Independent Component Analysis for Speech Enhancement with Missing TF Content

Doru-Cristian Balcan[1] and Justinian Rosca[2]

[1] Carnegie Mellon University, Computer Science Department,
5000 Forbes Ave., Pittsburgh, PA 15123, USA
dbalcan@cs.cmu.edu
[2] Siemens Corporate Research, Princeton NJ 08540, USA
justinian.rosca@scr.siemens.com

Abstract. We address the problem of Speech Enhancement in a setting where parts of the time-frequency content of the speech signal are missing. In telephony, speech is band-limited and the goal is to reconstruct a wide-band version of the observed data. Quite differently, in Blind Source Separation scenarios, information about a source can be masked by noise or other sources. These masked components are "gaps" or missing source values to be "filled in". We propose a framework for unitary treatment of these problems, which is based on a relatively simple "spectrum restoration" procedure. The main idea is to use Independent Component Analysis as an adaptive, data-driven, linear representation of the signal in the speech frame space, and then apply a vector-quantization-based matching procedure to reconstruct each frame. We analyze the performance of the reconstruction with objective quality measures such as log-spectral distortion and Itakura-Saito distance.

1 Introduction

Speech enhancement for real-world environments is still a signal processing challenge in spite of steady progress over the last forty years [1]. The work reported here addresses this problem in a setting where signal corruption consists of "hiding", zeroing out, various parts of the time-frequency (TF) content. Alternatively, we refer to this TF information as "missing" from the observed signal spectrogram. Our objective is to reconstruct the signal, or at least to infer the missing content, using the observable TF content and clean speech training data.

Two very different speech enhancement applications motivate this approach. On one side, present digital telephony operates at the minimum bandwidth requirements (e.g. $300\text{Hz} - 4k\text{Hz}$) although speech is much richer. Wide-band speech is desirable. This is the bandwidth extension problem (BWE) [2] and is typically approached using clean speech models. Linear predictive coding analysis decomposes the estimation problem into the extension of the excitation signal and the extension of the spectral envelope. The excitation can be extended e.g. with a spectral copy of the low-frequency excitation or by bandpass modulated Gaussian noise [3,4]. The spectral envelope can be extended using pattern recognition techniques relying on a Hidden Markov Model (HMM) of clean speech.

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 552–560, 2006.

On the other side, source extraction using blind source separation (BSS) techniques results in a limited fidelity reconstruction of the speech sounds in complex auditory scenarios. One reason is that many of the BSS techniques make simplifying assumptions about the underlying signal or mixing models, e.g. anechoic or single-path propagation models or sparsity and disjointness of TF representations, in order to be able to deal with more difficult realistic cases [5]. Such techniques do not presently take advantage of prior statistical speech models [6]. Statistical models of clean speech (e.g. [7]) capture the principal spectral shapes, or speech units, and their dynamics in order to be able to distinguish and recognize the speech units. This is the case of Hidden Markov Models in present Automatic Speech Recognition technology.

In our approach, the pattern of missing data could be either time invariant or variable in time, however the conceptual approach for restoring the signal does not depend on it and is rather general. We organize the description of the method as follows. The formal description of the problem in Section 2 is followed by the training-based procedure for the inference of missing data, and by the analysis of alternative choice of transforms. Section 5 presents experimental results. Section 6 highlights the main contributions and future work.

2 Formal Description and Notations

Given an observed signal $\tilde{x}$ as an altered version of a "clean" signal x_c, we want to reconstruct x_c. Let $\tilde{X}$ and respectively X_c be the Short-Time Fourier Transform (STFT) versions of $\tilde{x}$ and x_c, where the frame length is denoted by L, the overlap by δ, and the length of $\tilde{x}$ is N. We can therefore consider $\tilde{X}$ and X_c as $L \times \lfloor (N-\delta)/(L-\delta) \rfloor$ complex matrices.

To intuitively describe the missing data locations, we introduce the concept of "mask" associated with an observed signal. A mask M is simply a binary (0-1) matrix over TF domain, of the same size as $\tilde{X}$, with the obvious meaning: $M(\omega, t) = 0$, if and only if the TF point at "location" (ω, t) in the spectrogram $\tilde{X}$ is missing. In this case, we regard $\tilde{X}$ as being a "masked" version of X_c:

$$\tilde{X} = M \odot X_c \tag{1}$$

where $\odot$ is point-wise multiplication. The equation above immediately implies, for each time frame t:

$$\tilde{X}(:,t) = M(:,t) \odot X_c(:,t) = diag(M(:,t)) \cdot X_c(:,t) \tag{2}$$

Let us denote $\mathcal{F}$ and $\mathcal{F}^{-1}$ to represent the direct and the inverse Fourier Transform operators [1] respectively, and define $\mathbf{x}(t) = \mathcal{F}^{-1}\left(\tilde{X}(:,t)\right)$, and analogously $\mathbf{x}_c(t) = \mathcal{F}^{-1}(X_c(:,t))$. It follows that:

[1] Here, they are considered $L \times L$ complex matrices, normalized such that $\mathcal{F}^{-1} = \bar{\mathcal{F}}$.

$$\mathbf{x}(t) = \mathcal{F}^{-1} \cdot diag(M(:,t)) \cdot \mathcal{F} \cdot \mathbf{x}_c(t) \tag{3}$$

From here it follows that the observed speech frame is the filtered version of some clean speech frame. However, $\mathbf{x}_c$ can not be uniquely recovered just from equation 3. To reconstruct the clean frame $\mathbf{x}_c$, we will exploit the valuable information given by mask M about how the original clean data was transformed.

We can identify various particular cases of the problem, by inspecting the distribution of 1-entries in the mask. For example, if all the columns of M are identical, we denote this as the "constant missing data pattern" case, otherwise we refer to M as the description of a "variable missing data pattern". If all the entries of M corresponding to low-frequency entries in $\tilde{X}$ (up to some cut-off frequency F_{cut}) are equal to 1, we have a particular sub-case of the "constant pattern" case, namely BWE. Alternatively, we may obtain the mask by applying the BSS algorithm in [5]. Namely, when a mixture of 3 or more speech signals is given, the TF contents of these sources are assumed mutually disjoint, and a maximum-likelihood map is generated for each of them, to indicate the most informative positions of the sources in the TF plane. In this case, the masks we work with will have a variable, random-looking aspect, and the source restoration will fall under the "variable pattern" case [2].

In the following, we show how clean speech can be used to aid in the speech enhancement problem, present a unified methodology to treat these different situations, and identify the virtues and the limitations of our approach.

3 Inference of Missing Data

To restore the spectrogram of the corrupted speech signal, we need to design a procedure for the reconstruction of each data frame. As we noted in the previous section, we have good knowledge about how the clean data is processed to produce the observed data. Unfortunately, the information we have is not sufficient, since the frame filtering involved here is not an invertible process except in the trivial case when the entries in the corresponding mask column are all 1. Equation 3 has in general infinitely many solutions for $\mathbf{x}_c(t)$.

Fortunately, not all these possible choices are likely to correspond to real speech data frames. Furthermore, two (or more) consecutive observed frames should have originated in consecutive clean speech frames, and the smoothness of clean speech TF representation induces strong additional constraints on the reconstructed speech signal. Clean speech training data will help us restrict our search, by giving valuable hints of optimal (or plausible) solutions.

[2] In the constant pattern case, it is possible to give a very intuitive meaning to $\mathbf{x}(t)$, namely $\mathbf{x}(t) = \mathcal{W} \odot \left(\tilde{x}\left((t-1)(L-\delta)+1, ..., (t-1)(L-\delta)+L\right)^T\right)$ is a windowed version of the t^{th} frame extracted from $\tilde{x}$. In the variable pattern case, approximating consecutive frames in $\tilde{x}$ with filtered versions of consecutive frames in x_c is generally not accurate.

Now, we introduce several notations that are useful in describing the proposed reconstruction procedure. Let $\mathbf{X}_{train}$ denote a $L \times T$ matrix whose columns are speech frames, extracted at random positions from clean speech example sentences. We will denote by $\mathbf{x}^m_{train}(t)$ the result of filtering $\mathbf{x}_{train}(t)$, (the t^{th} column of $\mathbf{X}_{train}$) using mask vector m, as in Eqn. 3:

$$\mathbf{x}^m_{train}(t) = \mathcal{F}^{-1} \cdot diag(m) \cdot \mathcal{F} \cdot \mathbf{x}_{train}(t) \tag{4}$$

The procedure we propose to infer the missing data is relatively simple. For a certain observed frame $\mathbf{x}(t)$, and known $M(:,t)$ we filter the training frames accordingly, thus obtaining $\mathbf{x}^{M(:,t)}_{train}(t_1), t_1 = 1, ..., T$. Then, for a distance criterion d of our choice, we find the index:

$$t_* = \arg\min_{t_1} d\left(\mathbf{x}(t), \mathbf{x}^{M(:,t)}_{train}(t_1)\right) \tag{5}$$

To obtain the reconstructed frame $\hat{\mathbf{x}}(t)$, we simply add the "missing" part of $\mathbf{x}_{train}(t_*)$ to the observed frame $\mathbf{x}(t)$:

$$\hat{\mathbf{x}}(t) = \mathbf{x}(t) + \left(\mathbf{x}_{train}(t_*) - \mathbf{x}^{M(:,t)}_{train}(t_*)\right) = \mathbf{x}(t) + \mathbf{x}^{1-M(:,t)}_{train}(t_*) \tag{6}$$

The matching step (that is, finding t_* in Eqn. 5) is solvable by exhaustive search for each test frame in the training database[3]. However, matching can be efficiently defined and performed by vector quantization [8].

The quality of the reconstruction is strongly dependent on the distance criterion we choose. To address this problem, let us emphasize first that the signal representation (set of vectors in L-dimensional real or complex space) lends itself to very natural distance measures, among which the most intuitive is the Euclidean distance. It should be very clear though that this distance measure is only useful if the distribution of the data is suitably described in the "frame space" (for instance, if it is axis-aligned). Otherwise, a small distance might not necessarily mean good perceptual similarity and the best match might actually be completely erroneous from reconstruction point of view. There are two ways (not necessarily orthogonal) to avoid this obstacle.

One way is to choose several best matches (e.g. best ten) and combine them to produce a better reconstruction of the current frame. The second one refers to finding and using the best description of the data as linear combination of vectors in the space.

Independent Component Analysis (ICA) is an adaptive, linear representation method. When applied to speech frames, ICA provides a linear representation that maximizes the statistical independence of its coefficients, and therefore finds the directions with respect to which the coefficients are as sparsely distributed as possible. It is therefore a good idea to employ an ICA algorithm such as FastICA [9], which learns the ICA representation matrices from the clean training data. We denote the ICA basis function matrix by $\mathbf{A}$, and the filter matrix by $\mathbf{W}$.

[3] From now on, we refer to the observed data frames as "test data", to fit into the "testing vs. training" Machine Learning terminology.

The next step is to transform both the testing data and the filtered training data into the ICA domain (by multiplying each frame vector by matrix $\mathbf{W}$) and obtain the new vectors:

$$\mathbf{s}(t) = \mathbf{W} \cdot \mathbf{x}(t) \; ; \qquad \mathbf{s}_{train}^{M(:,t)} t) = \mathbf{W} \cdot \mathbf{x}_{train}^{M(:,t)}(t) \tag{7}$$

Next, find the best match by using the Euclidean distance in this space rather, than in the frame space (further referred to as "time domain"). Further justification and analysis of the inference in the ICA domain is given in [10].

Using ICA to represent speech has already been investigated, and its characteristics have been reported in several papers (e.g. [11], etc.). The ICA filters (rows of $\mathbf{W}$) tend to have high sensitivity to relatively narrow frequency bands, and therefore the ICA basis functions (columns of A) will tend to represent primarily frequency information within these bands. We can easily demonstrate this behavior in Figure 1.

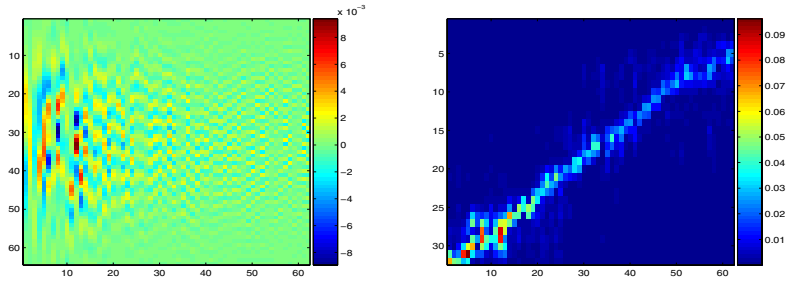

Fig. 1. Frequency selectivity of ICA basis functions: representation of both $\mathbf{A}$ and the Fourier domain representation of the columns of $\mathbf{A}$; Columns are sorted according to the peak frequency response

As a general practice, we note that we can obtain completely new distance criteria simply by mapping the test data and the filtered training data into some new space (either by a linear, or by a nonlinear mapping) and compute the Euclidean distance between the mapped vectors. In the following, we investigate the performance of computing these distances in various domains.

4 Alternative Domain Transforms

Let us examine particular cases of distance matching criteria, as well as their performance in speech signal reconstruction. Alternatively, we can regard a particular distance choice as the Euclidean distance in a certain transformed space, and we can talk about the choice of transform, instead of the choice of distance.

4.1 Euclidean Distance in Time-Domain

Let $\Lambda = diag\left(M\left(:,t\right)\right)$. The expression of this distance is:

$$\begin{aligned}
\left\|\mathbf{x}\left(t\right)-\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right\|_2^2 &= \left(\mathbf{x}\left(t\right)-\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right)^T\overline{\left(\mathbf{x}\left(t\right)-\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right)}\\
&= \left(\mathbf{x}\left(t\right)-\mathcal{F}^{-1}\cdot\Lambda\cdot\mathcal{F}\mathbf{x}_{train}\left(t_1\right)\right)^T\cdot\overline{\left(\mathbf{x}\left(t\right)-\mathcal{F}^{-1}\cdot\Lambda\cdot\mathcal{F}\mathbf{x}_{train}\left(t_1\right)\right)}\\
&= \left(\mathcal{F}\mathbf{x}\left(t\right)-\Lambda\cdot\mathcal{F}\mathbf{x}_{train}\left(t_1\right)\right)^T\cdot\left(\mathcal{F}^{-1}\right)^T\cdot\overline{\mathcal{F}^{-1}}\cdot\overline{\left(\mathcal{F}\mathbf{x}\left(t\right)-\Lambda\cdot\mathcal{F}\mathbf{x}_{train}\left(t_1\right)\right)}\\
&= \left(\Lambda\cdot\mathcal{F}\mathbf{x}_c\left(t\right)-\Lambda\cdot\mathcal{F}\mathbf{x}_{train}\left(t_1\right)\right)^T\cdot\overline{\left(\Lambda\cdot\mathcal{F}\mathbf{x}_c\left(t\right)-\Lambda\cdot\mathcal{F}\mathbf{x}_{train}\left(t_1\right)\right)}\\
&= \left(\mathcal{F}\mathbf{x}_c\left(t\right)-\mathcal{F}\mathbf{x}_{train}\left(t_1\right)\right)^T\cdot\Lambda\cdot\overline{\left(\mathcal{F}\mathbf{x}_c\left(t\right)-\mathcal{F}\mathbf{x}_{train}\left(t_1\right)\right)}\\
&= \left(\mathbf{x}_c\left(t\right)-\mathbf{x}_{train}\left(t_1\right)\right)^T\cdot\left(\mathcal{F}\Lambda\mathcal{F}^{-1}\right)\cdot\overline{\left(\mathbf{x}_c\left(t\right)-\mathbf{x}_{train}\left(t_1\right)\right)}
\end{aligned}$$

As we can see, this distance reflects somehow the "closeness" between the clean frame (from which the observed data supposedly originated) and the clean training data, but the weighting matrix $\mathcal{F}\cdot\Lambda\cdot\mathcal{F}^{-1}$ alters this natural closeness according to the missing data pattern $M\left(:,t\right)$ in a rather non-intuitive way. Speech signals do not have a too informative description in the (frame-based) time domain representation and our setting is apparently quite sensitive to this issue, especially when training data is scarce.

4.2 Euclidean Distance in Fourier Domain

Instead of $\left\|\mathbf{x}\left(t\right)-\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right\|_2$, we use $\left\|\mathcal{F}\mathbf{x}\left(t\right)-\mathcal{F}\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right\|_2$:

$$\begin{aligned}
\left\|\mathcal{F}\mathbf{x}\left(t\right)-\mathcal{F}\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right\|_2^2 &= \left(\mathcal{F}\mathbf{x}\left(t\right)-\mathcal{F}\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right)^T\cdot\overline{\left(\mathcal{F}\mathbf{x}\left(t\right)-\mathcal{F}\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right)}\\
&= \left(\mathbf{x}\left(t\right)-\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right)^T\overline{\left(\mathbf{x}\left(t\right)-\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right)} = \left\|\mathbf{x}\left(t\right)-\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right\|_2^2
\end{aligned}$$

Under our working assumptions, using Euclidean distance in time or Fourier domain will produce exactly the same results.

4.3 Euclidean Distance in ICA Domain

One problem with the previous two distance measures is that neither representation is the most appropriate one for describing the distribution of speech frames. One reason for their limited power is that they are both fixed, data independent representations. Since the distribution of speech frames is different from one speaker to another it is meaningful to consider adaptive representations of the data. We mentioned earlier ICA as a data-driven linear representation, whose objective is to maximize the statistical independence of the coefficients. Under this assumption, the basis functions will point into directions that determine maximally sparse (sharply peaked, with heavy tails) marginal distributions of the coefficients. Let us inspect what the Euclidean distance is in the ICA domain:

$$\left\|\mathbf{s}\left(t\right)-\mathbf{s}_{train}^{M(:,t)}\left(t_1\right)\right\|_2^2=\left(\mathbf{s}\left(t\right)-\mathbf{s}_{train}^{M(:,t)}\left(t_1\right)\right)^T\cdot\overline{\left(\mathbf{s}\left(t\right)-\mathbf{s}_{train}^{M(:,t)}\left(t_1\right)\right)}$$
$$=\left(\mathbf{W}\mathbf{x}\left(t\right)-\mathbf{W}\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right)^T\overline{\left(\mathbf{W}\mathbf{x}\left(t\right)-\mathbf{W}\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right)}$$
$$=\left(\mathbf{x}\left(t\right)-\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right)^T\cdot\mathbf{W}^T\mathbf{W}\cdot\overline{\left(\mathbf{x}\left(t\right)-\mathbf{x}_{train}^{M(:,t)}\left(t_1\right)\right)}$$

Since matrix $\mathbf{W}$ is in general not orthogonal, the weighted distance it induces will be different from the previous two distance measures. Furthermore, we can use the Mahalanobis distance (in either time or ICA domains) to normalize the data covariance, and thus to reduce the dependency of the distance on the relative scaling along the principal axes of the corresponding space.

5 Experimental Results

To demonstrate the performance of the proposed method, we tested it in several special settings, on data taken from the NTT database. The signals represent sentences, several seconds long, sampled at $F_s = 16k$Hz, and read by either male or female native English language speakers. We illustrate the reconstruction performance both qualitatively, by displaying the spectrogram of corrupted and reconstructed speech signals, and quantitatively, by computing objective reconstruction quality measures, such as log-spectral distortion (LSD) [2] and Itakura-Saito distance (IS) [1].

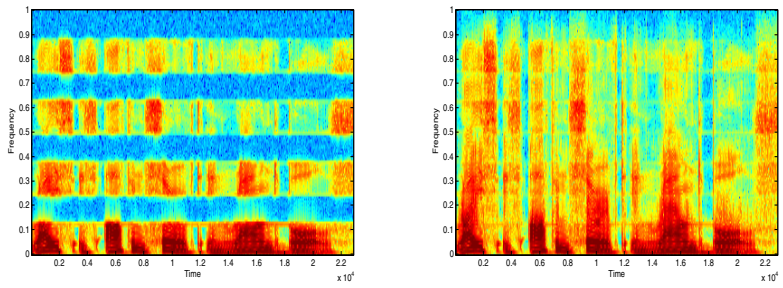

Fig. 2. Example of observed male speech data and the reconstructed signal

First, we applied the proposed method for BWE, where the observed signal was obtained by low-pass filtering a wideband speech signal. As the value for the cut-off frequency, we used $F_{cut} = 4k$Hz. We display the results obtained (see Table 1 (a)) when using Euclidean distance in time domain and in ICA domain. Although the difference is not big, the experiment shows that the performance in the latter case was consistently better. Also, by increasing the number of training speech frames, the performance is improved. Overall, results are comparable in terms of LSD with the state-of-the-art [2].

Table 1. (a) Left - Numerical values of mean values for LSD and IS, in the case of BWE ($F_{cut} = 4k$Hz). (b) Right - Mean LSD and IS values, in the case of alternative constant pattern setting. Amount of training data is either 20000 or 100000 frames of clean speech.

(a)	20,000		100,000	
Domain	LSD	IS	LSD	IS
Time	5.728	0.166	5.518	0.0756
ICA	5.574	0.157	5.462	0.0670

(b)	20,000		100,000	
	LSD	IS	LSD	IS
	5.935	0.164	5.827	0.132
	5.829	0.147	5.670	0.106

In the second experiment, we chose a fixed, multi-band pattern and again we tried to reconstruct a wideband signal (see example in Figure 2). The missing frequency bands were $1k$Hz$-2k$Hz, $3k$Hz$-4k$Hz, $5k$Hz$-6k$Hz, and $7k$Hz$-8k$Hz. The numerical values obtained for LSD and IS resulted after the bandwidth extension are shown in 1 (b). The results display the same behavior we noticed in the previous experiment.

6 Conclusions

This paper addresses the problem of speech enhancement in a setting where parts of the time-frequency content of the speech signal are missing. We analyze a simple but general algorithm for inferring the missing spectral data, and demonstrate the wide applicability of the method to bandwidth expansion, to a multi spectral band inference problem, and show applicability to enhancing speech resulting from TF BSS methods. Interesting areas of further work are coding for VQ-based training data compression and algorithm speed-up, applications to BSS, and extensions of the algorithm to exploit models of temporal relationships between successive frames.

References

1. Deng, L., O'Shaughnessy, D.: Speech Processing. M. Dekker (2003)
2. Jax, P.: Ch. 6. In: Audio Bandwidth Extension. J. Wiley and Sons (2004)
3. Qian, Y., Kabal, P.: Combining equalization and estimation for bandwidth extension of narrowband speech. ICASSP (2004) I–713–I–716
4. Unno, T., McCree, A.: A robust narrowband to wideband extension system featuring enhanced codebook mapping. In: ICASSP 2005, (DSPS R&D Center, Texas Instruments, Dallas, TX) I–805 – I–808
5. Rickard, S., Balan, R., Rosca, J.: Real-time time-frequency based blind source separation. (In: ICA 2001, San Diego, CA) 651–656
6. Attias, H., Platt, J., Acero, A., Deng, L.: Speech denoising and dereverberation using probabilistic models. In: Neural Information Processing Systems 13. (2001) 758–764
7. Roweis, S.T.: One microphone source separation. In: Neural Information Processing Systems 13 (NIPS). (2000) 793–799

8. Gray, R., Neuhoff, D.: Quantization. IEEE Trans. on Information Theory **44** (1998) 2325–2383
9. Hyvarinen, A., Karhunen, J., Oja, E.: Independent component analysis. John Wiley and Sons (2001)
10. Rosca, J., Gerkmann, T., Balcan, D.: Statistical inference with missing data in the ica domain. (In: Submitted to ICASSP 2006)
11. Lewicki, M.: Efficient coding of natural sounds. Nature Neurosci. **5** (2002) 356–363