Statistical Recovery of Wideband Speech from Narrowband Speech

Yan Ming Cheng, Douglas O'Shaughnessy, and Paul Mermelstein

***Abstract*—We present an algorithm to generate wideband speech from a narrowband version of the same. The main body of the algorithm is a statistical recovery function (SRF), which predicts the highband spectrum based solely on the narrowband spectrum. The performance of the algorithm has been measured both in terms of spectral distortion and spectral signal-to-noise ratio (SNR). We obtained a 3 dB gain in SNR for the reconstructed wideband speech as compared to the narrowband speech.**

I. Introduction

Wideband speech (in our experiments, covering the range 0.3–8 kHz) has generally a more pleasant quality compared with narrowband (0.3–3.75 kHz) speech. Most transmission lines carry only narrowband speech for economic reasons, and some existing communication networks do so for historical reasons. Because of the human preference for wideband speech, a solution to generate wideband speech from a narrowband transmission appears attractive. We develop here a tool to recover the spectral highband difference between wideband and narrowband speech, without the use of any additional transmitted information. The feasibility of such a tool depends on the validity of the assumption that the difference signal is closely correlated with, and is a nonlinear function of, the narrowband speech. Our experiments support the validity of this assumption.

In this paper, we present a preliminary study toward the realization of such a speech-recovery tool. The approach we adopt is to implement a recovery function at the receiver of a coded speech transmission. The function maps narrowband speech to a spectral difference signal, which is considered here only as highband (3.75–8 kHz) speech. To reconstruct the wideband speech, we add the highband component to the received narrowband speech. The recovery function is based on a statistical dependence between the narrowband and highband speech spectra and applies in a speaker-independent fashion.

II. The Statistical Recovery Function (SRF) and Its Use

A. Background

A speech signal can be segmented and assigned into classes, such as phonemes or broad phonetic classes. We interpret each class to have a distinct pattern in both low and high frequencies. If a signal in a narrowband spectrum is recognized as belonging to a certain class, the highband signal can be approximately determined by the corresponding pattern of the class. This idea can be generalized as that of a narrowband class being mapped to any highband class with a certain probability. In order to avoid hard decisions in the classification, we introduce the notion of a random source instead of class. Each random source has a probability density function (pdf), characterized by a mean and a covariance matrix. A signal of a class can be considered as, for instance, a random signal emitted from a random source with the highest probability; a transitional signal is emitted jointly by several random sources.

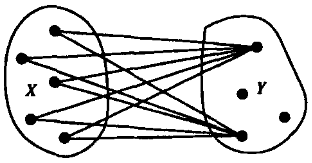

Fig. 1. Graphical illustration of a statistical recovery function. The dots in the space $\mathcal{X}$ represent random sources λ_i and those in the space $\mathcal{Y}$ represent θ_j. The lines connecting the sources in the two spaces represent the cross-correlation probabilities.

B. An Iterative Training Algorithm via the EM Algorithm

Consider a sample vector or frame of K narrowband speech samples, $\mathbf{x} = [x_0, x_1, \cdots, x_K]^T$, in a multidimensional space $\mathcal{X}$, and a sample vector of highband speech, $\mathbf{y} = [y_0, y_1, \cdots, y_K]^T$, in a space $\mathcal{Y}$. We assume that the ensemble of $\mathbf{x}$ is generated by a combination of N random sources, $\lambda_i, 1 \leq i \leq N$, and the ensemble of $\mathbf{y}$ by M random sources, $\theta_j, 1 \leq 1 \leq M$. The probability of source θ_j contributing to the highband speech, while source λ_i contributes to the narrowband speech, is defined by $\alpha_{ij} = p(\theta_j|\lambda_i)$, a cross-correlation probability. A graphical illustration is given in Fig. 1. Given a set of parameters, $A = \{\alpha_{ij}\}, \Lambda = \{\lambda_i\}$, and $\Theta = \{\theta_j\}$, and a vector of narrowband speech, $\mathbf{x}$, a recovery function f yields highband speech $\mathbf{y} = f(\mathbf{x}, A, \Lambda, \Theta)$. In the remainder of this section, we derive a training algorithm and a procedure for highband speech estimation.

Let us consider a joint pdf for the speech and individual sources at time t in the speech signal

$$\begin{aligned} p(\mathbf{y}_t, \mathbf{x}_t, \lambda_i, \theta_j) &= p(\mathbf{y}_t|\theta_j)p(\theta_j|\lambda_i)p(\mathbf{x}_t|\lambda_i)p(\lambda_i) \\ &= p(\mathbf{y}_t|\theta_j)\alpha_{ij}p(\mathbf{x}_t|\lambda_i)p(\lambda_i) \end{aligned} \tag{1}$$

since, by definition, $\mathbf{y}$ depends only on θ_j and θ_j only on λ_i. Following the frequent assumption of an all-pole (autoregressive) model for speech signals in linear predictive analysis [1], we will, for computational convenience, use pth- and qth-order autoregressive Gaussian sources to describe the random sources, both λ_i and θ_j, respectively. In cases where assuming an autoregressive model for the signal is not suitable, a standard Gaussian pdf can be used without any loss. The conditional autoregressive Gaussian pdf's of observing $\mathbf{x}_t$ and $\mathbf{y}_t$, given their underlying sources, $p(\mathbf{x}_t|\lambda_i)$ and $p(\mathbf{y}_t|\theta_j)$, can be found in [1], [2]. We use an energy-normalized version of the above pdf's, since the absolute energy is independent of the sources. In order to reconstruct energy information for a given signal, we will explicitly build up the recovery function for the energy ratio.

Given a pair of training sequences, $(\mathbf{X}, \mathbf{Y}) = \{(\mathbf{x}_t, \mathbf{y}_t)\}$ with $1 \leq t \leq T$, and a set of parameters $\Xi = \{A, \Lambda, \Theta\}$, the joint probability

$$\begin{aligned} p(\mathbf{X}, \mathbf{Y}) = p(\Xi, \mathbf{X}, \mathbf{Y}) &= \prod_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{M} p(\mathbf{x}_t, \mathbf{y}_t, \lambda_i, \theta_j) \\ &= \sum_{k=1}^{(NM)^T} \prod_{t=1}^{T} p(\mathbf{x}_t, \mathbf{y}_t, s_k(t)) \\ &= \sum_{k=1}^{(NM)^T} p(\Xi, \mathbf{X}, \mathbf{Y}, s_k) \end{aligned} \tag{2}$$

Manuscript received August 17, 1992; revised December 31, 1993. This work was supported by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Y. M. Cheng and P. Mermelstein are with Bell-Northern Research, Nuns' Island, Quebec, Canada H3E 1H6.

D. O'Shaughnessy is with INRS-Telecommunications, Nuns' Island, Quebec, Canada H3E 1H6.

IEEE Log Number 9403984.

1063-6676/94$04.00 © 1994 IEEE

where $s_k(t) = [\lambda_i, \theta_j]$ is a state-vector at time t and on the kth state-path of a treillis (the number of distinct state-vectors are NM and that of a distinct state-paths are $(NM)^T$). Given the same conditions, the conditional joint pdf of both λ_i and θ_j at time t is

$$p(t : \lambda_i, \theta_j | \mathbf{X}, \mathbf{Y}) = \frac{p(t : \lambda_i, \theta_j, \mathbf{X}, \mathbf{Y})}{p(\mathbf{X}, \mathbf{Y})} = \frac{p(\mathbf{x}_t, \mathbf{y}_t, \lambda_i, \theta_j)}{\sum_{k=1}^{N} \sum_{l=1}^{M} p(\mathbf{x}_t, \mathbf{y}_t, \lambda_k, \theta_l)} \tag{3}$$

and the conditional pdf of λ_i contributing to the speech at time t is

$$p(t : \lambda_i | \mathbf{X}, \mathbf{Y}) = \frac{p(t : \lambda_i, \mathbf{X}, \mathbf{Y})}{p(\mathbf{X}, \mathbf{Y})} = \frac{\sum_{j=1}^{M} p(\mathbf{x}_t, \mathbf{y}_t, \lambda_i, \theta_j)}{\sum_{k=1}^{N} \sum_{j=1}^{M} p(\mathbf{x}_t, \mathbf{y}_t, \lambda_k, \theta_j)}. \tag{4}$$

In order to estimate the parameters, we use the EM algorithm [3] to maximize the likelihood, $p(\Xi, \mathbf{X}, \mathbf{Y})$. In the E-step we compute the expectation of the log likelihood (or an auxiliary function in [4]), $Q(\Xi | \Xi_0) = E(\log p(\Xi, \mathbf{X}, \mathbf{Y}, s_k) | \Xi_0)$, over state-paths (see (2)), based on an initial guess of parameters, $\Xi_0 = \{A_0, \Lambda_0, \Theta_0\}$. In the M-step, we maximize the expectation function, $Q(\,)$ by adjusting Ξ. Using Lagrange optimization and the constraint $\Sigma_{j=1}^{m} p(\theta_j | \lambda_i) = 1$, it is not hard to derive (see [4])

$$\alpha_{ij} = p(\theta_j | \lambda_i) = \frac{C_{ij}}{\sum_{j=1}^{M} C_{ij}}$$

where $C_{ij} = \Sigma_{k=1}^{(NM)^T} p(\Xi_0, \mathbf{X}, \mathbf{Y}, s_k) c_{ij}(s_k)$ is the expectation of $c_{ij}(s_k)$, which is the count of a state-vector, $[\lambda_i, \theta_j]$, on the state-path, s_k. Since the expected count can be also efficiently computed as $C_{ij} = \Sigma_{t=1}^{T} p(t : \lambda_i, \theta_j, \mathbf{X}, \mathbf{Y}) = \Sigma_{t=1}^{T} p(t : \lambda_i, \theta_j | \mathbf{X}, \mathbf{Y}) p(\mathbf{X}, \mathbf{Y})$, and since $p(\mathbf{X}, \mathbf{Y})$ is independent of i and j, thus (5), at the bottom of this page. Similarly, an updating formula of the *a priori* pdf of the source, $p(\lambda_i)$, can be derived as the expected count of λ_i on a state-path divided by the total count on a state path (6), at the bottom of this page.We can also update the autocorrelation sequences of sources

$$r_{\lambda_i}(k) = \frac{\sum_{t=1}^{T} r_{x,t}(k) p(\lambda_i | \mathbf{x}_t, \mathbf{y}_t)}{\sum_{t=1}^{T} p(\lambda_i | \mathbf{x}_t, \mathbf{y}_t)} \tag{7a}$$

$$r_{\theta_j}(k) = \frac{\sum_{t=1}^{T} r_{y,t}(k) p(\theta_j | \mathbf{x}_t, \mathbf{y}_t)}{\sum_{t=1}^{T} p(\theta_j | \mathbf{x}_t, \mathbf{y}_t)} \tag{7b}$$

and a ratio of highband signal energy versus narrowband energy

$$\beta_{\theta_j} = \frac{\sum_{t=1}^{T} \left(\frac{\mathcal{E}(\mathbf{y}_t)}{\mathcal{E}(\mathbf{x}_t)} \right)^{1/2} p(\theta_j | \mathbf{x}_t, \mathbf{y}_t)}{\sum_{t=1}^{T} p(\theta_j | \mathbf{x}_t, \mathbf{y}_t)} \tag{7c}$$

where $p(\lambda_i | \mathbf{x}_t, \mathbf{y}_t)$ and $p(\theta_j | \mathbf{x}_t, \mathbf{y}_t)$ can be easily derived from $p(\lambda_i, \theta_l, \mathbf{x}_t, \mathbf{y}_t)$ and its marginal pdf's; $\mathcal{E}(\mathbf{x}_t)$ and $\mathcal{E}(\mathbf{y}_t)$ are the energies of $\mathbf{x}_t$ and of $\mathbf{y}_t$, respectively. The reason for using the energy ratio here in steady absolute energy value is that the latter value varies from signal to signal (e.g., on different telephone lines) and has little direct utility; the ratio, however, contains sufficient energy information for reconstruction of the highband signal. A set of updated autoregressive coefficients of sources can be obtained through the usual Levinson-Durbin recursive algorithm and $r_{\lambda_i}(k)$ and $r_{\theta_j}(k)$.

The following list summarizes the training algorithm:

1) Initialize parameters $\Xi_0 = \{A_0, \Lambda_0, \Theta_0\}$.
2) Iteration loop
   a) Time loop: t from 1 to T
      i) Compute $p(\mathbf{y}_t, \mathbf{x}_t, \lambda_i, \theta_j)$ according to (1).
      ii) Compute recursively the necessary cumulatives in (5), (6), and (7).
   b) End the time loop and update $\Xi_0 = \{A_0, \Lambda_0, \Theta_0\}$ according to (5), (6), and (7).
   c) Test if the stop criterion is satisfied. If yes, stop the iteration.

The EM algorithm guarantees that the developed updating formulas converge to a critical point.

$$\alpha_{ij} = \frac{\sum_{t=1}^{T} p(t : \lambda_i, \theta_j | \mathbf{X}, \mathbf{Y})}{\sum_{t=1}^{T} p(t : \lambda_i | \mathbf{X}, \mathbf{Y})} = \frac{\sum_{t=1}^{T} \frac{p(\mathbf{x}_t, \mathbf{y}_t, \lambda_i, \theta_j)}{\sum_{k=1}^{N} \sum_{l=1}^{M} p(\mathbf{x}_t, \mathbf{y}_t, \lambda_k, \theta_l)}}{\sum_{t=1}^{T} \frac{\sum_{l=1}^{M} p(\mathbf{x}_t, \mathbf{y}_t, \lambda_i, \theta_l)}{\sum_{k=1}^{N} \sum_{l=1}^{M} p(\mathbf{x}_t, \mathbf{y}_t, \lambda_k, \theta_l)}} \tag{5}$$

$$p(\lambda_i) = \frac{1}{T} \sum_{t=1}^{T} p(t : \lambda_i | \mathbf{X}, \mathbf{Y}) = \frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{j=1}^{M} p(\mathbf{x}_t, \mathbf{y}_t, \lambda_i, \theta_j)}{\sum_{k=1}^{N} \sum_{j=1}^{M} p(\mathbf{x}_t, \mathbf{y}_t, \lambda_k, \theta_j)} \tag{6}$$

C. Minimum Mean Square Estimation (MMSE) of Highband Speech

An estimation of energy-normalized highband speech through minimum mean square estimation (MMSE) is a conditional expectation

$$\begin{aligned}\hat{\mathbf{Y}} &= E(\mathbf{Y}|\mathbf{X}) = \int_{\mathcal{Y}^{T'}} \mathbf{Y} p(\mathbf{Y}|\mathbf{X})\, d\mathbf{Y} \\ &= \int_{\mathcal{Y}^{T'}} \mathbf{Y} \prod_{t=1}^{T'} p(\mathbf{y}_t|\mathbf{x}_t)\, d\mathbf{Y} \end{aligned} \tag{8}$$

where, using (1)

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{x}_t) &= \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{p(\mathbf{y}_t, \mathbf{x}_t, \lambda_i, \theta_j)}{p(\mathbf{x}_t)} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{p(\mathbf{y}_t|\theta_j)\alpha_{ij} p(\mathbf{x}_t|\lambda_i)p(\lambda_i)}{\sum_{i=1}^{N} p(\mathbf{x}_t|\lambda_i)p(\lambda_i)} \end{aligned}$$

and T' is the length of the highband speech to estimate. Since we assume that observations of both the highband and narrowband speech are independent in different time frames t, the above equation can be written in a form of vector concatenation, presented by the sign ×. Then, we calculate the highband speech estimate as

$$\hat{\mathbf{Y}} = \sum_{j=1}^{M} \frac{\hat{\mathbf{y}}_1^{(j)}}{p(\mathbf{x}_1)} \times \cdots \sum_{j=1}^{M} \frac{\hat{\mathbf{y}}_t^{(j)}}{p(\mathbf{x}_t)} \cdots \times \sum_{j=1}^{M} \frac{\hat{\mathbf{y}}_{T'}^{(j)}}{p(\mathbf{x}_{T'})} \tag{9}$$

where $\hat{\mathbf{y}}_t^{(j)} = \Sigma_{i=1}^{N} \mathbf{y}_t^{(j)} \alpha_{ij} p(\mathbf{x}_t|\lambda_i)p(\lambda_i), \mathbf{y}_t^{(j)}$ is the mean-vector of the random source θ_j at time t. Since $\mathbf{y}_t^{(j)}$is also a qth-order autoregressive process, then we have

$$[y(n)]^T = \left[\sum_{k=1}^{q} a_k^{(j)} y(n-k) + G_t^{(j)} \epsilon_t(n)\right]^T \tag{10}$$

where $\epsilon_t(n)$ is white Gaussian noise excitation at time t with zero mean and unity variance, and $a_k^{(j)}$ are the autoregressive coefficients. The same excitation sequence has to be applied to all sources θ to guarantee an identical initial phase and a smooth phase evolution in time. $G_t^{(j)}$ is a gain factor applied to the θ_j sources and estimated as (11), at the bottom of this page. Thus the estimation of $\hat{G}_t^{(j)}$ is independent of $p(\mathbf{x}_t)$. We assumed implicitly in the above highband speech generation that the highband speech exhibits no periodic behavior (i.e., pitch-periodicity is absent).

In Fig. 2, we show a diagram of the current system to recover the wideband speech. Each filter represents a random source's autoregressive spectrum in the highband. The input of each filter is weighted by a factor, $f(t,j) = G_t^{(j)} \zeta_{j,t}$, where $\zeta_{j,t} = \Sigma_{i=1}^{N} \alpha_{ij} p(\mathbf{x}_t|\lambda_i)p(\lambda_i)$.

III. Experimental Results

A. Speech Material

The speech database used contained phonetically-balanced wideband speech sampled at 16 kHz with an antialiasing filter cutting off at 7.8 kHz. The database was split into two parts. Part one, used

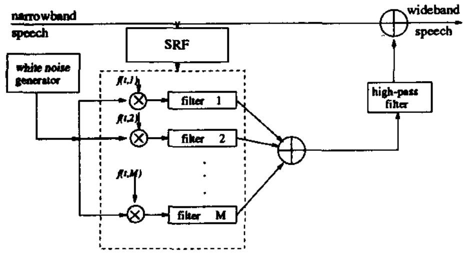

Fig. 2. Diagram of the wideband speech recovery system.

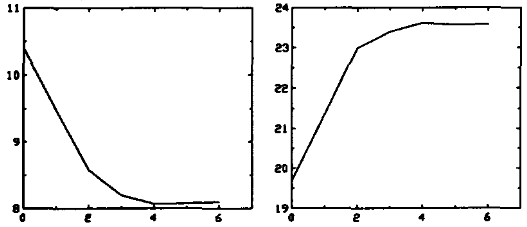

Fig. 3. Performance as a function of the number, M, of sources θ_j. The left panel shows the rms of log spectra; the right panel shows segmental SNR. The vertical axes are in dB; the horizontal axis shows $\log_2 M$ (i.e., M in bits).

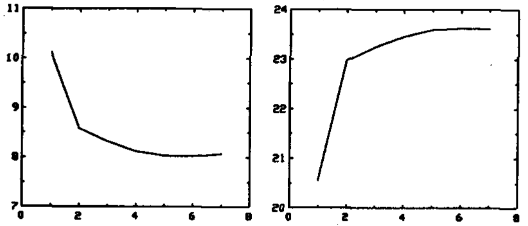

Fig. 4. Performance as a function of number, N, of sources λ_i. The left panel shows the rms of log spectra; the right panel shows the segmental SNR. The vertical axes are in dB; the horizontal axis shows $\log_2 N$ (i.e., bits).

to train the statistical recovery function, consisted of speech from four male and four female speakers. The data to test our algorithm consisted of speech from four separate speakers (two male and two female). Thus, the algorithm can be viewed as operating speaker-independently. The narrowband speech was generated by passing the wideband speech through a 0.3–3.75 kHz Chebychev bandpass filter. The frame length was 20 ms and the frame advance was 10 ms. The orders of linear prediction (autoregressive) analysis were sixteen (i.e., $p = 16$ and $q = 16$).

B. Experiments for the Training Procedure

For the training procedure, there are two factors that attracted most of our concern: Iteration convergence and initialization. For the

$$\begin{aligned}\hat{G}_t^{(j)} &= \arg\min_{G_t^{(j)}} \left(\hat{\mathcal{E}}_{y,t}^{(j)} - \frac{\hat{\mathbf{y}}_t^{(j)T}\hat{\mathbf{y}}_t^{(j)}}{p(\mathbf{x}_t)^2}\right)^2 \\ &= \arg\min_{G_t^{(j)}} \left(\mathcal{E}(\mathbf{x}_t)\left[\sum_{i=1}^{N} \beta_{\theta_j} p(\theta_j|\lambda_i)p(\mathbf{x}_t|\lambda_i)p(\lambda_i)\right]^2 - \hat{\mathbf{y}}_t^{(j)T}\hat{\mathbf{y}}_t^{(j)}\right)^2. \end{aligned} \tag{11}$$

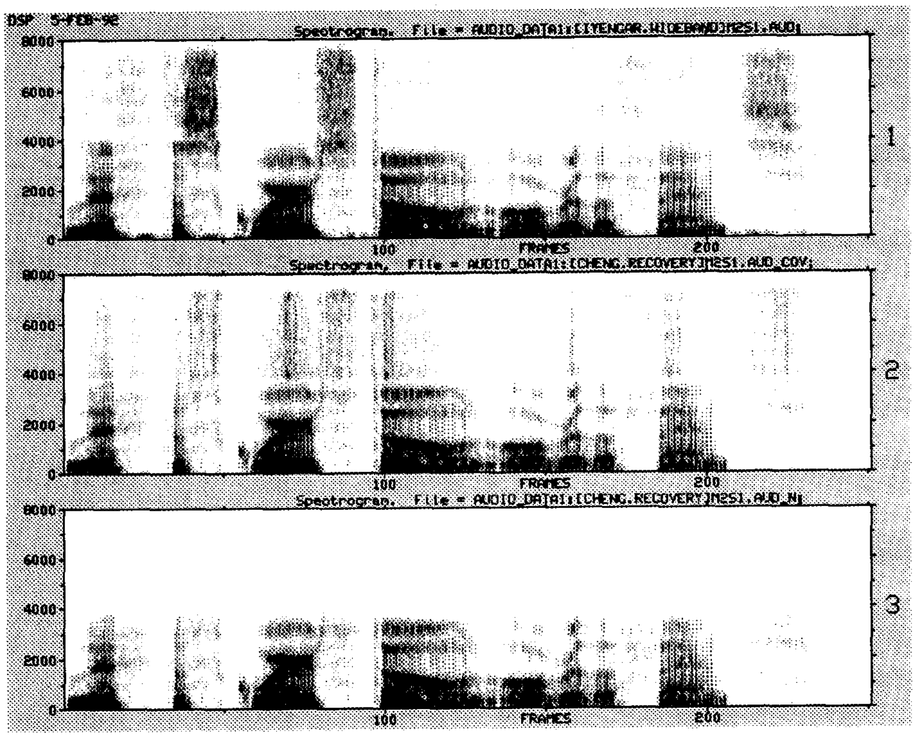

Fig. 5. Spectrograms of the original wideband speech (top), reconstructed wideband speech (middle), and narrowband speech (bottom) for a sentence, "Lift the square stone over the fence."

initialization, we had two options in these experiments: (1) vector quantization (VQ) [5] initialization and (2) bootstrap initialization. In the above two initializations, the cross-correlations are always initialized as $\alpha_{ij} = 1/M$. We have observed that for both VQ and bootstrap initialization the log likelihood increased with each iteration, thus the training convergence is practically demonstrated. Both VQ and bootstrap initializations, however, have log likelihood values very close to each other after about ten iterations. We may say that the initialization has little influence on the resulting mapping function, at least in terms of likelihood.

A more analytical way to study the training procedure is to use fully controlled data. For this purpose, we simulated the assumed data-generation process described in Section II. The parameters were $K = p = q = 2$ and four random sources at $\mathcal{X}$, the originating space, and three random sources at $\mathcal{Y}$, the destination space. The statistical coefficients of the data generation and their estimation after fifteen iterations were:

1) The mean vectors of the four sources in $\mathcal{X}$ and their estimations

$$\begin{bmatrix}-1.0\\-1.0\end{bmatrix}, \begin{bmatrix}-2.0\\-1.0\end{bmatrix}, \begin{bmatrix}-2.0\\-2.0\end{bmatrix}, \begin{bmatrix}-1.0\\-2.0\end{bmatrix},$$
$$\begin{bmatrix}-1.00\\-0.99\end{bmatrix}, \begin{bmatrix}-2.00\\-0.99\end{bmatrix}, \begin{bmatrix}-2.01\\-2.01\end{bmatrix}, \begin{bmatrix}-1.00\\-2.01\end{bmatrix}.$$

2) The mean vectors of the three sources in $\mathcal{Y}$ and their estimations

$$\begin{bmatrix}1.0\\0.0\end{bmatrix}, \begin{bmatrix}0.0\\1.0\end{bmatrix}, \begin{bmatrix}2.0\\1.0\end{bmatrix}, \qquad \begin{bmatrix}1.00\\0.00\end{bmatrix}, \begin{bmatrix}0.01\\1.00\end{bmatrix}, \begin{bmatrix}2.00\\1.00\end{bmatrix}.$$

3) The active probabilities of the four sources in $\mathcal{X}$ and their estimations

$$[0.3, 0.2, 0.2, 0.3], \qquad [0.29, 0.19, 0.21, 0.31].$$

4) The correlation matrix and their estimations

$$\begin{bmatrix}0.80 & 0.10 & 0.10\\0.30 & 0.50 & 0.20\\0.20 & 0.50 & 0.30\\0.10 & 0.10 & 0.80\end{bmatrix}, \qquad \begin{bmatrix}0.79 & 0.11 & 0.10\\0.32 & 0.50 & 0.18\\0.20 & 0.52 & 0.28\\0.10 & 0.10 & 0.81\end{bmatrix}.$$

From this controlled experiment, we have shown that the proposed algorithm estimates a correct statistical structure.

C. Experiment for the Recovery of Wideband Speech

For an assessment of the recovery algorithm we used, as criteria, spectral log rms, $D_{\rm rms}$, and segmental spectral signal-to-noise ratio (SNR), $L_{\rm SNR}$.

In the first experiment, we were very interested in the performance as a function of the number of random sources for the highband speech. The number of sources for the narrowband speech was preset to a large number ($N = 128$ in practice), which may not be efficient by was certainly sufficient. We see from Fig. 3 that the spectral log rms decreases and segmental spectral SNR increases as M increases. Above $M = 16$ (i.e., 4 bits), further changes were not significant. Secondly, fixing M at 16 we increased gradually the number of sources for narrowband speech. As N increased, a decrease in log rms and an increase in segmental spectral SNR were also observed (see Fig. 4). $N = 64$ (i.e., 6 bits) was reasonable. Compared with narrowband speech ($M = 0$), the reconstructed wideband speech with $N = 64$ and $M = 16$ showed a gain of about 3 dB in segmental spectral SNR. We note, however, that SNR is a flawed measure to evaluate performance in this context. We, thus, examined spectrograms.

In Fig. 5, we show spectrograms for an example of original wideband speech, reconstructed wideband speech, and narrowband speech. Most of the highband speech was successfully reconstructed; however, the reconstruction is not fully accurate, especially for the fricatives, /f/ and /s/. This weakness is mainly due to such fricatives' concentrating their information at highband; the narrowband versions of such fricatives do not allow easy discrimination.

IV. Conclusion

We developed a statistical recovery function (SRF) to recover wideband speech from the narrowband speech available at receivers in most communications networks. We obtained encouraging results in our preliminary study. Reconstructed wideband speech showed a gain of 3 dB in segmental SNR compared with narrowband speech, with no more than narrowband speech as input.

References

[1] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 23, no. 1, pp. 67–72, Feb. 1975.

[2] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE,* vol. 77, no. 22, pp. 257–289, 1989.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Ann. Royal Stat. Soc.,* pp. 1–38, 1977.

[4] L. E. Baum, "An inequality and associated maximization technique in satistical estimation for probabilistic functions of Markov processes," *Inequalities,* vol. 3, pp. 1–8, 1972.

[5] A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 28, no. 5, pp. 562–574, 1980.