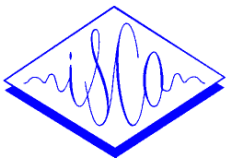

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

STATISTICAL RECOVERY OF WIDEBAND SPEECH FROM NARROWBAND SPEECH

Yan Ming Cheng[1,2], *Douglas O'Shaughnessy*[1] *and Paul Mermelstein*[1,2]

INRS-Telecommunications[1] and Bell-Northern Research[2],
16 Place du Commerce, Nuns' Island, Quebec H3E 1H6 Canada

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 on the narrowband spectrum. Assuming that bandpass portions of the speech are generated completely by a fixed number of random sources, the SRF explores the dependency among the random sources. 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. Informal perceptual experiments indicate a significant preference for the reconstructed speech.

I. Introduction

Wideband speech (in our experiments, covering the range from 0.3 to 8 kHz) has generally a more pleasant quality compared with narrowband (0.3–3.75 kHz) speech. It brings conversational partners subjectively closer together than is attained with narrowband speech and it may also improve intelligibility. 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 and because of increasing demands to improve speech communication quality toward that of face-to-face conversation, 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, either without the use of any additional transmitted information or with a little side-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. Unfortunately, as far as we know, there have been no experiments to support this assumption directly, and few efforts devoted to research in this area. Our experiments, however, 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 (from 3.75 to 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. In the next section, we present the derivation of the recovery function, algorithms for training the function, and a procedure for wideband speech reconstruction. In the third section, we study experimentally the training procedure and the recovery performance. Our conclusions are presented in the final section.

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

Consider a sample vector of narrowband speech samples, $\mathbf{x}$, in a multidimensional space $\mathcal{X}$, and a sample vector of highband speech, $\mathbf{y}$, in a space $\mathcal{Y}$. We assume that the ensemble of $\mathbf{x}$ is generated by N random sources, λ_i, $1 \le i \le N$, and the ensemble of $\mathbf{y}$ by M random sources, θ_j, $1 \le j \le M$. 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. 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.

II.A) An iterative training algorithm

Let us consider a joint pdf (probability density function) for the speech and individual sources at time t in the speech signal,

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

Since, by definition, $\mathbf{y}$ depends only upon θ_j and θ_j only upon λ_i, eq. (1) can be simplified as:

$$\begin{aligned} p(\mathbf{y}_t, \mathbf{x}_t, \lambda_i, \theta_j) &= p(\mathbf{y}_t|\theta_j)p(\theta_j|\lambda_i)p(\mathbf{x}_t|\lambda_i)p(\lambda_i) \\ &= p(\mathbf{y}_t|\theta_j)\alpha_{ij}p(\mathbf{x}_t|\lambda_i)p(\lambda_i). \end{aligned} \quad (2)$$

Furthermore, we have

$$p(\mathbf{y}_t, \mathbf{x}_t, \lambda_i) = \sum_{j=1}^{M} p(\mathbf{y}_t, \mathbf{x}_t, \lambda_i, \theta_j) \quad (3.a)$$

$$p(\mathbf{y}_t, \mathbf{x}_t, \theta_j) = \sum_{i=1}^{N} p(\mathbf{y}_t, \mathbf{x}_t, \lambda_i, \theta_j) \quad (3.b)$$

$$p(\mathbf{y}_t, \mathbf{x}_t) = \sum_{i=1}^{N}\sum_{j=1}^{M} p(\mathbf{y}_t, \mathbf{x}_t, \lambda_i, \theta_j). \quad (3.c)$$

General speaking, random sources of speech signals can often be fairly approximated by Gaussian sources [1]. Thus, for computational convenience, we use p^{th} and q^{th} order autoregressive Gaussian sources to describe the random sources, both λ_i and θ_j, respectively. The pdfs of observing $\mathbf{x}_t$ and $\mathbf{y}_t$, given their underlying sources, are [1]

$$p(\mathbf{x}_t|\lambda_i) = \exp\{-r_{x,t}(0) - 2\sum_{k=1}^{p} r_{x,t}(k)r_{a,\lambda_i}(k)\} \quad (4.a)$$

$$p(\mathbf{y}_t|\theta_j) = \exp\{-r_{y,t}(0) - 2\sum_{k=1}^{q} r_{y,t}(k)r_{a,\theta_j}(k)\}, \quad (4.b)$$

10.21437/ICSLP.1992-392

where $r_{x,t}(k)$ and $r_{y,t}(k)$ are autocorrelation sequences of the speech sample vectors $\mathbf{x}_t$ and $\mathbf{y}_t$, respectively, and $r_{a,\lambda_i}(k)$ and $r_{a,\theta_j}(k)$ are autocorrelation sequences of λ_j's and θ_j's autoregressive coefficients, respectively, computed from the autocorrelation sequences for the sources λ_i and θ_j, respectively.

Given a pair of training sequences (corresponding lowband and highband speech) over a time period T, $(\mathbf{X},\mathbf{Y}) = \{(\mathbf{x}_t,\mathbf{y}_t)\}$ with $1 \le t \le T$, and a set of parameters, A, Λ and Θ, the conditional joint pdf of both λ_i and θ_j at time t is

$$\begin{aligned} p(t:\lambda_i,\theta_j|\mathbf{X},\mathbf{Y}) &= \frac{p(t:\lambda_i,\theta_j,\mathbf{X},\mathbf{Y})}{p(\mathbf{X},\mathbf{Y})} \\ &= \frac{\prod_{\tau=1}^{t-1} p(\mathbf{x}_\tau,\mathbf{y}_\tau)p(\mathbf{x}_t,\mathbf{y}_t,\lambda_i,\theta_j)\prod_{\tau=t+1}^{T} p(\mathbf{x}_\tau,\mathbf{y}_\tau)}{\prod_{\tau=1}^{T} p(\mathbf{x}_\tau,\mathbf{y}_\tau)} \\ &= \frac{p(\mathbf{x}_t,\mathbf{y}_t,\lambda_i,\theta_j)}{p(\mathbf{x}_t,\mathbf{y}_t)} \\ &= \frac{p(\mathbf{x}_t,\mathbf{y}_t,\lambda_i,\theta_j)}{\sum_{k=1}^{N}\sum_{l=1}^{M} p(\mathbf{x}_t,\mathbf{y}_t,\lambda_k,\theta_l)}. \end{aligned} \tag{5}$$

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

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

Based on eqs. (5) and (6), a reasonable updating of the cross-correlation, α_{ij}, can be computed as the expected number of observations of both λ_i and θ_j being active divided by that of λ_i being active in the training sequence:

$$\begin{aligned} \alpha_{ij} &= \frac{\sum_{t=1}^{T} p(t:\lambda_i,\theta_j|\mathbf{X},\mathbf{Y})}{\sum_{t=1}^{T} p(t:\lambda_i,|\mathbf{X},\mathbf{Y})} \\ &= \frac{\sum_{t=1}^{T} \frac{p(\mathbf{x}_t,\mathbf{y}_t,\lambda_i,\theta_j)}{\sum_{k=1}^{N}\sum_{l=1}^{M} p(\mathbf{x}_t,\mathbf{y}_t,\lambda_k,\theta_l)}}{\sum_{t=1}^{T} \frac{\sum_{l=1}^{M} p(\mathbf{x}_t,\mathbf{y}_t,\lambda_i,\theta_l)}{\sum_{k=1}^{N}\sum_{l=1}^{M} p(\mathbf{x}_t,\mathbf{y}_t,\lambda_k,\theta_l)}}. \end{aligned} \tag{7}$$

The updating formula can be derived from a powerful EM algorithm [2]. Similarly, an updating of the *a priori* pdf of the source, $p(\lambda_i)$, can be computed as the expected number of λ_i being active divided by the total number of vector pairs in the training sequence:

$$\begin{aligned} p(\lambda_i) &= \frac{1}{T}\sum_{t=1}^{T} p(t:\lambda_i|\mathbf{X},\mathbf{Y}) \\ &= \frac{1}{T}\sum_{t=1}^{T} \frac{\sum_{j=1}^{M} p(\mathbf{x}_t,\mathbf{y}_t,\lambda_i,\theta_j)}{\sum_{k=1}^{N}\sum_{j=1}^{M} p(\mathbf{x}_t,\mathbf{y}_t,\lambda_k,\theta_j)}. \end{aligned} \tag{8}$$

We can also update the autocorrelation sequences of sources,

$$r_{\lambda_i}(k) = \frac{\sum_{t=1}^{T} r_{x,t}(k)p(\lambda_i|\mathbf{x}_t,\mathbf{y}_t)}{\sum_{t=1}^{T} p(\lambda_i|\mathbf{x}_t,\mathbf{y}_t)}, \tag{9.a}$$

$$r_{\theta_j}(k) = \frac{\sum_{t=1}^{T} r_{y,t}(k)p(\theta_j|\mathbf{x}_t,\mathbf{y}_t)}{\sum_{t=1}^{T} p(\theta_j|\mathbf{x}_t,\mathbf{y}_t)}, \tag{9.b}$$

where

$$p(\lambda_i|\mathbf{x}_t,\mathbf{y}_t) = \frac{\sum_{j=1}^{M} p(\lambda_i,\theta_j,\mathbf{x}_t,\mathbf{y}_t)}{\sum_{k=1}^{N}\sum_{j=1}^{M} p(\lambda_k,\theta_j,\mathbf{x}_t,\mathbf{y}_t)}$$

and

$$p(\theta_j|\mathbf{x}_t,\mathbf{y}_t) = \frac{\sum_{i=1}^{N} p(\lambda_i,\theta_j,\mathbf{x}_t,\mathbf{y}_t)}{\sum_{i=1}^{N}\sum_{l=1}^{M} p(\lambda_i,\theta_l,\mathbf{x}_t,\mathbf{y}_t)},$$

and a ratio of highband energy versus narrowband energy,

$$\beta_{\theta_j} = \frac{\sum_{t=1}^{T} \frac{\mathcal{E}(\mathbf{y}_t)}{\mathcal{E}(\mathbf{x}_t)} p(\theta_j|\mathbf{x}_t,\mathbf{y}_t)}{\sum_{t=1}^{T} p(\theta_j|\mathbf{x}_t,\mathbf{y}_t)} \tag{9.c}$$

where

$$p(\theta_j|\mathbf{x}_t,\mathbf{y}_t) = \frac{\sum_{i=1}^{N} p(\lambda_i,\theta_j,\mathbf{x}_t,\mathbf{y}_t)}{\sum_{i=1}^{N}\sum_{l=1}^{M} p(\lambda_i,\theta_l,\mathbf{x}_t,\mathbf{y}_t)};$$

$\mathcal{E}(\mathbf{x}_t)$ and $\mathcal{E}(\mathbf{y}_t)$ are the energies of $\mathbf{x}_t$ and of $\mathbf{y}_t$, respectively. A set of updated autoregressive coefficients of sources can be obtained through the usual Levinson-Durbin recursive algorithm and $r_{\lambda_i}(k)$ and $r_{\theta_j}(k)$.

Our training algorithm consists of performing eqs. (5)-(9) iteratively until reaching a certain stopping criterion. In the next section, we will experimentally show the convergence of the training procedure and show that the joint log likelihood, $\log p(\mathbf{X},\mathbf{Y})$, is increased at each iteration.

II.B) Minimum Mean Square Estimation (MMSE) of highband speech

For a given statistical recovery function and for a sequence of narrowband speech observations, $\mathbf{X}$, let us consider the conditional pdf of highband speech $\mathbf{Y}$,

$$p(\mathbf{Y}|\mathbf{X}) = \frac{p(\mathbf{Y},\mathbf{X})}{p(\mathbf{X})} = \frac{\prod_{t=1}^{T'} p(\mathbf{y}_t,\mathbf{x}_t)}{\prod_{t=1}^{T'} p(\mathbf{x}_t)} = \prod_{t=1}^{T'} p(\mathbf{y}_t|\mathbf{x}_t),$$

where 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 (to simplify the mathematics). Using eqs. (1) and (2), we have

$$p(\mathbf{y}_t|\mathbf{x}_t) = \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{p(\mathbf{y}_t,\mathbf{x}_t,\lambda_i,\theta_j)}{p(\mathbf{x}_t)} = \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{p(\mathbf{y}_t|\theta_j)\alpha_{ij}p(\mathbf{x}_t|\lambda_i)p(\lambda_i}{p(\mathbf{x}_t)} \tag{10}$$

where $p(\mathbf{x}_t) = \sum_{i=1}^{N} p(\mathbf{x}_t,\lambda_i) = \sum_{i=1}^{N} p(\mathbf{x}_t|\lambda_i)p(\lambda_i)$. An estimate of highband speech through Minimum Mean Square Estimation (MMSE) is a conditional expectation,

$$\hat{\mathbf{Y}} = E(\mathbf{Y}|\mathbf{X}) = \int_{\mathbf{y}^{T'}} \mathbf{Y}p(\mathbf{Y}|\mathbf{X})d\mathbf{Y} = \int_{\mathbf{y}^{T'}} \mathbf{Y}\prod_{t=1}^{T'} p(\mathbf{y}_t|\mathbf{x}_t)d\mathbf{Y}, \tag{11}$$

Introducing eq. (10) into (11), we calculate the highband speech estimate as

$$\begin{aligned} \hat{\mathbf{Y}} &= \int_{\mathbf{y}}\int_{\mathbf{y}}\cdots\int_{\mathbf{y}} \mathbf{y}_1 \times\cdots\times \mathbf{y}_{T'} \prod_{t=1}^{T'} p(\mathbf{y}_t|\mathbf{x}_t)d\mathbf{y}_1 \times\cdots\times d\mathbf{y}_{T'} \\ &= \int_{\mathbf{y}} \mathbf{y}_1 p(\mathbf{y}_1|\mathbf{x}_1)d\mathbf{y}_1 \times\cdots\times \int_{\mathbf{y}} \mathbf{y}_{T'}p(\mathbf{y}_{T'}|\mathbf{x}_{T'})d\mathbf{y}_{T'} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{\mathbf{y}_1^{(j)}\alpha_{ij}p(\mathbf{x}_1|\lambda_i)p(\lambda_i)}{p(\mathbf{x}_1)} \\ &\quad\times\cdots\times \\ &\quad \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{\mathbf{y}_{T'}^{(j)}\alpha_{ij}p(\mathbf{x}_{T'}|\lambda_i)p(\lambda_i)}{p(\mathbf{x}_{T'})} \\ &= \hat{\mathbf{y}}_1 \times\cdots\times \hat{\mathbf{y}}_{T'} \\ &= \{\, \hat{\mathbf{y}}_t \,\}, \end{aligned} \tag{12}$$

where $\mathbf{y}_t^{(j)}$ is a sample vector coming from the random source θ_j at time t and is a q^{th}-order autoregressive Gaussian process,

$$\mathbf{y}_t^{(j)} = \big[y(n)\big]^{\#} = \Big[\sum_{k=1}^{q} a_k^{(j)} y(n-k) + G_t\,\epsilon_t(n)\Big]^{\#}, \tag{13}$$

where $\epsilon_t(n)$ is white Gaussian noise at time t with zero mean and unity variance, the $a_k^{(j)}$'s are the autoregressive coefficients of the source θ_j, the sign $^\#$ stands for vector transposition, and G_t is a gain factor. The gain factor applies to all of the θ_j sources and is estimated as follows

$$\hat{G}_t = \arg\min_{G_t}\left(\hat{\mathcal{E}}_{y,t} - \hat{\mathbf{y}}_t^{\#}\hat{\mathbf{y}}_t\right)^2$$
$$= \arg\min_{G_t}\left(\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\beta_{\theta_j}\mathcal{E}(\mathbf{x}_t)p(\theta_j|\lambda_i)p(\mathbf{x}_t|\lambda_i)p(\lambda_i)}{p(\mathbf{x}_t)} - \hat{\mathbf{y}}_t^{\#}\hat{\mathbf{y}}_t\right)^2 \quad (14)$$

The $p(\mathbf{x}_t)$ in eq. (12) is similar to $\hat{G}_t$ in that both are constant factors across all i and j. Therefore, $p(\mathbf{x}_t)$ can be effectively discarded from eq. (14), in terms of its effect on the minimization of G_t.

We assumed implicitly in the above highband speech generation that the highband speech exhibits no periodic behavior (i.e., pitch-periodicity is absent). This assumption is motivated by the observation that the intensity of the periodic component in voiced glottal excitation decreases with increasing frequency. In practice, a small periodic component can be observed in the highband spectrum. The assumption of no highband periodicity can be easily removed by using an innovation sequence generated from narrowband speech, such as the technique in residual-excited linear prediction [3].

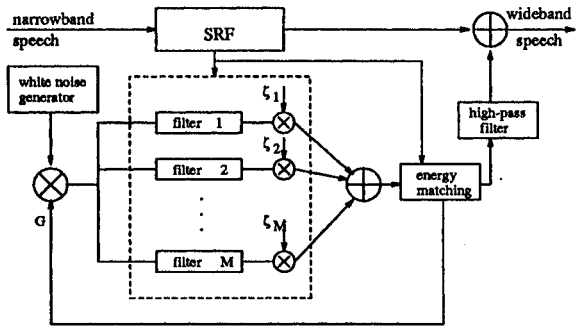

Fig. 1 Diagram of the wideband speech recovery system

In Fig. 1, we show a diagram of the current system to recover wideband speech. The SRF box computes all the pdfs for the highband speech energy, given narrowband speech as input. A random number generator produces Gaussian white noise to excite a bank of autoregressive filters. Each filter represents a random source in the highband. The output of each filter is weighted by a factor, $\zeta_j = \sum_{i=1}^{N}\alpha_{ij}p(\mathbf{x}_t|\lambda_i)p(\lambda_i)$. The component labelled "energy matching" estimates the gain factor (see eq. (14)) based on the estimate of highband speech energy.

The linear prediction analysis of the highband speech that we used introduced spectral ripples at the edges of the highband spectrum. We found that the ripples at the low frequency edge severely distorted the recovered wideband speech. To diminish the ripples' effect, the highband speech used to train the recovery function was the output of a 3–7.8 kHz Chebychev bandpass filter with the wideband speech as input, and a 4–7.8 kHz Chebychev bandpass filter was applied to the reconstructed highband speech. Finally, the recovered wideband speech was obtained through adding the 4–7.8 kHz highband speech to the narrowband speech.

III. Experimental Results

III.A) Speech material

The speech database used contained phonetically-balanced wideband speech sampled at 16 kHz with an anti-aliasing filter cutting off at 7.8 kHz. The database was split into two parts. Part one, used 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 different 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 $q = 16$).

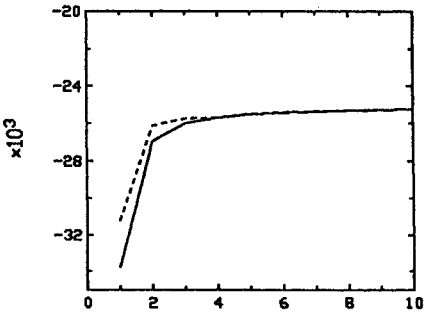

Fig. 2 Comparison of different initializations in the SRF training procedure. The solid line represents log likelihood $\log p(\mathbf{X}, \mathbf{Y})$ as a function of the number of training iterations with bootstrap initialization. The dashed line corresponds to vector quantization initialization. The vertical axis shows log likelihood in an arbitrary scale, the horizontal axis shows the number of iterations.

III.B) Experiments for the training procedure

For the training procedure, there are two factors which attracted most of our concern: iteration convergence and initialization. For the initialization, we had two options in these experiments: (1) Vector Quantization (VQ) initialization – to use the LBG method [4] to partition spaces $\mathcal{X}$ and $\mathcal{Y}$ into N and M cells, respectively, then to use the cells' coefficients as corresponding source coefficients to initialize the training procedure; (2) Bootstrap initialization – to choose randomly N and M vectors from the training $\mathbf{X}$ and $\mathbf{Y}$, respectively, then to use the coefficients of these vectors to initialize the training procedure. In the above two initializations, the cross-correlations are always initialized as $a_{ij} = \frac{1}{M}$. Fig. 2 shows the log likelihood of the joint pdf between the narrowband and highband speech in the training procedure as a function of the number of iterations for both the VQ and bootstrap initialization methods. The training convergence is thus practically demonstrated, because the log likelihood increased with each iteration. VQ initialization starts the training procedure with a higher log likelihood than the bootstrap method. Both VQ and bootstrap initializations, however, have log likelihood values very close to each other after about ten iterations. We conclude that the initialization has little influence on the resulting mapping function. From the practical point of view of minimizing computation, the training procedure with bootstrap initialization was simpler to implement and quicker to compute than that with

VQ initialization. Therefore, we utilized the training procedure with the bootstrap initialization in the remaining experiments of this paper.

III.C) Experiment for the recovery of wideband speech

For an assessment of the recovery algorithm, we used, as criteria, spectral log rms,

$$D_{rms} = \frac{1}{T'} \sum_{t=1}^{T'} \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[20 \log_{10} \frac{\hat{Y}_t(\omega)}{Y_t(\omega)} \right]^2 d\omega \right\}^{1/2}$$

and segmental spectral SNR (signal-to-noise ratio),

$$L_{SNR} = \frac{1}{T'} \sum_{t=1}^{T'} 10 \log_{10} \frac{\int_{-\pi}^{\pi} |Y_t(\omega)|^2 d\omega}{\int_{-\pi}^{\pi} |Y_t(\omega) - \hat{Y}_t(\omega)|^2 d\omega},$$

where $Y_t(\omega)$ and $\hat{Y}_t(\omega)$ are the amplitude spectra of the original and reconstructed wideband speech, respectively, at time t.

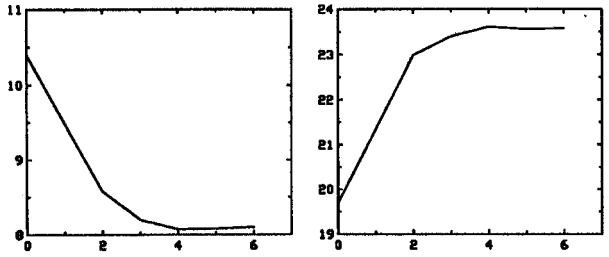

Fig. 3 Performance as a function of the number, M, of sources θ_j. The left panel shows the rms of log spectra; the right panel shows segmental SNR. The vertical axes are in dB; the horizontal axis shows $\log_2 M$ (i.e., M in bits).

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 but 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) proved a reasonable value. 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. This demonstrated that our algorithm of statistical recovery of wideband speech was successful.

Informal listening tests indicate that the reconstructed wideband speech has a perceptually more pleasant and natural quality than the hollow-sounding narrowband speech. The reconstructed wideband speech, while an improvement compared to the narrowband speech, is nonetheless clearly distinguishable from the original wideband speech. The partial reason of the unperfect reconstruction, we have thought, is that the highband speech energy is not perfectly reconstructed in the present SRF, since the mechanism of the SRF emphasizes the recovery of spectral features of the random sources. To compensate this weakness, we have used two-bits per frame to encode temporal diference of highband speech energy. We gained another 2 dB of spectral segmental

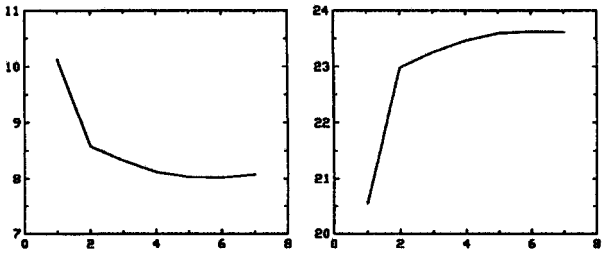

Fig. 4 Performance as a function of number, N, of sources λ_i. The left panel shows the rms of log spectra; the right panel shows the segmental SNR. The vertical axes are in dB; the horizontal axis shows $\log_2 N$ (i.e., bits).

SNR. In this case, $D_{rms} = 7.3$ dB and $L_{SNR} = 25.5$ dB. The perceptual quality of the reconstructed wideband speech was significantly improved.

IV. Conclusion

We developed a statistical recovery function (SRF) to recover wideband speech from the narrowband speech available at receivers in most communication 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. From a practical viewpoint, the developed SRF requires very little computation and can be easily embedded in the implementation of a speech decoder. Since the number of random sources in our model is not high in practice, the speech material needed for SRF training is limited. We have found the importance of better reconstructionn of highband speech energy to increase the quality of obtained wideband speech. The current simple solution, adding two-bits per frame, could be improved by more complex scheme, such as *backward prediction.* In the current system, the excitation of the highband speech is assumed to be computer-generated white Gaussian noise. Such a noise source obviously lacks correlation with narrowband speech, especially in periodic aspects. Thus, an excitation generated on-line from the narrowband speech (such as in [3]) could improve significantly the perceptual quality. Our further efforts will be directed to search for a better scheme for the reconstruction of both excitation and energy of highband speech.

Acknowledgment

This work was supported by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Reference

1. F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. ASSP*, Vol. 23, No.1, pp. 67-72, Feb. 1975.
2. A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Ann. of the Royal Stat. Society*, pp. 1-38, 1977.
3. J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding system," *Proc. IEEE Conf. ASSP*, pp. 216-219, 1979.
4. A. Buzo, A.H. Gray Jr., R.M. Gray and J.D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. ASSP*, Vol. 28, No. 5, pp. 562-574, 1980.