BANDWIDTH EXPANSION OF NARROWBAND SPEECH USING NON-NEGATIVE MATRIX FACTORIZATION

Dhananjay Bansal[1], Bhiksha Raj[2], Paris Smaragdis[2]

1.Convergys Corporation, Reston, VA, USA
2.Mitsubishi Electric Research Laboratories, Cambridge, MA, U.S.A.

ABSTRACT

In this paper, we present a novel technique for the estimation of the high frequency components (4-8kHz) of speech signals from narrow-band (0-4 kHz) signals using convolutive Non-Negative Matrix Factorisation (NMF). The proposed technique utilizes a brief recording of simultaneous broad band and narrow band signals from a target speaker to learn a set of broad-band non-negative "bases" for the speaker. The low-frequency components of these bases are used to determine how the high-frequency components must be combined in order to reconstruct the high-frequency components of new narrow-band signals from the speaker. Experiments reveal that the technique is able to reconstruct broadband speech that is perceptually virtually indistinguishable from true broadband recordings.

1. INTRODUCTION

Broad-band speech, i.e. speech signals that contain information up to 7Khz or higher, are naturally better sounding and more intelligible than narrow-band speech signals, i.e. signals that have frequencies only upto 4Khz or less, e.g. telephone quality speech. In this paper we address the problem of bandwidth-expansion, i.e. imputing high-frequency components of narrow-band signals, in order to improve their perceptual quality, if not their intelligibility.

Various solutions have already been proposed for this problem. Aliasing based methods, e.g. [1], derive high-frequency components by aliasing low frequencies into high frequencies by various means. Codebook mapping techniques (e.g. [2]) map the spectrum of the narrow-band signal onto a codeword in a codebook, and derive the upper frequencies from a corresponding high-frequency codeword. Statistical approaches utilize the statistical relationship of lower-band (< 4Khz) and upper-band (> 4Khz) frequency components to derive the latter from the former. Cheng et. al. [3] model the lower-band and upper-band components of speech as the outcome of mixtures of random processes. Mixture weights derived from the narrow-band signals are applied to the high-frequency processes to generate the upper-band frequencies in the signal. Other researchers have modelled the statistical relationships between the lower-band and upper-band components of the signal through statistical models such as Gaussian mixture models, HMMs or multi-band HMMs (e.g. [4]). Finally, linear model approaches (e.g. [5]) attempt to derive upper-band frequency components as linear combinations of lower-band components.

In this paper we present a new approach to bandwidth-expansion of narrow-band speech signals based on non-negative matrix factorisation (NMF) [7] of the magnitude spectra of speech that does not fall into any of the above categories. We represent sequences of broad-band magnitude spectral vectors as linear combinations of non-negative "bases". Bases are automatically learnt from training data. Given any new narrow-band signal, the combination of the lower-band of these bases that best explains the narrow-band signal is estimated. Upper-band frequencies are derived by combining the upper band of the bases in an identical manner.

The derivation of high-frequency components from the lower-band frequencies in a signal is a non-trivial problem. This is because the mutual information between the lower-band and upper-band frequencies in a speech signal is relatively low within any frame of speech [6]. It is therefore difficult, if not impossible to accurately predict the high-frequency components of several sounds, particularly sibilants and fricatives such as 'f' and 's'. While the use of cross-frame correlations that might enable better prediction of high frequencies, these must often be derived from complex time-series models such as HMMs, or by explicit interpolation, within the current framework of bandwidth expansion techniques.

The NMF based algorithm proposed in this paper, on the other hand attempts to learn cross-frame contextual dependencies through the use of Convolutive NMF [8], that actually learns spectral patches that are several frames wide, rather than spectral vectors, as bases. Since long-term patterns are more evident in the envelope of spectra than in the fine detail, we model the two differently, using wider spectral bases for the envelope and narrower ones for the harmonic structure. Experiments conducted on narrow-band speech signals derived from wide-band signals indicate that the proposed method can result in reconstruction of broad-band signals that sound virtually identical to the original wide-band signal.

We note that although we have described the NMF-based technique in opposition to prior methods in this section, we do not intend to claim supersession over existing techniques. Rather, we present it as yet another alternative that might potentially avoid some of the pitfalls of current techniques. The current work has several shortcomings. The learned spectral bases are speaker-specific, and hence the reconstruction is effective only for the specific speaker that the bases have been trained on. Our plans for future work include the extension of the technique to the speaker-independent scenario. No allowance has currently been made for external noise; this remains as future work.

The rest of the paper is arranged as follows: In Section 2 we briefly describe convolutive NMF. In Section 3 we describe the proposed bandwidth-expansion technique in detail. In Section 4 we present experimental results, and finally in Section 5 we present our conclusions

2. CONVOLUTIVE NON-NEGATIVE MATRIX FACTORISATION

Matrix factorisation algorithms try to decompose an $M \times N$ matrix V into two matrices W and H as

$$V \approx W.H \tag{1}$$

where W is an $M \times R$ matrix, H is an $R \times N$ matrix, and $R \leq M$, such that the error of reconstruction of V is minimized. In such decomposition, the columns of the matrix W may be interpreted as a set of basis vectors and the columns of H as the coordinates of the columns of V in terms of these bases. Alternately, the columns of H represent weights with which the basis vectors in W must be combined to obtain the closest approximation to the columns of V.

Conventional factorization techniques such as principal component analysis (PCA) and independent component analysis (ICA), etc., allow the bases vectors to comprise both positive and negative terms, and the interaction between them as specified by the components of H to be both positive and negative. In strictly non-negative data sets such as matrices that represent sequences of magnitude spectral vectors, neither the allowance for negative components in the bases nor that of negative interaction between them carries any physical meaning - magnitude cannot be negative.

In [7] Lee and Seung present a Non-negative matrix factorization (NMF) algorithm that constrains the elements of W and H to be strictly non-negative. Empirically, the basis vectors derived by NMF are often found to be physically meaningful. E.g. they are often found to represent parts of faces and text [7], or individual notes in musical pieces [8].

The NMF algorithm of Lee and Seung treats all column vectors in V as a combination of R vectors, implicitly assuming that it is sufficient to explain the structure within individual vectors to explain the entire data set. This effectively assumes that the order in which the vectors are arranged within V is irrelevant. However, these assumptions are clearly invalid in data sets such as sequences of magnitude spectral vectors, where there structural patterns are evident across multiple vectors, and the order in which the vectors are arranged is clearly important.

In [8] Smaragdis presents a convolutive version of the NMF algorithm (CNMF), wherein the bases used to explain any matrix V are not merely vectors, but actually comprise short sequences of vectors. This operation can be symbolically represented as:

$$V \approx \sum_{t=0}^{\tau} W_t^T . \overrightarrow{H^t}^T \tag{2}$$

where each W_t is a non-negative $M \times R$ matrix, H is a non-negative $R \times N$ matrix as before, and the $(\overrightarrow{t})$ operator represents a right shift operator that shifts the columns of H t positions to the right. The T in the superscript of Equation 2 represents a transposition operator. The size of H is maintained by introducing zero valued columns at the leftmost position to account for columns that have been shifted out of the matrix.

If we represent the j^{th} vector in W_t as W_t^j, each of the sets of vectors $W^j = \{W_t^j\}, t = 1..\tau$ forms a sequence of spectral vectors, or a spectral patch. These spectral patches form the bases that are used to explain the data in V. Equation 2 approximates V as the superposition of the convolution of these patches with the corresponding rows of H (i.e. the contribution of j^{th} spectral patch to the approximation of V is obtained by convolving it with the j^{th} row of H). Note that if $\tau = 1$ this reduces to conventional NMF. In order to estimate the appropriate matrices W_t and H to estimate V, we can use the already existing framework of NMD. We define a cost function as:

$$D = \left\| V \otimes \ln\left(\frac{V}{\Lambda}\right) + \Lambda - V \right\|_F \tag{3}$$

where the norm on the right side is a Froebinus norm. $\otimes$ represents a Hadmard component by component multiplication, the matrix division to the right is also per-component, and Λ is the approximation to V given by the right hand side of Equation 2. The cost function of Equation 3 is identical to the modified Kullback-Leibler cost function described by Lee and Seung, with the variation that the approximation is given by the convolutive NMF decomposition of Equation 2, instead of the linear decomposition of Equation 1.

Equation 2 can also be viewed as a set of NMF operations that are being summed to produce the final result. From this perspective, the chief distinction between Equations 1 and 2 is that the latter decomposes V in to a combination of $\tau + 1$ matrices, while the former uses only 2. This interpretation permits us to obtain an iterative procedure for the estimation of the W_t and H matrices through a simple modification of the NMF update equations of Lee and Seung The resulting update equations are given by:

$$H = H \otimes \frac{\sum_t W_t^T . \left[\frac{V}{\Lambda}\right]^{\overleftarrow{t}}}{\sum_t W_t^T . 1} \tag{4}$$

$$W_t = W_t \otimes \frac{\left[\frac{V}{\Lambda}\right] . \overrightarrow{H^t}^T}{1 . \overrightarrow{H^t}^T} \tag{5}$$

where $\otimes$ represents a Hadamard (component-by-component) multiplication, and the division operations are also component-into-component. The $(\overleftarrow{t})$ operator represents a left shift operator akin to the right shift operator in Equation 2. The overall procedure for estimating the W_t and H matrices is thus as follows: Initialize all matrices somehow (random initialization is effective); thereafter iteratively update all terms using Equations 4 and 5.

The spectral patches W^j (comprising the j^{th} columns of all the W_ts) learnt through CNMF are observed to capture salient spectrographic structures in the signal [8]. As we show in Section 4, when applied to speech, the learned bases often represent relevant phonemic or sub-phonetic structures.

3. RECONSTRUCTING HIGH FREQUENCY STRUCTURES OF A BAND LIMITED SIGNAL

The procedure for reconstructing upper-band frequencies of a narrow-band signal has three components: (i) a signal processing component where we derive separate representations for the low- and high-resolution spectra of the signal (which we refer to as the envelope and harmonic spectra respectively in the rest of this paper), (ii) a learning component, where we learn non-negative spectral bases for both the envelope and harmonic spectra, and finally (iii) a reconstruction component that performs the actual reconstruction of upper-band frequencies for narrowband signals.

3.1. Signal Processing

We assume generically that the sampling frequency for all signals is sufficient to capture both lower and upper band frequencies. Test data that have been sampled at lower frequencies must be upsampled to this rate. In this paper we have assumed a sampling frequency of 16 Khz, and all window sizes etc. are given with reference to this number.

We compute a short-time Fourier transform of the signal using a Hanning window of 512 samples (32ms) with an overlap of 256 samples between adjacent frames. Let S represent the sequence of complex Fourier spectra for a speech signal. Let Φ represent the phase and V the component-wise magnitude of S. V thus represents the magnitude spectrogram of the signal. Φ and V may be viewed as matrices, where each column represents the phase and magnitude spectra of a single 32ms frame of speech. If there are M unique points in the Fourier spectrum for any frame, and there are N frames in the signal, V and Φ are $M \times N$ matrices.

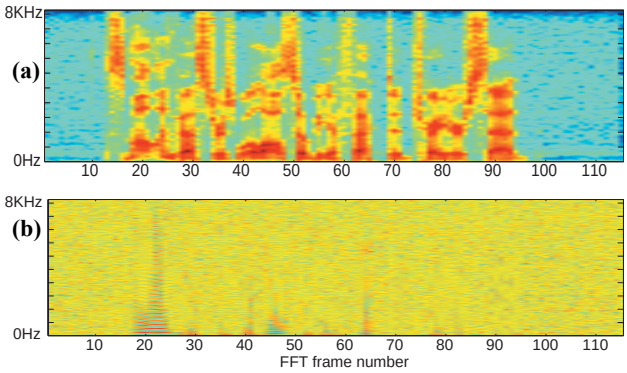

Figure 1. (a) Envelope spectrum of a typical broadband utterance. Note the formant structures. (b) Harmonic spectrum of the same utterance.

We compute the envelope and harmonic spectra of the signal by liftering V. Let V_e represent the sequence of envelope spectra derived from V, and V_h the sequence of corresponding harmonic spectra. V_e and V_h are $M \times N$ matrices derived from V as:

$$V_h = \exp(IDCT(DCT((\log(V)) \otimes Z_h))) \quad (6)$$

$$V_e = \exp(IDCT(DCT((\log(V)) \otimes Z_e))) \quad (7)$$

where Z_e is an $M \times N$ matrix such that the lower K components of each row are set to 1 and the rest of the components are set to 0. Z_h is similar to R_e except that the upper $M-K$ components are set to 1 and the rest to 0, i.e. $Z_h = 1 - Z_e$. The DCT and IDCT operations in Equations 6 and 7 are applied separately to each row of their matrix arguments.

By appropriate selection of the K, V_h and V_e can be set to model the envelope and fine-detailed harmonic structure of the spectrum of the signal. In our work we have found $K = M/3$ to be an effective setting. Figure 1 shows typical envelope and harmonic spectrograms derived with this setting for K.

The first stage of the processing scheme is to train various parameters from a corpus of training data. We compute V_e, V_h and Φ from a corpus of training data. In practice, these matrices are obtained in a two step process: In the first step, the training signals are filtered to the frequency band expected in the narrow-band test data, downsampled to the expected sampling rate of the narrow-band test data, and finally upsampled again to the sampling frequency of the full-bandwidth data. This results in signals that are a close approximation to the signals that will be obtained by upsampling narrow-band test data. Harmonic, envelope and phase spectral matrices V_h^n, V_e^n and Φ^n are obtained from the upsampled narrow band training data. Parallely, envelope, harmonic and phase spectral matrices V_e^w, V_h^w and Φ^w are also derived from the original wide-band signals in the training data. The final V_h, V_e and Φ matrices are formed from lower frequency components (below a cutoff F) from the spectral matrices for the narrow band signal and the higher frequency components of the matrices derived from the broadband data as:

$$\begin{aligned} V_e &= Z_w V_e^w + Z_n V_e^n \\ V_h &= Z_w V_h^w + Z_h V_e^n \\ \Phi &= Z_w \Phi^w + Z_n \Phi^n \end{aligned} \quad (8)$$

where Z_w is a square matrix where the first L diagonal elements are 1 and the rest of the elements are 0, and Z_n is similar matrix where the last $M-L$ diagonal elements are 1 and the rest of the elements are 0. L is set at the frequency index that corresponds to the cutoff frequency F.

3.2. Learning Spectral Bases

Spectral patch bases $W_t^e : t = 1..\tau_e$ are derived for V_e using the iterative algorithm specified by Equations 4 and 5. The H matrix also derived from this procedure is discarded. A set of *lower-band spectral envelope bases*, $W_t^{e,l}$, are derived from W_t^e are obtained by truncating all the matrices at the L^{th} row, such that each of the resulting matrices is of size $L \times R$:

$$W_t^{e,l} = Z_L W_t^e \quad (9)$$

where Z_L is an $L \times M$ matrix where the L leading diagonal terms are 1 and the rest of the terms are 0. A set of lower-band spectral harmonic bases, $W_t^{h,l}$ are also similarly obtained. The set of matrices, W_t^e, $W_t^{e,l}$, W_t^h and $W_t^{h,l}$ matrices form the spectral patch bases to be used for reconstruction.

The matrix Φ is similarly separated into an $L \times N$ low-frequency matrix Φ_l and an $(M-L) \times N$ high-frequency matrix Φ_u. A linear regression A_Φ between them is obtained as:

$$A_\Phi = \Phi_u . \text{pseudoinverse}(\Phi_h) \quad (10)$$

3.3. Reconstructing broadband signals

A narrow-band test signal is first upsampled to the sampling frequency of the broadband training signal, and phase, envelope and harmonic spectral matrices Φ, V_e and V_h are derived from it. The lower frequency components of the matrices are separated out as $V_e^l = Z_L V_e$ and $V_h^l = Z_L V_h$.

CNMF approximations are obtained for V_e^l and V_h^l based on the $W_t^{e,l}$ and $W_t^{h,l}$ bases obtained from the training data. This approximates V_e^l and V_h^l as:

$$V_h^l \approx \sum_{t=0}^{\tau_h} (W_t^{h,l})^T . (\overrightarrow{{}^t H_h})^T \text{ and } V_e^l \approx \sum_{t=0}^{\tau_e} (W_t^{e,l})^T . (\overrightarrow{{}^t H_e})^T \quad (11)$$

The H_h and H_e matrices are obtained through iterations of Equation 4.

Complete wide-band spectrograms are reconstructed by applying the estimated H_h and H_e matrices to the complete bases W_t^e and W_t^h learned during training:

$$\overline{V}_h = \sum_{t=0}^{\tau_h} (W_t^h)^T . (\overrightarrow{{}^t H_h})^T \text{ and } \overline{V}_e = \sum_{t=0}^{\tau_e} (W_t^e)^T . (\overrightarrow{{}^t H_e})^T \quad (12)$$

The upper-band frequencies of $\overline{V}_h$ and $\overline{V}_e$ are overlaid onto V_h and V_e as

$$\hat{V}_h = Z_w \overline{V}_h + Z_h V_h \text{ and } \hat{V}_e = Z_w \overline{V}_e + Z_h V_e \quad (13)$$

The complete magnitude spectrum for the signal is obtained as $\hat{V} = \hat{V}_h \otimes \hat{V}_e$. The phase for the reconstructed signal is:

$$\hat{\Phi} = (Z_h + Z_U A_\Phi Z_L)\Phi \quad (14)$$

where Z_U is an $M \times L$ matrix whose (M-L) leading diagonal elements are 1 and the rest of the elements are 0. Note that the lower frequency components of both $\hat{\Phi}$ and $\hat{V}$ are identical to the corresponding components of the original Fourier spectral matrix obtained from the test signal. The complete broadband signal is obtained by computing the inverse short-time Fourier transform of $\hat{V}e^{j\hat{\Phi}}$.

4. EXPERIMENTAL RESULTS

Experiments were conducted on two sets of signals, the first

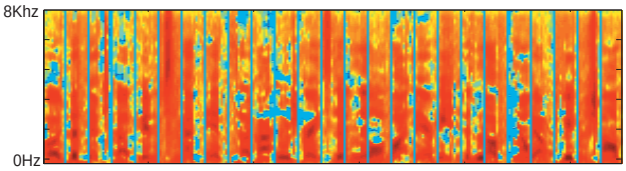

Figure 2. Examples of spectral patch bases derived from the envelope spectra for a speaker from the WSJ corpus. Note that many of the bases appear to capture entire spectral trajectories for a part of a phoneme.

consisting of all the recordings from one of the speakers in the "speaker dependent" component of the Wall street journal corpus, and the second on open-mic recordings obtained from a male speaker. All signals were sampled at 16Khz. In each set, five minutes of full-bandwidth recordings from the speakers were utilized as training data; the rest of the speech was used as test data. The test data were filtered and downsampled to 4Khz.

All signals were analyzed with 32ms windows, both for training and testing, resulting in a 512 point Fourier spectrum with 257 unique points. Envelope and harmonic spectra were obtained by utilizing a K value of 85 in Equations 5 and 6. 50 spectral patch bases of width $\tau_e = 8$ were obtained for the envelope spectra. 100 spectral bases with $\tau_h = 1$ were obtained for the harmonic spectra. Figure 2 shows a number of the envelope bases. Several of them are observed to capture spectral trajectories of phoneme segments. In particular, some are seen to capture fricatives.

On the wall street journal data, it was assumed that the narrowband signals extended until 4000 Hz (i.e. the cutoff frequency F was assumed to be 4Khz). For the bandwidth expansion, all frequencies from 4-8 Khz were reconstructed. Figure 3a shows the spectrum of a signal reconstructed in this manner. On the second data set, the cutoff frequency below which reliable low-frequency components could be obtained from the narrowband signal was set to a more realistic 3.7Khz. Bandwidth expansion only reconstructed frequencies upto 6500Hz in this case. Figure 3b shows and example of a signal reconstructed in this manner. Additional audio samples may be downloaded from http://www.cs.cmu.edu/~bhiksha/audio

5. Observations and Conclusions

As can be observed from Figure 3, the bandwidth expansion technique proposed in this paper is able to reconstruct higher frequencies of the signal very accurately. As the audio samples demonstrate, the reconstructed signals are perceptually indistinguishable from the original wide-band signals that the test data were derived from. The proposed method thus promises to be highly effective for bandwidth expansion of narrowband speech.

However, the algorithm as presented here can only be considered preliminary. The current implementation is speaker-specific: CNMF bases were derived from speaker specific training data. It remains to be determined if the technique will work in a speaker independent manner. Encouragingly, excellent results have been obtained even though the training data used for the speakers in all of our experiments was typically less than 5 minutes. This leads us to speculate that reliable speaker-independent bases can be obtained from multi-speaker corpora, since the characteristics of any speaker or group of speakers can be captured from relatively small amounts of data.

It must also be determined if speaker-independent bases so obtained are language independent since the bases appear to capture spectral trajectories that resemble phonemes. The test data in our experiments were relatively noise free. The effect of noise on the proposed technique remains to be evaluated. Finally, all test recordings were obtained by downsampling broadband

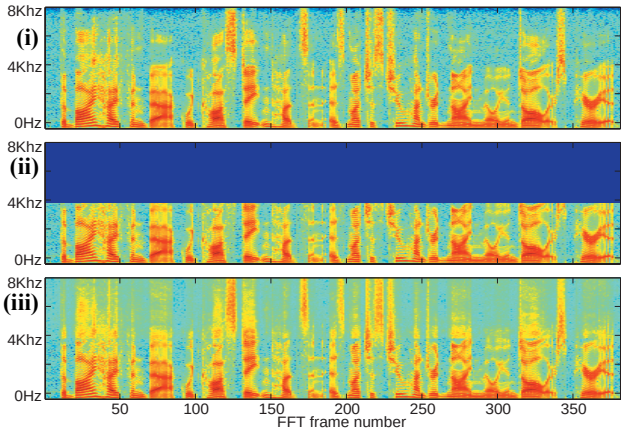

Figure 3a. i) original broadband signal. ii) upsampled narrow-band signal with frequencies up to 4Khz iii) reconstructed broadband signal.

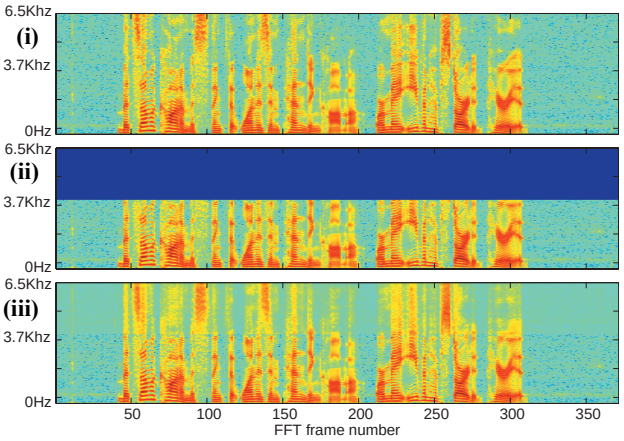

Figure 3b. i) original broadband signal. ii) upsampled narrow-band signal with frequencies up to 3.7Khz. iii) reconstructed broadband signal with frequencies up to 6.5Khz.

speech. The effectiveness of the algorithm on real telephone recordings remains to be evaluated. We expect to cover all these aspects in future work.

REFERENCES

1. Yasukawa, H. (1996). "Signal Restoration of Broad Band Speech Using Nonlinear Processing", Proc. European Signal Processing Conf. (EUSIPCO-96), pp 987-990.

2. Chennoukh, S., Gerrits, A., Miet, G. and Sluijter, R. (2001). "Speech Enhancement via Frequency Bandwidth Extension using Line Spectral Frequencies"., Proc ICASSP-95.

3. Cheng, Y.M., O'Shaugnessey, D.O., and Mermelstein, P. (1994). "Statistical Recovery of Wideband Speech from Narrowband Speech", IEEE Trans., ASSP, Vol 2., pp 544-548, 1994.

4. Hosoki, M., Nagai, T. and Kurematsu, A. (2002). "Speech Signal Bandwidth Extension and Noise Removal Using Subband HMM"., Proc. ICASSP 2002.

5. Avendano, C., Hermansky, H., and Wand, E.A. (1995). "Beyond Nyquist: Towards the Recovery of Broad-bandwidth Speech from Narrow-bandwidth Speech.". Proc. Eurospeech-95.

6. Nilsson, M., Andersen, S.V. and Kleijn, B. (2000). "On the Mutual Information Between Frequency Bands in Speech", Proc. ICASPP 2000, pp. 1327-1330.

7. Lee, D.D and H.S. Seung. "Learning the parts of objects with nonnegative matrix factorisation". In Nature 401:788-791, 1999.

8. P. Smaragdis, "Discovering Auditory Objects Through Non-Negativity Constraints," SAPA 2004, Jeju, Korea, October 2004.