COMPENSATING FOR DENOISING ARTIFACTS

Rita Singh

Carnegie Mellon University, Pittsburgh, PA, USA.

rsingh@cs.cmu.edu

ABSTRACT

Noise degrades speech signals, affecting their perceptual quality, intelligibility, as well as their downstream processing, e.g. coding or recognition. One obvious solution to this is to denoise the signals, but denoising algorithms filter out an *estimate* of noise, which is often inexact. As a result, denoising can attenuate spectral components of speech, which may enhance perceptual quality but further reduce it's intelligibility. We address the latter issue and propose a method to restore lost spectral components in denoised speech. Our algorithm modifies the standard NMF formulation to represent clean speech as a composition of bases, and denoised speech as a composition of *distortions* of these bases. By decomposing the denoised signal into a composition of the distorted bases, the corresponding clean signal can be estimated as an identical composition of the clean bases.

Index Terms: Denoising, Restoration, Intelligibility, Non-negative matrix factorization, Spectral decomposition

1. INTRODUCTION

Speech intended for transmission or recognition is typically recorded in realistic, noisy environments. In addition to reducing both the perceptual quality and intelligibility of the signal, noise negatively affects the performance of the downstream processing mechanisms (such as codecs), which are optimized for efficient performance on *clean* speech. For this reason, it is often necessary to denoise speech before further processing.

A large number of denoising algorithms have been proposed in the literature. These typically estimate the noise first and then eliminate it, either by direct subtraction [1, 2], or filtering [3, 4]. The problem is that noise estimates are usually inexact, especially when the noise is time-varying. As a result, when the estimated noise is removed from the signal, not only is residual noise left behind, but information-carrying spectral components are also attenuated. An example is shown in Figure 1, where a state-of-art denoising algorithm (from a commercial vendor) has been used to denoise a speech signal corrupted by automobile noise. We see that the high-frequency components of fricated sounds such as /S/ and very-low frequency components of nasals and liquids, such as /M/, /N/ and /L/ have been damaged. This happens because automotive noise is dominated by high and low frequencies, and cancelling the noise attenuates these frequency components in the signal.

Thus, although noise reduction results in a signal with improved perceptual quality, the intelligibility of the signal often does not improve, *i.e.*, while the denoised signal sounds cleaner, the ability of listeners to make out what was spoken is not enhanced. In fact, particularly when the denoising is aggressive or when the noise estimate is affected by nonstationarity, the denoised signal is *less* intelligible than the original noisy signal.

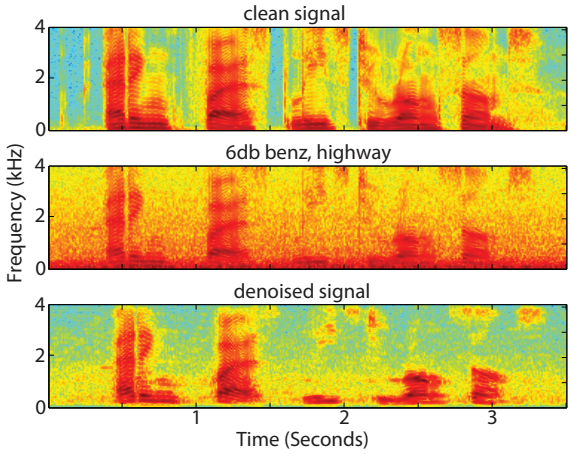

Fig. 1. Top: Spectrogram of a clean signal. Middle: The same signal corrupted by additive automotive noise to 6dB. Bottom: Denoised version of signal in the middle panel. Speech spectral components that were masked by the noise have been attenuated or deleted.

Although this is an artificially created problem resulting from imperfect processing, it is nevertheless a real one faced by producers of after-market spoken-interface devices that incorporate third-party denoising hardware or software. The algorithms are often black-boxes that are integrated into the sound capture mechanism itself, and we only have the output of the the denoiser. It then becomes important to somehow restore the speech information that the denoising algorithm has excised.

In this paper we propose an algorithm for restoring lost spectral components of denoised signals with the intention of enhancing its intelligibility. Our solution is constrained by the practical aspects of the problem: we assume that a) the denoising algorithm is a black-box and the manner in which it estimates noise, and the actual cancellation algorithm are unknown and b) it is impractical to record the noise itself separately, and no external estimate of the noise is available to guess how the denoising has affected any segment of the speech. Additionally, any processing we do must restore lost spectral components of the speech signal without reintroducing noise into the signal.

Our solution utilizes a *compositional* characterization of the signal that assumes that the signal can be represented as a constructive composition of additive units. We obtain this characterization via non-negative matrix factorization [5], although other techniques *e.g.* [6] may also be employed. We assume that the manner in which each of the additive constituents of the signal is affected by the denoising is relatively constant, and can be learned from training data compris-

ing stereo pairs of undistorted and distorted (by denoising) signals. By determining how the denoised signal is represented in terms of these additive constituents, the attenuated spectral structures can be estimated from the undistorted versions of these constituents.

Note that the proposed solution has several analogues in the literature. NMF has frequently been used for separation of mixed signals [7], or even denoising speech [8]. Closer to our solution, in [9, 10] compositional models have been used to extend the bandwidth of bandlimited signals. However, no current literature exists that addresses the specific problem mentioned in this paper, to the best of our knowledge.

This paper is organized as follows. Sections 2 3 and 4 present our basic model, feature representation used, and the proposed restoration algorithm respectively. Sections 5 and 6 present our experimental results and conclusions.

2. MODEL OF THE PROBLEM

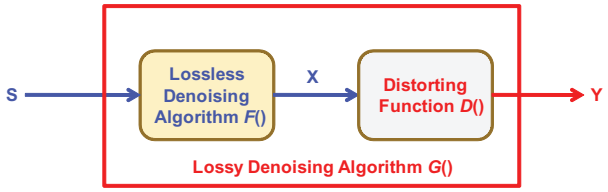

Fig. 2. Model for lossy denoising algorithm. Noisy speech S is processed by an ideal "lossless" denoising algorithm $F()$ to produce a "lossless" denoised signal X, which passes through a distortion function $D()$ to produce the"lossy" signal Y.

We model the *lossy* denoising algorithm that inappropriately attenuates spectral components of speech as a combination of a *lossless* denoising mechanism that attenuates the noise in the signal without attenuating any speech spectral components, and a *distortion* that modifies the losslessly denoised signal to produce the lossy signal. This model in illustrated by Figure 2. Noisy speech S is processed by an ideal "lossless" denoising algorithm $F(S)$ to produce a (hypothetical) lossless denoised signal X. X is passed through a distortion function $D()$ that attenuates speech spectral components to produce the lossy signal Y. Our goal is to estimate X, given only Y.

We assume that the lossless signal X can be expressed as a *composition* of unit elements, which we will call "bases", *i.e.* X can be expressed as:

$$X = \sum_{i=1}^{K} w_i B_i \tag{1}$$

where B_i are the bases and w_i are the weights with which they combine to compose X. The bases B_i are assumed to represent uncorrelated building blocks that constitute the individual spectral structures that compose X.

The distortion function $D()$ distorts the bases to modify the spectral structure they represent. Thus any basis B_i is transformed by the distortion to $B_i^{distorted} = D(B_i)$. We assume that the distortion transforms any basis independently of other bases, *i.e.* $D(B_i|B_j : j \neq i) = D(B_i)$, where $D(B_i|B_j : j \neq i)$ represents the distortion of B_i given that other bases $B_j, j \neq i$ are also concurrently present. Note that this assumption is not truly valid unless the bases represent non-overlapping, complete spectral structures. We nevertheless make this assumption to simplify the algorithm. We also assume that the *manner* in which the bases combine to compose the signal is not modified by the distortion. Again, this is a simplifying assumption that is not necessarily valid.

The implication of the above assumptions is that

$$\begin{aligned} &Y = D(X) \Rightarrow \\ &X = \sum_i w_i B_i \quad \Leftrightarrow \quad Y = \sum_i w_i B_i^{distorted} \end{aligned} \tag{2}$$

Equation 2 implies that if all bases B_i and their distorted versions $B_i^{distorted}$ are known, and if the manner in which the distorted bases compose Y could be determined (*i.e.* if the weights w_i could be estimated) then X could be estimated.

3. REPRESENTING THE SIGNAL

The model of Sec. 2 is primarily a *spectral* model. It characterizes signals as a composition of uncorrelated signals, which naturally leads to spectral characterization of all signals, since the power spectra of uncorrelated signals add. We therefore represent all signals as magnitude spectrograms[1] that are obtained by computing short-time Fourier transforms (STFT) of the signals. We found the optimal analysis window for the STFT to be 40-64ms, and used 64ms windows in our experiments.

S, X, and Y thus represent magnitude spectrograms of the noisy speech, losslessly denoised speech and lossy denoised speech respectively. The bases B_i, as well as their distorted versions $B_i^{distorted}$ represent magnitude spectral vectors. The magnitude spectrum of the t^{th} analysis frame of X, which we represent as $X(t)$ is assumed to be composed from the lossless bases B_i as $X(t) = \sum_i w_i(t) B_i$ and the magnitude spectrum of the corresponding frame of the lossy signal Y is given by $Y(t) = \sum_i w_i(t) B_i^{distorted}$. Also, the *weights* w_i *are now all non-negative*, since the signs of the weights in the model of Equation 1 simply get incorporated into the *phase* of the spectra for the bases, and do not appear in the relationship between magnitude spectra of the signals and the bases.

Our restoration algorithm estimates the lossless magnitude spectrogram X from that of the lossy signal, Y. The estimated spectrogram is inverted to a time-domain signal using the phase borrowed from the complex spectrogram of the lossy signal.

4. THE RESTORATION ALGORITHM

For restoration, in an initial training phase, the lossless bases B_i for X and the corresponding lossy bases $B_i^{distorted}$ for Y are learned from training data. These bases are then employed to estimate X. We give the details of the procedure below.

4.1. Learning the Bases

Since $D()$ is unknown, we jointly learn B_i and $B_i^{distorted}$ from analysis of joint recordings of X and the corresponding Y. For this, we need joint recordings of X and Y in the training phase. Since X is not directly available, we use the following approximation instead: we artificially corrupt clean speech signals C with digitally added noise to obtain noisy signal S. We then process S with the denoising algorithm to obtain the corresponding Y. The "losslessly denoised" signal X is a hypothetical entity that cannot be known. Instead we use the original *clean* signal C as a proxy for X.

[1] Although in theory it is the *power* spectra that add, we have empirically found additivity to hold better for magnitude spectra.

The denoising algorithm (and the distortion) will usually introduce a delay into the signal – the signals for Y and C may be shifted with respect to one another. Since the model of Equation 2 assumes a one-to-one correspondence between each segment of X and the correponding segment of Y, we must first align the recorded samples of C and Y to eliminate any relative shifts introduced by the denoising. We estimate this shift by cross-correlating each C and the corresponding Y.

The bases B_i are assumed to be the *composing* units for X. It has been shown that such bases can be obtained by analysis of magnitude spectra of signals using non-negative factorization methods, *e.g.* [6]. However, we have an additional constraint – the distorted bases $B_i^{distorted}$ must be reliably known to actually be distortions of their clean counterparts B_i.

We therefore use an *example based* model [11] where such correspondence is assured. We randomly select a large number of magnitude spectral vectors from C as the bases B_i for X. We select the corresponding vectors from the training instances of Y as $B_i^{distorted}$. The procedure ensures that $B_i^{distorted}$ is indeed a near-exact distorted version of B_i. Since bases represent spectral structures in speech, and the potential number of spectral structures in speech is virtually unlimited, we select a large number of bases. The model of Equation 1 thus becomes *overcomplete*, combining many more elements than the dimensionality of the signal itself.

4.2. Estimating weights

As a first step to restoring a test signal Y, we determine how each spectral vector $Y(t)$ of Y is composed by the distorted bases. We have $Y(t) = \sum_i w_i(t) B_i^{distorted}$. If we represent the set of all bases as a matrix: $\bar{\mathbf{B}} = [\{B_i^{distorted}\}]$, and the set of weights $\{w_i(t)\}$ as a vector: $W(t) = [w_1(t) w_2(t) \ldots]^\top$, we can write:

$$Y(t) = \bar{\mathbf{B}} W(t) \tag{3}$$

$W(t)$ is constrained to be non-negative during estimation. While a variety of update rules have been proposed to learn the weights [5], for speech and audio signals we have found it most effective to use the update rule that minimizes the generalized Kullback-Leibler distance between $Y(t)$ and $\bar{\mathbf{B}} W(t)$ [5]:

$$W(t) \leftarrow W(t) \otimes \frac{\bar{\mathbf{B}}^\top \frac{Y(t)}{\bar{\mathbf{B}} W(t)}}{\bar{\mathbf{B}}^\top \mathbf{1}} \tag{4}$$

Here $\otimes$ represents component-wise multiplication. All divisions are also component-wise. Since the representation is overcomplete (*i.e.* there are more bases than there are dimensions in $Y(t)$), the equation is underdetermined and multiple solutions for $W(t)$ exist that explain $Y(t)$ equally well. Generally in this situation the constraint that $\{w_i(t)\}$ must be *sparse*, *i.e.* have minimal non-zero (or significant) entries, is enforced, usually by minimizing the L_1 norm of $W(t)$. In our problem, enforcing sparsity did not improve results, so we do not enforce it.

4.3. Estimating restored speech

Once the weights $W(t) = [w_1(t) w_2(t) \ldots]^\top$ are determined for any $Y(t)$, by Equation 2 the corresponding lossless spectrum $X(t)$ can simply be estimated as $X(t) = \sum_i w_i(t) B_i$. Note that since the estimation procedure is iterative, the exact equality of Equation 3 is never achieved. Instead $\bar{\mathbf{B}} W(t)$ is only an approximation to $Y(t)$. To account for the entire energy in Y, we therefore use the following Wiener filter formulation to estimate the spectral vectors of X:

$$X(t) = (Y(t) + \epsilon) \otimes \frac{\sum_i w_i(t) B_i}{\sum_i w_i(t) B_i^{distorted} + \epsilon} \tag{5}$$

All divisions and multiplications are component-wise. $\epsilon > 0$ ensures that attenuated spectral components at $Y(t) = 0$ can still be restored.

4.4. Expanding the Bandwidth

If the recorded and denoised signal is *bandwidth reduced* (*e.g.* if it is telephone speech), we can extend the procedure above to (re-)introduce high-frequency spectral components into the signal. This is also expected to improve the intelligibility of the signal. To expand the bandwidth we use a procedure similar to that in [9]: the training data now include *wideband* signals for C. The training recordings for C and Y are aligned, and STFT analysis for them is performed using identically long (in time) analysis windows. This ensures that in any joint recording there is a one-to-one correspondence between the spectral vectors for C and Y. Consequently, while the bases $B_i^{distorted}$ (drawn from training instances of Y) represent reduced-bandwidth signals, the corresponding B_i represent wideband signals and include high-frequency components. When signals are restored, low-frequency components are restored using Equation 5. The high-frequency components are obtained as $X(t, f) = \sum_i w_i(t) B_i(f), f \in \{high\ frequency\}$, where f is an index to specific frequency components of $X(t)$ and B_i.

This estimate only computes spectral magnitudes. In order to invert the magnitude spectrum to a time-domain signal, phase is also required. The phase for low-frequency components is borrowed directly from the reduced-bandwidth lossy denoised signal. For higher frequencies we have found it sufficient to simply replicate the phase terms from the lower frequencies.

5. EXPERIMENTAL RESULTS

Our experiments used a commercial denoising algorithm, which we treated as a blackbox. The denoising system was part of a data capture system and it's output was bandlimited to 4kHz. The system was quite effective at improving the signal-to-noise ratio (SNR) of the recording, but, as with all denoising systems, resulted in a loss of intelligibility of the signal. The problem was particularly observable when the speech signal was corrupted by automotive noise, which has relatively high energy at low frequencies; high frequency components of fricatives such as /S/, /F/ and /HH/ and low frequency components of sounds such as /N/, /M/ and /B/ were attenuated in the denoised signal.

We obtained our training data by digitally adding various types of automotive noises at various SNRs to clean speech recordings. The noise-added signals were then denoised, and then aligned by correlation to the clean signal to eliminate relative delays. Training data consisted of several pairs of clean and denoised-and-aligned counterpart signals from multiple speakers. A set of 6000 bases were randomly drawn from the training set and used for signal restoration of test data. We obtained two different sets of bases. In the first, where the goal was simple signal enhancement, the training data were bandlimited. In the second, where we also expanded the bandwidth of the signal, the training data were fullbandwidth signals.

The test data too were obtained by digitally adding noise to clean signals. In our experiments the test speakers were represented in the training set since we only had a limited amount of data available for

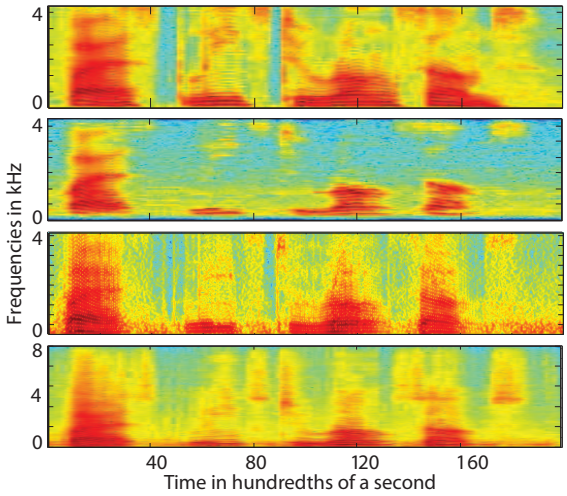

Fig. 3. Example of signal enhanced by proposed algorithm. Top panel: Actual clean signal. Second panel: Denoised narrowband signal. Third: Restored narrowband signal. Fourth: Restored *wideband* signal.

logistic reasons. However in other tests [8] we have shown that NMF need not be a speaker dependent algorithm. We therefore expect to generalize to truly speaker independent situations.

The test data were denoised by the denoising algorithm, and the signals were then processed by the proposed restoration algorithm. Figure 3 shows an example of a signal that has been enhanced by this algorithm. We note that even in the case where no bandwidth expansion is attempted, some of the spectral structures attenuated by denoising have been restored to some degree. In the case of the bandwidth-expanded signal, additional structure has been introduced into higher frequencies.

A fundamental problem with evaluating the output of our algorithm objectively is that it is intended to restore spectral components of a denoised signal. Denoising schemes typically improve the SNR of a signal. Adding new spectral components to such signals may in fact reduce the SNR. Adding spectral components, on the other hand, is expected to improve the intelligibility of the signal. While a complete MOS test to determine the intelligibility of a reconstituted signal was beyond the scope of this paper, we did compute PESQ [12] scores for the signals, which are nominally supposed to predict voice quality, under the assumption that the processing would also reflect in improvement of the quality of denoised signal, as measured by the PESQ score. While voice quality is not a perfect proxy for intelligibility, there is nevertheless a correlation between the two [13].

Firstly, we noticed that on nearly all signals the denoising algorithm improved the PESQ score by 0.5 or greater. We also observed that on a majority of the denoised signals, the additional processing did not significantly affect PESQ score, typically modifying it by ± 0.1. Additional informal listening tests showed that these signals had also not lost intelligibility significantly from the denoising and the proposed restoration algorithm did not affect the intelligibility of the denoised signals further.

However, on a small fraction of the signals (slightly over 10% of the signals), a significant improvement in PESQ score, ranging from 0.3-0.5 was obtained from the restoration. Listening tests showed that on these signals the denoising algorithm had significantly degraded the intelligibility of the signal. The proposed restoration algorithm improves the intelligibility of the signal, often significantly, on these instances. The example of Figure 3 is one of these examples. Audio examples and experimental details can be found on our website: "http://mlsp.cs.cmu.edu/projects/audio/denoisingrestoration". In no instance did the the proposed restoration algorithm actually degrade the intelligibility of the signal in informal listening tests. For bandwidth enhanced signals, it was not possible to compute PESQ scores. But listening intelligibility improved in every case, sometimes significantly. Examples can be heard on the website.

6. CONCLUSIONS

The algorithm in its current form is intended to be lightweight, in order to run on small devices. However, we believe significant improvements may be obtained by imposing temporal constraints and other priors at the cost of computation. Preliminary experiments show that imposition of such constraints results in further improvement of intelligibility, while adding some distortion.

7. REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on signal processing*, vol. 27(2), pp. 113–120, 1979.

[2] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in cars," *Speech Communication*, vol. 11(2), pp. 215–228, 1992.

[3] J. Hansen and M. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. on signal processing*, vol. 39(4), pp. 795–805, 1991.

[4] Y. Ephraim, "Statistical model based speech enhancement systems," *Proceedings of IEEE*, vol. 80(10), pp. 1526–1555, 1992.

[5] D. D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401(6755), pp. 788–791, 1999.

[6] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as non-negative factorizations," *Computational intelligence and Neuroscience*, vol. 2008, 2008.

[7] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. on audio, speech and language processing*, vol. 15(1), pp. 1–12, 2007.

[8] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Proc. INTERSPEECH*, 2010.

[9] D. Bansal, B. Raj, and P. Smaragdis, "Bandwidth expansion of narrowband speech using non-negative matrix factorization," in *Proc. Interspeech*, 2005.

[10] P. Smaragdis, B. Raj, and M. Shashanka, "Missing data imputation for time-frequency representations of audio signals," *Journal of signal processing systems*, pp. 1–10, 2010.

[11] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse nonparameteric models for single-channel separation of known sounds," in *Neural Info. Processing Systems (NIPS)*, 2009.

[12] ITU, "P.862 : Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[13] W. M. Liu, K. A. Jellyman, J. S. D. Mason, and N. W. D. Evans, "Assessment of objective quality measures for speech intelligibility estimation," in *Proc. ICASSP*, 2006.