STATISTICAL INFERENCE OF MISSING SPEECH DATA IN THE ICA DOMAIN

Justinian Rosca — Siemens Corporate Research, Princeton, NJ 08550, USA, justinian.rosca@siemens.com

Timo Gerkmann — Bochum University, Bochum, Germany, timo.gerkmann@ruhr-uni-bochum.de

Doru-Cristian Balcan — Carnegie Mellon University, Pittsburgh, PA 15213, USA, dbalcan@cs.cmu.edu

ABSTRACT

We address the problem of speech estimation as statistical estimation with "missing" data in the independent component analysis (ICA) domain. Missing components are substituted by values drawn from "similar" data in a multi-faceted ICA representation of the complete data. The paper presents the algorithm for the inference of missing data in the case of a fixed pattern of missing data. We apply our approach to the problem of bandwidth extension, or where speech is degraded by a fixed filtering process and show the capability of the algorithm to reconstruct fine missing details of the original data with little artifacts. The evaluation is done using objective distortion measures on speech samples from the NTT database.

1. INTRODUCTION

Over the last forty years, remarkable progress has been made in the area of speech separation and enhancement, however accurate estimation of clean speech for real-world environments is still a challenge (for a comprehensive review of ideas see [1]). Extraction of speech sources typically exploits the diversity from multi-microphone measurements and the statistical independence of the sources, in so called blind source separation (BSS) or independent component analysis (ICA) [2]. Some of these techniques make simplifying assumptions about the data e.g. anechoic or single-path propagation models or sparsity and disjointness of TF representations. They have been successful on real data [3], however they are limited in the fidelity of reconstruction of the speech sounds in complex auditory scenarios. One should take advantage of prior statistical speech models [4].

One problem motivating this work is spectral extension. Present digital telephony systems operate at the minimum requirements of analog speech communication (e.g. 300Hz to 4kHz bandwidth) although speech is much richer. Another problem is source separation from mono or multi-channel data: in the presence of masking sources of sound, only inaccurate source reconstruction is possible with present approaches. In both cases, can information of the source signal be modeled and used in order to recreate a natural sounding source in adverse conditions? These problems have much in common. Data (in some appropriately chosen space) may not really be missing but be masked or very noisy, which makes statistical estimation difficult and non-robust. An abstract statement of our problem is to perform statistical model-based inference of the "missing" data and rely on prior models of clean speech [5]. Also, it is useful to characterize properties of the corresponding statistical inference procedures.

The paper presents and evaluates algorithms for implicit modeling and inference to deal with a fixed pattern of missing data and applications to the spectral extension problems. This can be generalized to a random pattern of missing data. Next section presents related work motivating our approach. Section 3 introduces the algorithm for statistical estimation with a fixed data template, corresponding to the spectral extension problem. Section 4 presents experimental results on bandwidth extension problems and Section 5 concludes and highlights future work directions.

2. RELATED WORK

An auditory scene is generally a mixture of auditory objects, environment noise, music, etc. Computational auditory scene analysis (CASA) approaches [6] attempt to extract and use information or features about auditory sources in order to ultimately separate the source of interest. Model-based mono source separation approaches confine the signal space to possibilities given by an explicit or implicit model of the source. Under the model assumptions, it is possible to identify the model and therefore recover an estimate of the source [7, 8, 9].

Statistical models of clean speech (e.g. [8]) capture the principal spectral shapes, or speech units, and their dynamics in order to be able to distinguish and recognize the speech units (i.e. Hidden Markov Models in present Automatic Speech Recognition technology). Instead, we depart from the goal of recognizing speech units and investigate the possibility to model global characteristics of speech regardless of the sequence of speech units. Independent component analysis is what we need [2]. In this work we use ICA speech models to infer missing or masked components in speech data and reconstruct speech better.

Speech features are characterized by statistical dependence across space and time. However we will not use explicit features defined with fixed, or data-independent transformations. This is the case e.g. in [10], where a multitude of features and grouping cues (continuity, pitch, timbre) are used to train

Authors performed this work while visiting Siemens Corporate Research a spectral clustering procedure. Related examples are denoising sounds by sparse code shrinkage [2], as such signals result in sparse representations due to their time-frequency distributions. Alternatively, sparseness has been explicitly used in the choice of representations. Compared with some commonly used transformations such as discrete Fourier, discrete cosine, and wavelet transforms, ICA is a data-driven transformation adapted to the structure of speech training data. The literature exploits this assumption formally by considering that speech features have Laplacian priors, and by using ICA to derive data-dependent features [11].

Present bandwidth extension (BWE) literature is also relevant, as BWE fundamentally relies on clean speech models. BWE algorithms use Linear Predictive Coding (LPC) analysis to decompose the estimation problem into the extension of the excitation signal and the extension of the spectral envelope. The excitation can be extended with a spectral copy of the low-frequency excitation [12], or by bandpass modulated Gaussian noise [13, 14]. The spectral envelope can be extended using pattern recognition techniques relying on a Hidden Markov Model (HMM) of speech. Relevant information about the spectral envelope of the extension band is extracted from narrow band speech. Other approaches forego the LPC analysis and copy the narrow band speech spectrum [15].

3. SPECTRAL EXTENSION FOR FIXED PATTERN OF MISSING DATA

3.1. ICA Domain Data Transformation

We denote a time domain, clean, mono speech signal by $x(t)$, where t is the discrete time index. A set of independent basis functions, learned in unsupervised manner by an ICA algorithm, represents both the *frequency* and the *phase* spectrum of the clean speech signal.

The first step in learning the ICA representation of $x(t)$ is to take a large number L of frames of length N at random starting points in training speech data $x(t)$, transform each frame using a Hamming window, and lay them as a column of the clean speech matrix $\mathbf{X}_{train}$. An ICA algorithm, such as FastICA [2], learns a set of basis vectors (columns of matrix $\mathbf{A}$) from $\mathbf{X}_{train}$. The transformation into statistically independent components (ICs), $\mathbf{S}_{train}$, captures informative features of the natural speech in the form of a higher order structure [16] (in contrast to second order statistics):

$$\mathbf{S}_{train} = \mathbf{A}^{-1}\mathbf{X}_{train} \quad \text{or} \tag{1}$$

$$\mathbf{X}_{train} = \mathbf{A}\mathbf{S}_{train} \tag{2}$$

The rows of $\mathbf{A}^{-1}$ are filters, acting as feature detectors at various frequency combinations on the training data $x(t)$. The basis functions i.e. the columns of $\mathbf{A}$ are waveforms to which the filters respond optimally. Figure 1 illustrates the frequency localization properties of the basis functions.

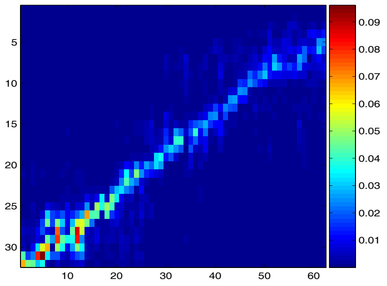

Fig. 1. Frequency localization property of the ICA basis functions $\mathbf{A}$ is shown by the Fourier domain representation of the columns of $\mathbf{A}$; Columns are sorted according to the peak frequency response per column. Training data is described in Section 4.

3.2. Bandwidth Extension Problem

The bandwidth extension problem is reduced to a problem of *statistical inference with missing data* in the ICA domain. We are given testing data $\mathbf{X}_{test}$, representing limited bandwidth speech (e.g. a cut-off frequency of 4 kHz). Enhancement targets the natural quality and intelligibility of speech (see a good review in [12]). Consider the data set:

$$\mathbf{S} = \left[\mathbf{A}^{-1}\mathbf{X}_{train}\,;\,\mathbf{A}^{-1}\mathbf{X}_{test}\right] , \tag{3}$$

where the first part is computed from wide band speech while the latter is "incomplete", missing high-frequency band information. This represents the set of data on which we apply statistical inference to recover the missing data, in this case the ICA components of the high pass bands in $\mathbf{A}^{-1}\mathbf{X}_{test}$.

Figure 2 illustrates the essence of the missing data problem. The missing values (crossed in the figure) are statistically imputed from the ICA representation of wide band data.

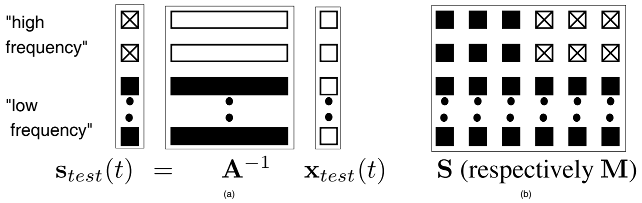

Fig. 2. (a) The missing data problem for BWE using the ICA representation: Assume that the ICA basis are given in a first approximation by Fourier basis. High frequency feature detectors (white horizontal vectors) result in zero level values for the corresponding independent component coefficients of a limited band (e.g. low pass) data frame. We consider these coefficients as "missing." (crossed in the figure) (b) $\mathbf{S}$ and the missing data pattern $\mathbf{M}$.

3.3. Statistical Inference of Missing IC Coefficients

We deal with the problem of missing data according to the treatment in [5]. Assume the data $\mathbf{S} = (s_{ij})$ in equation (3), where one column of $\mathbf{S}$ corresponds to independent component coefficients $\mathbf{s}(t)$ (where t is a generic index). $\mathbf{s}(t)$ originates either in the training data or the testing data. Consider the missing data indicator $\mathbf{M} = (m_{ij})$, where $m_{ij} = 1$ if

s_{ij} is missing and 0 otherwise. The missing data mechanism is given by the conditional density $f(\mathbf{M}|\mathbf{S})$. In our case the pattern of missing data is fixed, i.e. it does not depend on the values of the data: $f(\mathbf{M}|\mathbf{S}) = f(\mathbf{M})$. This case is denoted in [5] as *missing completely at random* . We base inferences of the missing values in $\mathbf{S}(t)$ on properties of the ICA model.

We adopt the idea of the "nearest neighbor hot-deck" imputation [5]. Missing values of $\mathbf{s}(t)$ are recovered by substitution with values drawn from a set of complete frames in the sample (i.e. corresponding to the training data). This hot-deck set is obtained by using a metric d such as Euclidean or Mahalanobis distance on frames, measuring how well complete data frames fit the incomplete data frame. Hot-deck refers to the set of matching frames to become "donors" of information, available for an incomplete instance in opinion surveys. The squared Mahalanobis distance is a weighted distance, computed by weighting different components of the data according to their covariance:

$$d^2_{Mah}(\mathbf{s}(t_0), \mathbf{s}(t)) = (\mathbf{s}(t_0) - \mathbf{s}(t))^T \mathbf{R}_{ss}^{-1} (\mathbf{s}(t_0) - \mathbf{s}(t)) \, , \quad (4)$$

where $\mathbf{R}_{ss}$ is the estimate of the covariance matrix on the set of completely available independent component (training) frames. The closest frame will contribute the missing data:

$$t_* \quad = \quad \text{argmin}_t \, d_{Mah}(\mathbf{s}(t_0), \mathbf{s}_{train}(t)) \qquad (5)$$

It can be shown that for simple hot-deck procedures (e.g. sampling with replacement) and under the missing completely at random assumption (our case) the estimators for the mean and variance of the inferred values are unbiased. This is important as inferred independent components will be transformed back to the time domain and integrated with the available data to enhance speech.

3.4. Spectral Extension Algorithm

Although wide band speech independent components could be used in the nearest-neighbor hot-deck matching process, we choose a variation of equation 3, namely:

$$\mathbf{S} \quad = \quad \left[\mathbf{A}^{-1}\mathbf{X}^{LP}_{train} \, ; \, \mathbf{A}^{-1}\mathbf{X}_{test}\right] \qquad (6)$$

where $\mathbf{X}^{LP}_{train}$ is the training data transformed with the same bandlimiting process (i.e. low pass) as the data to be extended. According to Figure 1, the reasoning is that IC contributions due to high frequency data will leak into the independent components roughly responsible for low frequency content. Thus, in the statistical matching process we make a fair comparison: Imputation is based on comparisons (using the metric of choice) between band limited test data frames and equivalent training frames. Similarly, the inferred data can be extracted from the high-pass representation of the training data at the corresponding matching index t_* (5). Let us introduce the following notations: Consider two filters $h_1(t)$ and $h_2(t)$ and define the convolutions $\otimes$ and signals below (where $x^I(t)$ and $x^{II}(t)$ have the role of x^{LP} and x^{HP}):

$$x^I(t) = h_1(t) \otimes x(t) \quad ; \quad \mathbf{S}^I = \mathbf{A}^{-1}\mathbf{X}^I \qquad (7)$$

$$x^{II}(t) = h_2(t) \otimes x(t) \quad ; \quad \mathbf{S}^{II} = \mathbf{A}^{-1}\mathbf{X}^{II} \qquad (8)$$

The final spectral extension algorithm is as follows:

```
1. Training phase:
     (a) Given clean wide band speech $x_{train}$, learn its
         ICA mixing matrix $\mathbf{A}$.
     (b) Obtain $\mathbf{S}^I_{train}$ and $\mathbf{S}^{II}_{train}$, the frequency band
         IC representations of $x^I(t)$ and $x^{II}(t)$
2. Testing phase:
     (a) Given spectral limited, possibly noisy speech
         $x_{test}$ obtain ICA domain representation $\mathbf{S}$
     (b) For each test data frame index $t_0$, find $t_*$,
         index of best match in training data (eq.  5)
     (c) Reconstruction rule.  Estimate ``extended''
         independent component values:
             $\hat{\mathbf{s}}(t) = \mathbf{s}^I(t) + \mathbf{s}^{II}_{train}(t_*)$      (9)
     (d) Spectrally extended signal is $\hat{\mathbf{x}}(t) = \mathbf{A}\hat{\mathbf{s}}(t)$
```

Let us prove that the reconstruction rule for extended independent components is correct.

Theorem: Consider two complementary filters $h_1(t)$ and $h_2(t)$ with Fourier transforms $H_1(\omega, t)$ and $H_2(\omega, t)$. The complementarity condition is:

$$H_1(\omega, t) + H_2(\omega, t) = 1 \quad \forall \omega, t \; . \qquad (10)$$

Also, assume perfect recoverability of independent components from training data, that is $\forall t$ in test data $\exists t_*$, s.t.

$$x(t) \quad = \quad x_{train}(t_*) \; , \qquad (11)$$

and t_* is given by equation (5). Then equation (9) perfectly recovers $x(t)$ from $x^{I,II}_{train}(t)$ and testing data $x^I(t)$.

Proof: According to the reconstruction rule (9), and (11)

$$\begin{aligned} \hat{\mathbf{s}}(t) &= \mathbf{s}^I(t) + \mathbf{s}^{II}_{train}(t_*) = \mathbf{A}^{-1}h_1(t) \otimes x(t) + \mathbf{A}^{-1}h_2(t) \otimes x(t) \\ &= \mathbf{A}^{-1}\mathcal{F}^{-1}\{H_1(\omega, t) + H_2(\omega, t)\} \otimes x(t) = \mathbf{A}^{-1}\mathcal{F}^{-1}\{1\} \otimes x(t) \\ &= \mathbf{A}^{-1}\delta(t) \otimes x(t) = \mathbf{A}^{-1}x(t) = \mathbf{s}(t) \quad \text{qed.} \end{aligned} \qquad (12)$$

Here $\mathcal{F}^{-1}\{\cdot\}$ is the inverse Fourier transform and δ is the Kronecker delta (unit impulse).

4. EXPERIMENTAL RESULTS

We tested the spectral extension algorithm in two variants: (1) BWE; (2) expansion with complete missing data over a fixed pattern in three or more frequency bands. We report some of these results below. In all cases we measured log-spectral distortion as in [12] and the Itakura-Saito distortion measure between wideband and estimated speech. All speech signals were taken from the NTT database, 16kHz sampling frequency. We used 64-sample frames, with a 4-sample time step, and we employed a total of 50000 frames for training. Each result averages 20 cases, for four different speakers on five different training-testing combinations, on clean data.

The inner-working of the essential matching part is depicted in Figure 3. Table 1 shows statistical results of test 1. Figure 4 shows sample results of the tests. We noticed that results improve with larger amounts of training data as expected. Overall, estimated speech had some artifacts, which decrease to imperceptible levels as more training data is used.

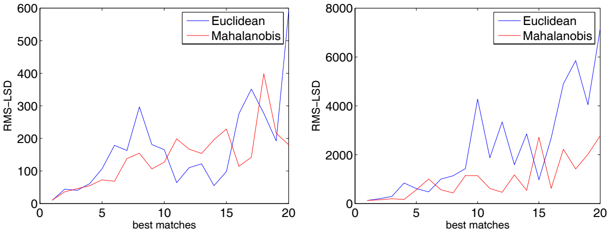

Fig. 3. Average LSD for BWE of hand chosen speech segments representing sound frames “a” and “sh”. The spectral expansion is done according to either t_* (best match, rank 1 on x-axis), or suboptimally, here just to explore effect of suboptiomal choice: match with rank 2 (second best), 3, etc.

Distance/Alg.	LSD	ISD
Euclidean	16.27 (1.412)	2.31 (0.56)
Mahalanobis	16.29 (1.513)	2.29 (0.52)

Table 1. Log-spectral distortion and Itakura-Saito distortion measures results of the BWE algorithm: means (std. deviation) for 20 cases. The algorithm used either Euclidean, or Mahalanobis distance for pattern matching.

5. CONCLUSION

This paper defines new algorithms for spectral estimation using statistics learned from data in an unsupervised way. The algorithms work in an appropriate domain, the ICA domain, by essentially performing statistical inference with missing data. A simple version of the algorithms implements bandwidth extension and is a form of nearest-neighbor hot-deck imputation for a fixed template of missing data. For the near future we suggest the following powerful generalization of the method: perform statistical spectral inference with missing data according to random patterns of missingness. This is possible and useful for speech enhancement, e.g. output from source separation methods based on time-frequency masking.

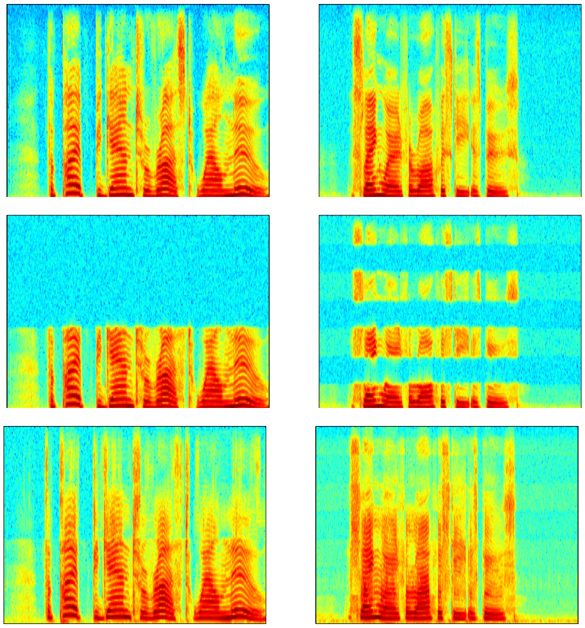

Fig. 4. (Column 1) Bandwidth extension example showing reconstruction of fine detail in a female voice signal. Top: original wide band signal; Middle: low pass signal with a cut-off frequency of 3.4 kHz; Bottom: bandwidth-extended signal. (Column 2) Generalization of the BWE for spectral estimation with multiple missing frequency bands shows reconstruction of missing data with little artifacts. In this example of a female voice three frequency bands are missing: 1140-2280, 3420-4560 and 5700-6850Hz. Original wide band signal (top), masked signal (middle), and extended signal (bottom).

6. REFERENCES

[1] L. Deng and D. O’Shaughnessy, *Speech Processing*, M. Dekker, 2003.

[2] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley and Sons, 2001.

[3] S. Rickard, R. Balan, and J. Rosca, “Real-time time-frequency based blind source separation,” in *ICA, San Diego, CA*, 2001, pp. 651–656.

[4] H. Attias, J.C. Platt, A. Acero, and L. Deng, “Speech denoising and dereverberation using probabilistic models,” in *Neural Information Processing Systems 13 (NIPS)*, 2000.

[5] R.J.A. Little and D.B. Rubin, *Statistical Analysis with Missing Data*, John Wiley and Sons, 2002.

[6] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, 1994.

[7] R. Balan, A. Jourjine, and J. Rosca, “Ar processes and sources can be reconstructed from degenerate mixtures,” in *ICA*, 1999, pp. 467–472, Aussois, France.

[8] S. T. Roweis, “One microphone source separation,” in *Neural Information Processing Systems 13 (NIPS)*, 2000, pp. 793–799.

[9] G. Hu and D. Wang, “Monaural speech separation,” in *Neural Information Processing Systems 15 (NIPS)*, S. Thrun S. Becker and K. Obermayer, Eds., pp. 1221–1228. MIT Press, Cambridge, MA, 2003.

[10] F.R. Bach and M.I. Jordan, “Blind one-microphone speech separation: A spectral learning approach,” in *Neural Information Processing Systems 17 (NIPS)*, 2004.

[11] M. Lewicki and T. Sejnowski, “Learning overcomplete representations,” *Neural Computation*, , no. 12, pp. 337–365, 2000.

[12] P. Jax, *Audio Bandwidth Extension*, chapter 6, J. Wiley and Sons, 2004.

[13] Y. Qian and P. Kabal, “Combining equalization and estimation for bandwidth extension of narrowband speech,” *ICASSP*, vol. I, pp. 713–716, 2004.

[14] T. Unno and A. McCree, “A robust narrowband to wideband extension system featuring enhanced codebook mapping,” in *ICASSP*, 2005, vol. I, pp. 805–808.

[15] L. Laaksonen, J. Kontio, and P. Alku, “Artificial bandwidth expansion method to improve intelligibility and quality of amr-coded narrowband speech,” in *ICASSP*, 2005, vol. I, pp. 809–812.

[16] A.J. Bell and T.J. Sejnowski, “Learning the higher order structure of a natural sound,” *Network: Computation in Neural Systems*, vol. 7, 1996.