BANDWIDTH EXPANSION WITH A PÓLYA URN MODEL

Bhiksha Raj†, *Rita Singh*‡, *Madhusudana Shashanka*†, *Paris Smaragdis*†

† Mitsubishi Electric Research Labs, Cambridge, MA, USA
‡ Haikya Corp., Watertown, MA, USA

ABSTRACT

We present a new statistical technique for the estimation of the high frequency components (4-8kHz) of speech signals from narrow-band (0-4 kHz) signals. The magnitude spectra of broadband speech are modelled as the outcome of a Pólya Urn process, that represents the spectra as the histogram of the outcome of several draws from a mixture multinomial distribution over frequency indices. The multinomial distributions that compose this process are learnt from a corpus of broadband (0-8kHz) speech. To estimate high-frequency components of narrow-band speech, its spectra are also modelled as the outcome of draws from a mixture-multinomial process that is composed of the learnt multinomials, where the counts of the indices of higher frequencies have been obscured. The obscured high-frequency components are then estimated as the expected number of draws of their indices from the mixture-multinomial. Experiments conducted on bandlimited signals derived from the WSJ corpus show that the proposed procedure is able to accurately estimate the high frequency components of these signals.

Index Terms— Signal restoration, Signal reconstruction, Speech enhancement

1. INTRODUCTION

In this paper we address the problem of *bandwidth expansion* – the automated imputation of absent frequency components of a band-limited speech signal. Numerous techniques for bandwidth expansion have been proposed in the literature. Typically, these techniques address the problem of constructing high-frequency components of telephone quality speech, since, as is well known that appropriate introduction of high-frequency components in such signals makes them perceptually more pleasing, although not necessarily more intelligible. Aliasing based methods, e.g. [1], construct the absent high-frequency components by aliasing low frequencies through non-linear transformations of the signal. 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. Linear model approaches (e.g. [3]) attempt to derive upper-band frequency components as linear combinations of lower-band components. Statistical approaches utilize the statistical relationships between the lower and higher-band frequency components of speech to derive the latter from the former. Typically, the statistical relationships are characterized through joint distributions of high- and low-frequency components, represented by models such as Gaussian mixture models, HMMs or multi-band HMMs (e.g. [4]). Alternately, they may be captured through dimensionality reduction techniques such as non-negative matrix factorization [5].

The approach presented in this paper is statistical in nature and follows the above-mentioned premise of exploiting interdependencies between the occurrence of frequency bands to estimate missing frequency components. The statistical model used, however differs from conventional statistical models in the definition of the underlying random variable. Conventional statistical models for speech model the distribution of spectral energies (or log energies) in various frequency bands. The random variable – the energy – is continuous in nature whose distribution must be characterized through hypothesized functional forms, such as Gaussian density functions.

In contrast, in this paper we define the *frequencies* in the speech signal (rather than the *energy* at any frequency) as the random variable. If spectral decomposition of the signal is achieved through a discrete Fourier transform, the frequencies are discrete, thus forming a discrete random variable. The magnitude spectrum of any segment of speech is modelled as the outcome of many draws of frequencies from a mixture multinomial distribution over the discrete frequency indices[1]. Every spectrum thus has an underlying mixture multinomial distribution. The component multinomials of the mixture are assumed to belong to a prespecified set; only the mixture weights with which the components combine are specific to the spectrum itself.

The set of component multinomials are learned from a corpus of broadband speech. In order to expand the bandwidth of a band-limited signal, the mixture multinomial distribution underlying the magnitude spectrum of each analysis window is estimated. Missing frequency bands are marginalized out of the component multinomials in order to estimate mixture weights. The missing frequencies are then estimated as the expected number of draws of these frequencies from the estimated mixture multinomial, given the number of draws of other observed frequencies. While the proposed method is suitable for the imputation of *any* set of absent frequency bands, we have specifically evaluated it in the context of expanding the bandwidth of telephone-quality speech. Perceptual and qualitative evaluations show that the technique is able to accurately reconstruct missing high-frequencies of band-limited signals, even for sounds such as low-energy fricatives for which bandwidth expansion has traditionally been considered difficult.

The rest of the paper is organized as follows. In Section 2 we describe our mixture multinomial model for speech spectra. In Section 3 we describe how absent frequencies in a spectrum may be estimated using the proposed model. In Section 4 we describe how we determine the phases of absent frequencies. In Section 5 we describe the complete bandwidth expansion algorithm in detail, and in Section 6 we present experimental results.

Although the proposed method is highly effective, it still has several shortcomings as noted in the conclusions in Section 7. The statistical models learned must be speaker-specific for the method to be most effective in its current form. Temporal correlations etc. are

[1]This may be viewed as an instance of a Pólya urn model with simple replacement

not being considered. Thus, the current paper must only be considered to be a presentation of the basic premise of a new technique. Various extensions that will address its current shortcomings will be devised in future work.

2. THE MIXTURE MULTINOMIAL MODEL

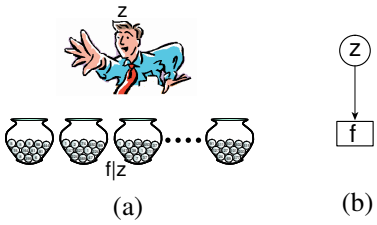

Fig. 1. a) Urn and ball illustration of mixture-multinomial model for spectra. A "picker" randomly selects urns and draws balls marked with frequency indices from the urns. The spectrum is a histogram of the draws. b) Corresponding graphical model. A latent variable z determines the probability with which frequency f is selected.

The mixture multinomial model described in this section models the structure of the magnitude spectral vectors (henceforth simply referred to as "spectral vectors") of speech. It is assumed that all speech signals are converted to sequences of spectral vectors through a short-time Fourier transform. The term "frequency" in the following discussion actually refers to the frequency indices of the DFT employed by the STFT.

We explain the mixture multinomial model for magnitude spectra through the urn-and-ball example of Figure 1a. A stochastic picker has a number of urns, each of which contains a number of balls. Every ball is marked with one of N frequency values. Each urn contains a different distribution of balls. The picker randomly selects one of the urns, draws a ball from it, notes the frequency on the ball and returns it to the urn. He repeats the process several times. He finally plots a histogram of the frequencies noted from the draws. The probability distribution of the balls from any urn in this example is a multinomial distribution. The overall distribution of the process is a mixture multinomial distribution. By our model, the number of times a particular frequency is drawn represents the value of the spectrum at that frequency. The complete histogram represents the magnitude spectrum of the analysis frame. Graphically, the mixture multinomial model may be represented by Figure 1b: a latent variable z determines the probability with which a frequency f is drawn. The latent variable z represent the urns and the probability of drawing a frequency $P(f|z)$ represents the probability with which f may be drawn from the z^{th} urn.

It must be noted that Figure 1 represents the mixture multinomial distribution *underlying* a single spectral vector – the spectral vector itself is obtained by several draws from the distribution. The parameters of the underlying model vary from analysis frame to analysis frame with one important constraint: we assume that the component multinomial distributions remain constant across all analysis frames, while the mixture weights for the components vary. In terms of the urn-and-ball simile, this means that the set of urns remains the same for all frames; however the picker selects urns according to a different probability distribution in every frame. Thus the overall mixture multinomial distribution model for the spectrum of the t^{th} frame is given by

$$P_t(f) = \sum_z P_t(z)P(f|z) \tag{1}$$

where $P_t(z)$ represents the *a priori* probability of z in the t^{th} analysis frame and $P_t(f)$ represents the multinomial distribution underlying the spectrum of the t^{th} frame.

The parameters of the distributions are learnt from a corpus of training speech signals through iterations of the following equations, that have been derived using the EM algorithm:

$$P_t(z|f) = \frac{P_t(z)P(f|z)}{\sum_{z'} P_t(z')P(f|z')} \tag{2}$$

$$P(f|z) = \frac{\sum_t P_t(z|f)S_{t,f}}{\sum_{f'}\sum_t P_t(z|f')S_{t,f'}} \tag{3}$$

$$P_t(z) = \frac{\sum_f P_t(z|f)S_{t,f}}{\sum_{z'}\sum_f P_t(z'|f)S_{t,f}} \tag{4}$$

where $S_{t,f}$ represents the f^{th} frequency band of the the t^{th} spectral vector in the training corpus.

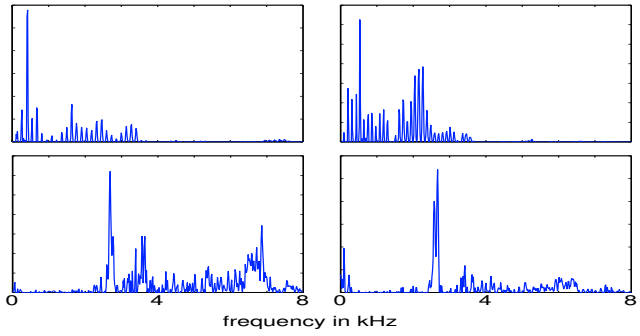

Fig. 2. Multinomial bases learnt for a speaker. The top panels show examples of bases that capture harmonic characteristics of voiced sounds. The lower panels show broadband bases that represent fricated components of speech.

The time-invariant multinomial distributions $P(f|z)$ represent the basic building blocks for the mixture multinomials underlying all spectral vectors. They may hence be viewed as the "basis vectors" that explain speech spectra. Figure 2 shows several basis vectors learnt from training examples for a male speaker.

In order to learn the generic spectral characteristics of all speech in a speaker independent manner, the training corpus must include speech from a large number of speakers, and a correspondingly large number of multinomial bases must be learnt. However, if the spectral vectors are obtained from N-point DFTs, no more than $N/2 + 1$ independent multinomial bases can be learnt, limiting the ability of the model to capture spectral patterns in a speaker-independent manner. To counter this problem, techniques that enable learning of *overcomplete* representations, (*e.g.* [6][2]) must be employed. In this paper however, we restrict ourselves to speaker-dependent modelling for simplicity.

[2]also submitted to ICASSP 2007

3. IMPUTING UNSEEN FREQUENCIES IN A SPECTRAL VECTOR

Once the parameters of the mixture multinomial model have been learned, it can be used to impute the values of unseen or obscured frequency components in a spectral vector. Let S represent a spectral vector whose components $S_f : f \in \mathcal{F}$ are observed, and the rest, $S_f : f \in \bar{\mathcal{F}}$ are obscured or missing. For example, for the spectrum of a frame of a telephone-bandwidth signal $\mathcal{F}$ would represent the set of all frequencies between 300Hz and 3.7Khz (that are actually present in the signal) and $\bar{\mathcal{F}}$ would represent all other frequencies (that are missing[3]).

The first step in the imputation process is the determination of the mixture multinomial distribution underlying the complete spectrum. This distribution is given by:

$$P_S(f) = \sum_z P_S(z)P(f|z) \tag{5}$$

where the multinomial bases $P(f|z)$ are the ones that have been learnt from training data. The mixture weights $P_S(z)$ are learnt from the partially observed spectrum by iterations of the following equations:

$$\begin{aligned} P_S(z|f) &= \frac{P_S(z)P(f|z)}{\sum_{z'} P_S(z')P(f|z')} \qquad \forall f \in \mathcal{F} \\ P_S(z) &= \frac{\sum_{f\in\mathcal{F}} P_S(z|f)S_f}{\sum_{z'}\sum_{f\in\mathcal{F}} P_S(z'|f)S_f} \end{aligned} \tag{6}$$

Equation 6 has been derived from Equations 3 and 4, with the distinction that all computation is now performed only over the set of *observed* frequencies $\mathcal{F}$.

The complete spectral vector represents the histogram of an unknown number of draws from the distribution of Equation 5. The expected number of total draws from the distribution can be estimated from the observed frequencies as

$$\hat{N} = \frac{\sum_{f\in\mathcal{F}} S_f}{\sum_{f\in\mathcal{F}} P_S(f)} \tag{7}$$

The unobserved frequency components of the spectrum can now be estimated as

$$\hat{S}_f = \hat{N}P_S(f) \quad \forall f \in \bar{\mathcal{F}} \tag{8}$$

4. PREDICTING THE PHASE OF UNSEEN FREQUENCIES

The bandwidth expansion algorithm must not only estimate the magnitude of the missing spectral components, but also their phase. The mixture multinomial model described in the earlier section is only effective at predicting the magnitudes of unseen frequency components of spectral vectors. A separate procedure is required to estimate their phase. It is known that the human ear is relatively insensitive to phase variations in higher frequencies. As a result, prior approaches to bandwidth expansion of narrow-band signals have used a variety of simplistic methods for the estimation of the phase of high-frequency components, such as the replication of the phase or lower-band components. Telephone bandwidth signals, however, are also missing very low frequencies, at which human sensitivity to phase is significant. At these frequencies, techniques such as phase duplication or random selection can result in artefacts in the bandwidth-expanded signal.

We have found that the most effective way for estimating the phase of frequency components is to model them through a linear transform of the phase of observed frequency components. Let $\Phi_{\mathcal{F}}$ represent a vector of the phases of the frequency components in $\mathcal{F}$. Similarly, let $\Phi_{\bar{\mathcal{F}}}$ represent the vector of phases of the unseen frequency components. We estimate $\Phi_{\bar{\mathcal{F}}}$ as

$$\Phi_{\bar{\mathcal{F}}} = A_\Phi \Phi_{\mathcal{F}} \tag{9}$$

where A_Φ is a matrix.

A_Φ is also learnt from the training corpus. Let $\mathbf{\Phi}_{\mathcal{F}}$ represent a matrix composed of phase vectors comprising the phases of frequency components in $\mathcal{F}$ of spectral vectors from the training data. Similarly let $\mathbf{\Phi}_{\bar{\mathcal{F}}}$ represent the matrix of the corresponding phase vectors from the training data representing frequencies in $\bar{\mathcal{F}}$. A_Φ is obtained as the following least-squared error estimate

$$A_\Phi = Pinv(\mathbf{\Phi}_{\mathcal{F}})\mathbf{\Phi}_{\bar{\mathcal{F}}} \tag{10}$$

where $Pinv(\mathbf{\Phi}_{\mathcal{F}})$ represent the pseudo inverse of $\mathbf{\Phi}_{\mathcal{F}}$.

5. COMPLETE BANDWIDTH EXPANSION ALGORITHM

We assume generically that the sampling frequency for all signals is sufficient to capture all desired frequencies (including 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 1024 samples (64ms) with an hop of 256 samples between adjacent frames. The magnitudes and phases of the frequency components are derived from the STFT.

In the training phase, a training corpus of broad-band speech is parameterized as described above. Mixture multinomial bases $P(f|z)$ are extracted from the magnitude spectra of the training speech using the algorithm described in Section 2. The linear transform matrix A_Φ that relates the phases of the frequency components that we expect to observe in the band-limited signal and the phases of frequencies that will not be observed is also estimated.

In the operational phase, any band-limited signal whose missing frequency components must be filled is first resampled, if necessary, to 16Khz and parameterized using an STFT as described above. Magnitude and phase components of the observed frequencies are obtained from the STFT. The magnitudes of missing frequency components of each spectral vector are estimated using the procedure described in Section 3. The phases of the missing frequency components are estimated as described in Section 4. The bandwidth expansion operation is performed separately for each spectral vector in the band-limited signal. Once the missing frequency components of all spectral vectors have been estimated, the now-complete STFT is inverted to obtain a full-bandwidth signal.

6. EXPERIMENTAL EVALUATION

Experiments were conducted on recordings from six speakers, three male and three female, from the "speaker independent" component of the Wall Street Journal Corpus. For each speaker, approximately ten minutes of full-bandwidth recordings were used to train mixture multinomial bases, while the rest were used as test data. The

[3] it is assumed that the signal is sampled at the same rate as the broadband signals from which multinomial bases have been learnt.

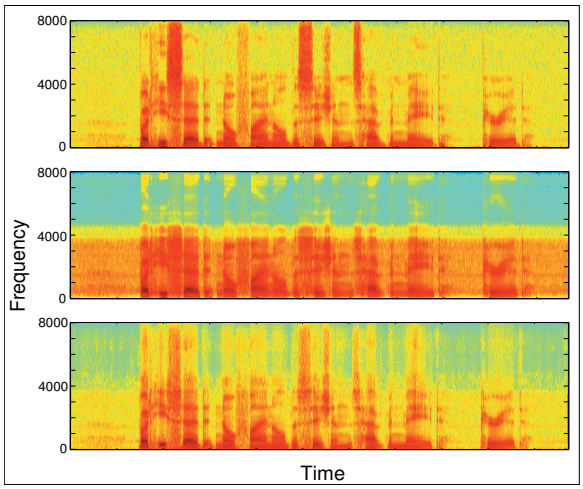

Fig. 3. The top panel shows the spectrogram of a broad-band speech signal from a male speaker. The center panel is shows the spectrogram of the signal after the 0-300Hz and 3700-8000Hz frequency bands have been filtered out. The bottom panel shows the spectrogram of the output of the bandwidth-expansion algorithm.

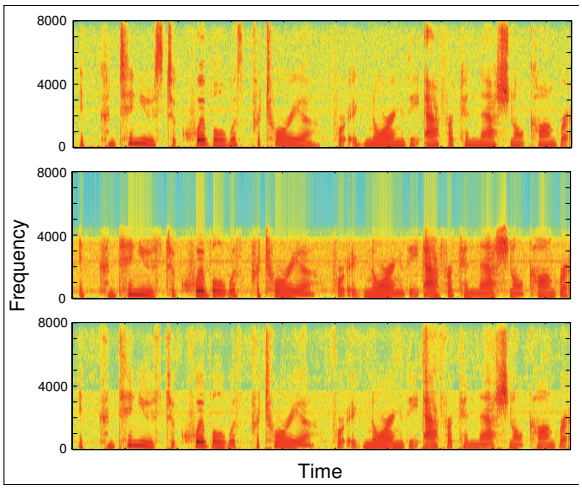

Fig. 4. Spectrograms of broad-band, narrow-band and bandwidth-expanded signals for a female speaker.

full-bandwidth training data are sampled at 16Khz. Test recordings were filtered using a 10th order Butterworth filter to only include frequencies in the range 300Hz-3700Hz, such as might be expected in signals captured over a telephone channel.

Both training and test signals were analyzed using 64ms analysis windows, corresponding to 1024 samples, resulting in Fourier spectra with 513 unique points. Adjacent frames overlapped by 768 points. 100 multinomial bases were computed for each speaker.

The missing frequency bands corresponded to the the frequency indices in the range 1-19 and 238-513. The magnitudes and phases of missing frequency bands were estimated and the complete bandwidth-expanded signals obtained as described in the paper.

Figure 3 shows the results of bandwidth expansion on a signal from a male speaker. Figure 4 shows a similar example from a female speaker. In both cases, the algorithm is able to reconstruct a very good facsimile of the missing upper ($>$3700Hz) and lower ($<$300Hz) frequencies. Perceptually, we find that the reconstructed signals are very close (although not identical) in quality to the original broadband signal. There are no discernible distortions. These and other example reconstructions can be downloaded from http://www.cs.cmu.edu/ bhiksha/audio.

7. CONCLUSIONS

The proposed bandwidth expansion technique is able to reconstruct higher frequencies of the signal very accurately. As the audio samples demonstrate, the reconstructed signals are perceptually very similar to the original broadband signals that the test data were derived from. However, the algorithm as presented here has several restrictions associated with it. In the experiments reported in Section 6, the bases used to expand any speaker's speech were speaker specific. For speaker independence, a large number of bases are required; however the maximum-likelihood formulation for the learning of bases that has been presented in this paper does not permit the learning of more bases than the number of independent frequency components in the spectrum. To learn a larger number of bases, as might be needed to sustain speaker-independent implementation of the algorithm, sparse overcomplete learning methods must be employed. The current implementation does not utilize temporal dependencies between spectral vectors. Such dependencies, however, are easily incorporated into the proposed model. The current work does not employ *priors* on the distribution of mixture weights for the mixture multinomial densities. The incorporation of priors into the proposed framework is also straightforward. We will be investigating these extensions in future work.

8. REFERENCES

[1] H. Yasukawa, "Signal restoration of broad band speech using nonlinear processing," in *Proc. European Signal Processing Conference (EUSIPCO-96)*, 1996.

[2] Gerrits A. Miet G. Sluijter R. Chennoukh, S., "Speech enhancement via frequency bandwidth extension using line spectral frequencies," in *Proc. IEEE Intl. Conf. on Acoustis Speech and Signal Processing (ICASSP-95)*, 1995.

[3] Hermansky H. Wand E.A. Avendano, C., "Beyond nyquist: Towards the recovery of broad-bandwidth speech from narrow-bandwidth speech," in *Proc. Eurospeech-95*, 1995.

[4] Nagai T. Hosoki, M. and A. Kurematsu, "Speech signal bandwidth extension and noise removal using subband hmm," in *Proc. IEEE Intl. Conf. on Acoustis Speech and Signal Processing (ICASSP-02)*, 2002.

[5] Raj B. Smaragdis P. Bansal, D., "Bandwidth expansion of narrowband speech using non-negative matrix factorization," in *Proc. Interspeech 2005*, 2005.

[6] Raj B. Shashanka, M.V.S and P. Smaragdis, "Sparse overcomplete decomposition for single channel speaker separation," in *Submitted to IEEE Intl. Conf. on Acoustis Speech and Signal Processing (ICASSP 2007)*, 2007.