "A novel approach to acoustic echo cancellation," in *Proceedings of the European Signal Processing Conference*, 2008.

[5] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[6] P. Smaragdis, B. Raj, and M.V. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[7] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," in *Advances in Neural Information Processing Systems 20*, J.C. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., pp. 1313–1320. MIT Press, Cambridge, MA, 2008.

[8] C. Févotte, R. Gribonval, and E. Vincent, "BSS EVAL toolbox user guide," Tech. Rep. 1706, IRISA, Rennes, France, http://www.irisa.fr/metiss/bss eval/, April 2005.