JOINT DICTIONARY TRAINING FOR BANDWIDTH EXTENSION OF SPEECH SIGNALS

Jishnu Sadasivan[1]*, Subhadip Mukherjee*[2]*, and Chandra Sekhar Seelamantula*[2]

[1]Department of Electrical Communication Engineering, [2]Department of Electrical Engineering
Indian Institute of Science, Bangalore 560012, India
Emails: jishnus@ece.iisc.ernet.in, subhadip@ee.iisc.ernet.in, chandra.sekhar@ieee.org

ABSTRACT

We address the problem of extending the bandwidth of speech signals, which is of importance to enhance the quality and intelligibility of the telephone speech. The low-pass filtering effect of the telephone communication channels eliminate the high-frequency components of the speech signal, and it is necessary to retrieve those to maintain the speech quality. We adopt a joint-dictionary training approach to recover the missing spectral information. By exploiting the sparsity of the spectrogram frames, the dictionaries for the wide-band (WB) and the corresponding narrow-band (NB) spectrogram frames are trained in a coupled manner in order to learn the mapping from NB to WB frames. We refer to this approach as the *joint dictionary training for bandwidth extension* (JDTBE). To ensure that the reconstructed bandwidth-extended speech is consistent with the measurement, we propose to apply a suitable affine transformation that depends on the properties of the telephone channel. We study the effect of the choice of sparsity on the quality of the reconstructed speech, for both male and female speakers. A comparison of the proposed JDTBE algorithm with a bandwidth extension technique based on stochastic modeling reveals the superiority of the JDTBE approach in terms of subjective listening test scores.

Index Terms— Joint dictionary learning, sparsity, bandwidth extension of speech, consistency criterion, K-SVD algorithm.

1. INTRODUCTION

Due to the limitations of acquisition and transmission systems, speech acquired or transmitted is usually limited to a particular band of frequencies. Constraints in bandwidth lead to loss of perceptual quality and intelligibility of speech. A typical example of this phenomenon can be observed in telephone quality speech, which is limited to the frequency band $0.3 - 3.4$ kHz. Absence of the high frequency components in telephone quality speech leads to a muffled sound, resulting in poor intelligibility. It is observed from the listening experiments that acoustic bandwidth significantly affects the perceived speech quality. The speech signals having a bandwidth of 50 Hz to 7 kHz obtain a mean opinion score higher than that obtained by telephone speech [1], almost by a margin of 1.3. Also, the intelligibility of the non-contextual syllables is reduced by approximately 90% [2] due to the action of the telephone channel. Therefore, it is important to retrieve the missing information in the higher frequency band to create perceptually pleasant and intelligible speech. We propose a joint dictionary training-based approach to recover the missing information in the narrowband (NB) speech. Before explaining the details of our technique, we provide an overview of the existing literature on bandwidth extension.

Literature review: Most of the existing algorithms for bandwidth extension are based on the source-filter model of human speech production. The vocal tract filter, which parameterizes the wideband (WB) spectral envelope, and the WB excitation signal are estimated from the observed NB signal. The effectiveness of estimation differs depending on the modeling accuracy of the joint probability density of the NB and WB features, along with the choice of those features. An analysis of the suitability of various features for bandwidth extension is available in [3, 4]. Various methods, such as spectral shifting [5], modulation [6], non-linear processing of NB excitation [7], etc., have been proposed to generate the WB excitation signal, which is used as the input to the estimated WB filter to reconstruct WB speech. To estimate the WB excitation signal, a method based on spectral mirroring and data-driven voice source modeling [8] have been proposed in [9].

Gaussian mixture model (GMM)-based approaches for modeling the joint distribution of the WB and NB features have been discussed in [10–12], where the spectral envelope parameters of the WB speech are estimated from the NB features using a Bayesian minimum mean-square error estimate. The idea of using a codebook to recover the WB spectral information has also been proposed in the literature [13–16]. Another popular technique to model the joint distribution of the features in order to retrieve the missing spectral components is the *hidden Markov model* (HMM) [7, 17–19]. Pulakka et al. [20] employed a neural network to estimate the WB mel-spectrum from the observed NB features. Katsir et al. [21] developed a bandwidth extension method that exploits the phonetic content, where each speech frame is classified into a specific phoneme using a HMM-based statistical model. A technique based on probabilistic mapping on subspaces was developed in [22], where one utilizes the inherent sparsity of the state map to generate the bases of the target subspace. An *expectation maximization* (EM) algorithm-based technique was deployed in [23] to obtain the joint statistics of the WB and the NB features.

Pulakka et al. [24] proposed a bandwidth extension technique by first performing spectral folding, followed by modifying the high-frequency magnitude spectra using spline curves, where the spline control points are determined by using the NB features and sound classification information. Some other notable algorithms for bandwidth extension of speech include temporal envelope model [25], non-negative matrix factorization [26], etc.

Our contribution: We propose a joint dictionary training approach for bandwidth extension (JDTBE) of speech, where dictionaries for the WB and the corresponding NB spectrogram frames are trained in a coupled manner, over a corpus of speech signals, to learn the joint sparse representation of the NB and the WB frames. Our method leverages the sparsity of the spectrogram frames in the trained dictionaries while recovering the missing spectral information. A detailed description of the proposed technique is provided in the following sections.

2. PROBLEM FORMULATION AND PROPOSED METHOD

Let s_{nb} denote the observed NB speech signal, which can be modeled as a linearly degraded version of the actual WB speech s_{wb}. Given the NB speech signal s_{nb}, which contains only the low frequency information of the actual WB signal, our objective is to recover the high-frequency components, and thus reconstruct the WB speech s_{wb}. The problem of bandwidth extension can be thought of as an inverse problem, where one wants to extrapolate a signal to higher frequencies from its low-frequency measurement. We next explain the two stages of the proposed JDTBE algorithm.

1) Training: It has been reported in the speech processing literature that the spectrogram frames of a speech signal admit a sparse representation in an appropriately chosen basis [27]. We exploit the sparse structure of the spectrogram frames to solve the inverse problem of bandwidth extension. The appropriate dictionary that yields the sparse representation is trained on a corpus of speech spectrograms. In the training phase, we train two dictionaries jointly on the concatenated database of WB and the corresponding NB spectrogram frames. This objective is accomplished by solving the optimization problem:

$$\min_{D_{wb},D_{nb},\mathbf{x}_i} \sum_{i=1}^{N} \left\| \begin{bmatrix} \mathbf{y}_{wb}^{(i)} \\ \mathbf{y}_{nb}^{(i)} \end{bmatrix} - \begin{bmatrix} D_{wb} \\ D_{nb} \end{bmatrix} \mathbf{x}_i \right\|_2^2 \text{ s.t. } \|\mathbf{x}_i\|_0 \leq s, \forall i, \tag{1}$$

where $\mathbf{y}_{wb}^{(i)}$ and $\mathbf{y}_{nb}^{(i)}$ denote the WB and the corresponding NB frames, respectively in the spectrograms used for training, with i indicating the frame index. The symbol s denotes the sparsity level of the spectrogram patches. The dictionaries D_{wb} and D_{nb} capture the sparsity of the WB and the corresponding NB frames using the same sparse coefficient vector $\mathbf{x}_i$ for each frame i in the spectrogram. The joint dictionary training problem in (1) is solved via an alternating minimization approach, where one starts with initial guesses for the dictionaries D_{wb} and D_{nb}. Subsequently, one alternates between the following two steps: (i) the sparse coefficient vectors $\mathbf{x}_i$s are updated for fixed dictionaries using the orthogonal matching pursuit (OMP) algorithm [28], and (ii) the dictionaries are updated using the K-SVD algorithm [29]. This process constitutes the joint dictionary training step of the JDTBE approach.

2) Reconstruction of WB speech: To extend the bandwidth of an observed NB speech signal, it is first transformed to the spectrogram domain. Then, for each spectrogram frame $\mathbf{y}_{nb}^{(j)}$ in the observed NB speech, the following sparse coding problem is solved, using the trained NB dictionary D_{nb}:

$$\hat{\mathbf{x}}_0^{(j)} = \arg\min_{\mathbf{z}} \left\| \mathbf{y}_{nb}^{(j)} - D_{nb}\mathbf{z} \right\|_2^2 \text{ subject to } \|\mathbf{z}\|_0 \leq s. \tag{2}$$

The estimate of the corresponding WB frame is obtained by $\mathbf{y}_{wb}^{(j)} = D_{wb}\hat{\mathbf{x}}_0^{(j)}$, using the trained dictionary D_{wb} for the WB frames. This process is repeated for every frame j in the measured NB spectrogram to reconstruct the corresponding WB speech.

2.1. Motivation for joint training

The fundamental assumption behind the JDTBE approach is that the spectrogram frames admit a sparse representation in an appropriate basis. Let $s_{wb}^{(i)}$ denote the i^{th} segment of WB speech signal, multiplied by a window. The effect of the telephone channel can be modeled using a low-pass filter $h(n)$ followed by a downsampler $\mathcal{D}$. The output of the low-pass filter in the time-domain is given by $s_{wb}^{(i)}(n) * h(n)$. Subsequently, the down-sampler acts on the output of $h(\cdot)$ and the resulting NB frame can be written as $s_{nb}^{(i)}(n) = s_{wb}^{(i)}(m) * h(m)\big|_{m=2n}$. In the frequency domain, the relation between the WB frames and their NB counterparts is given as $S_{nb}^{(i)}(\omega) = \frac{1}{2}H\left(\frac{\omega}{2}\right) S_{wb}^{(i)}\left(\frac{\omega}{2}\right)$. We train the dictionaries in the spectrogram domain, which is computed by taking the modulus of the short-time-Fourier-transform (STFT), and given by $\mathbf{y}_{nb}^{(i)}(\omega) = \frac{1}{2}\left|H\left(\frac{\omega}{2}\right)\right| \mathbf{y}_{wb}^{(i)}(\frac{\omega}{2})$. The relation between the WB spectrogram frames $\mathbf{y}_{wb}^{(i)}$ and their corresponding NB counterparts $\mathbf{y}_{nb}^{(i)}$ can be expressed as a linear operation, which we write as $\mathbf{y}_{nb}^{(i)} = A\mathbf{y}_{wb}^{(i)}$. Using the assumption that the WB frames in the spectrogram domain admit a sparse representation in a dictionary D_{wb}, one can write $\mathbf{y}_{nb}^{(i)} = AD_{wb}\mathbf{x}_i$, where $\mathbf{x}_i$ is the corresponding sparse coefficient vector. Therefore, we note that the i^{th} WB spectrogram frame and its corresponding NB frame can be represented using the same sparse coefficient $\mathbf{x}_i$. This observation forms the basis of the joint training approach in (1), where the objective is to learn a sparse representation of the WB and NB frames using two different dictionaries and an identical sparse coefficient vector. The assumption of common sparse representation of the WB and the NB frames is further exploited during the bandwidth extension stage using the trained dictionaries.

2.2. Consistency Criterion

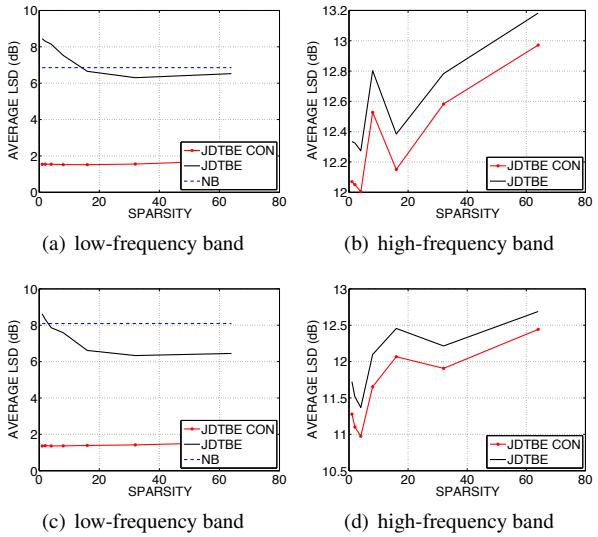
The NB part of the speech spectrogram contains considerable amount of information, which should be preserved in the process of extrapolation to the high-frequency components. Failure to accomplish this may lead to a loss of intelligibility in the enhanced speech. In order to ensure that the NB spectrogram is not tampered, one must enforce the condition that the reconstructed WB speech $\hat{s}_{wb}$ is consistent with the observed NB speech s_{nb}, that is, $\hat{s}_{wb}$ should exactly match with s_{nb} when subjected to low-pass filtering, followed by the downsampling effect of the telephone channel. Denoting the combined linear degradation operation of low-pass filtering and downsampling by $\mathcal{P}$, we want to enforce $\mathcal{P}\hat{s}_{wb} = s_{nb}$. This is achieved by solving the quadratic minimization problem with the desired linear equality constraint, given by

$$\hat{s}_{wb}^{(c)} = \arg\min_{s} \frac{1}{2}\|\hat{s}_{wb} - s\|_2^2 \text{ subject to } \mathcal{P}s = s_{nb}, \tag{3}$$

where $\hat{s}_{wb}^{(c)}$ is the reconstructed WB speech that satisfies the consistency criterion. The superscript c is used to emphasize that the reconstructed signal is consistent with the observed NB signal. The optimization problem in (3) can be solved in closed form and the solution is given by

$$\hat{s}_{wb}^{(c)} = \hat{s}_{wb} - \mathcal{P}^T\left(\mathcal{P}\mathcal{P}^T\right)^{-1}\left(\mathcal{P}\hat{s}_{wb} - s_{nb}\right). \tag{4}$$

The expression for final reconstructed WB speech $\hat{s}_{wb}^{(c)}$ in (4) is a combination of the dictionary based reconstruction $\hat{s}_{wb}$ and a correction term, which takes value zero if $\hat{s}_{wb}$ is consistent with the observed NB speech s_{nb}. The imposition of consistency can be interpreted as a projection operation, where one computes the final enhanced signal in such a way that it is closest to the one obtained using the trained dictionary and its NB spectrogram exactly matches the measurement. We refer to the approach of enforcing consistency, following the dictionary-based reconstruction as JDTBE-CON. Experimental results indicate that the imposition of the consistency criterion significantly improves the quality of the reconstructed speech.



(a) low-frequency band (b) high-frequency band

(c) low-frequency band (d) high-frequency band

Fig. 1. Performance of the JDTBE algorithm for dictionaries trained with different sparsity levels. The top and the bottom rows correspond to female and male speakers, respectively.

3. EXPERIMENTAL RESULTS

3.1. Implementation details

The experiments are conducted using speech files from the TIMIT database. In the simulations, the WB signals, containing components up to 8 kHz frequency, are filtered using a telephone channel filter to obtain the NB signal, having frequency components from 0.3 kHz up to 3.4 kHz. For computing the spectrogram, we consider frame length of 32 ms, with 50% overlap between adjacent frames, and a Hamming window. Discrete Fourier transforms (DFT) of 512 and 256 points are calculated to obtain the WB and the corresponding NB spectrograms, respectively. The frames of the WB and NB spectrograms, stacked on top of each other, are used as the feature vectors for learning the dictionaries. Since the DFT coefficients are conjugate symmetric, redundant points are removed while training. Consequently, the joint dictionary $D_{\text{joint}} = \begin{bmatrix} D_{wb} \\ D_{nb} \end{bmatrix}$ has atoms of size $1 + \frac{512}{2} + 1 + \frac{256}{2} = 386$, stacked as columns. We fix the number of atoms in the joint dictionary to be twice the atom dimension, so that D_{joint} is of size 386×772. We use $40,000$ spectrogram frames computed using the utterances of 10 different speakers for training. Separate dictionaries are trained for male and female speakers. Once the training stage is over, the top 257×772 and the bottom 129×772 blocks in D_{joint} are saved as the dictionaries for the WB and NB frames, respectively. While using the trained dictionaries for bandwidth extension, speech files and speakers different from the ones used for training are considered. We combine the replica of the observed NB phase with the spectral magnitude reconstructed in the high-frequency band using the dictionaries.

3.2. Performance of the JDTBE approach

To assess the quality of the enhanced speech obtained using the JDTBE method, we use the log-spectral distortion (LSD), averaged over 10 speech files, as the performance metric. We denote the LSD by d_{LS}, and it is defined as

$$d_{\text{LS}} = \left(\frac{1}{\omega_h - \omega_l} \int_{\omega_l}^{\omega_h} \left(20 \log \left(\frac{|P_y(\omega)|}{\left|\hat{P}_y(\omega)\right|} \right) \right)^2 \mathrm{d}\omega \right)^{\frac{1}{2}},$$

where $P_y(\omega)$ and $\hat{P}_y(\omega)$ denote the spectra of the original and the estimated WB speech. The frequencies ω_l and ω_h should be chosen depending on which part (low or high-frequency band) of the speech spectrum we want to compare against the corresponding part of the original WB speech. For example, to measure the proximity of the lower-band of the enhanced speech with that of the original, one must set $\omega_l = 0$ and $\omega_h = \frac{\pi}{2}$.

In Fig. 1, we show the average LSD values, obtained using the JDTBE approach for both male and female speakers, as a function of the sparsity level s. From Figures 1(a) and 1(c), we observe that the LSD over the low-frequency band decreases with increase in s. However, it almost saturates as s increases beyond 32. The LSD value over the low-frequency band for the measured NB signal is shown in dashed lines for facilitating a comparison. We observe that the low-frequency LSD of the reconstructed WB speech goes below that of the measured NB speech, as s exceeds 16 and 4, corresponding to female and male speakers, respectively. The reason behind the reduction of the LSD over the low-frequency band is due to the ability of the JDTBE approach to reconstruct missing spectral components in the range $0 - 0.3$ kHz and 3.4 to 4 kHz. Increasing s has the effect of making more atoms in the dictionary available for representing the NB spectrogram frames, thereby reducing the error over the low-frequency band. However, further increasing s results in marginal reduction in error, and consequently the LSD saturates. From Figures 1(b) and 1(d), we observe that, for both male and female speakers, the LSD over the high-frequency band attains a minimum for a sparsity level of $s = 4$, and it increases for higher values of s. The reason behind this phenomenon is explained as follows. In the bandwidth extension stage, one only has access to the observed NB spectra. By increasing s in (2), one allows more atoms to be chosen from D_{nb} to represent the NB spectrogram frames, thereby reducing the discrepancy over the low-frequency part of the spectra. However, since the same coefficients obtained from (2) are combined with D_{wb} to obtain the WB signal, increasing s beyond a point can potentially introduce error in the higher-frequency band owing to the use of unnecessary atoms from D_{wb}. We find experimentally that $s = 4$ yields optimum reconstruction of the missing high-frequency band. We also observe from the Figure 1 that imposing the consistency criterion further reduces the LSD values over both high- and low-frequency bands.

The spectrograms of the bandwidth-extended speech, using the JDTBE approach, without and with the consistency criterion imposed, are shown in Figures 2(c) and (d), respectively. The spectrograms of the input NB and the actual WB speech are also shown to facilitate visual comparison. The sparsity level s chosen for training as well as reconstruction is 4. We observe that the JDTBE algorithm recovers the missing high-frequency part of the input NB spectrogram reasonably well. After imposing consistency, we note that the spectrogram in the lower-band of the reconstructed signal closely matches that of the original WB signal. One can observe from the spectrogram of reconstructed signal that, in the regions $0 - 0.3$ kHz and $3.4 - 4$ kHz, the JDTBE-CON approach does a better job of recovering the spectral content, which is absent in the observed NB signal. This fact is also reflected in the LSD values of the reconstructed speech obtained after imposing the consistency criterion.

The boxes in the spectrograms are used to highlight a particular region where the reconstruction obtained using the JDTBE-CON approach matches the corresponding part of the WB spectrogram more closely than JDTBE.

3.3. Description of the listening test

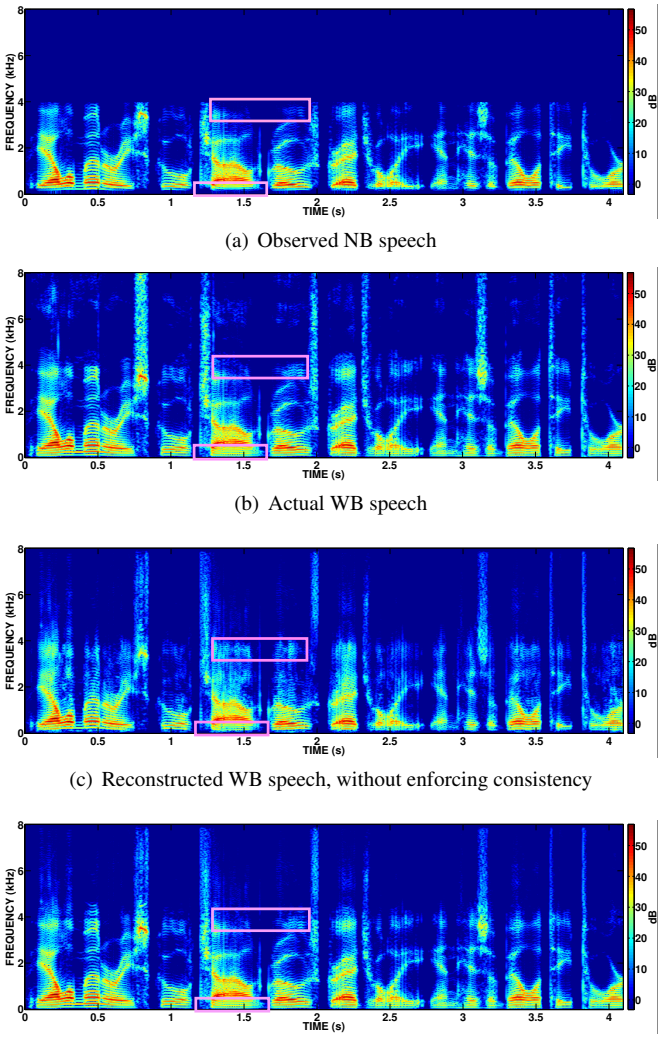
We performed a listening test using a high quality headphone (*Sennheiser HD* 215) to assess the perceptual quality of the bandwidth-extended speech using the JDTBE and the method proposed by Qian et al. [11]. A pairwise listening test is performed, where the listeners are asked to rate two speech signals A and B in the scale of -3 to $+3$, in steps of 1. The scores reflect the quality of B compared with A and their meanings are as follows: $+3$: much better, $+2$: better, $+1$: slightly better, 0 : about the same, -1 : slightly worse, -2 : worse, -3 : much worse. During the test, listeners were allowed to listen to the signals repeatedly, if needed. They were also given the flexibility to adjust the sound level to a comfortable value. We made use of 10 male and 10 female speech files in the experiment, and five listeners participated in the test. The final scores are calculated by averaging over all the listeners and the files. The speech files used for subjective evaluation are available online[1]. In Table 1, we show

A↔B (B compared with A)	Male	Female
NB↔JDTBE (4)	0.04	−1.34
NB↔JDTBE-CON (4)	2.02	1.79
NB↔JDTBE (64)	0.87	0.74
NB↔JDTBE-CON (64)	1.98	1.33
NB↔WB	2.7	2.72
NB↔SMBE	0.8	0.93
JDTBE(4)↔JDTBE-CON (4)	1.95	1.76
JDTBE (64)↔JDTBE-CON (64)	0.81	0.26
JDTBE-CON(64)↔JDTBE-CON (4)	0.14	0.04
JDTBE-CON(4)↔SMBE	−1.32	−0.82
WB↔WB	0.06	0.04
WB↔JDTBE-CON (64)	−0.5	−0.63

Table 1. Comparison in terms of listening test scores. The numbers inside the parentheses indicate the corresponding sparsity level s.

the listening test scores of the bandwidth-extended speech using the JDTBE and JDTBE-CON algorithms for two different values of s, in comparison with the NB, original WB, and the reconstructed WB speech obtained using the method proposed in [11], which we refer to as *stochastic-modeling-based bandwidth extension* (SMBE) in the table. We report the scores corresponding to two sparsity levels, $s = 4$ and $s = 64$. For both sparsity levels, the JDBTBE-CON achieves higher listening test scores, for both male and female speakers, as compared with the observed NB speech and the reconstructed WB speech using the SMBE technique. The JDTBE algorithm, for $s = 64$, is rated on par with the SMBE technique by the listeners. The score obtained by JDTBE for $s = 4$ is slightly inferior compared with that obtained by the measured NB input. We observe that the scores achieved with the JDBTE-CON approach are consistently better than that obtained with the JDBTE algorithm. However, as s increases, the improvement obtained by applying the consistency criterion tends to become marginal. Among the techniques compared, the JDBTE-CON approach with $s = 4$ is observed to perform better than the competing techniques in terms of the listening test scores.

[1] http://spectrumee.wix.com/abwe.



(a) Observed NB speech

(b) Actual WB speech

(c) Reconstructed WB speech, without enforcing consistency

(d) Reconstructed WB speech, with consistency (JDTBE-CON)

Fig. 2. (Color online) Spectrograms of the reconstructed WB speech, with and without enforcing the consistency criterion, for $s = 4$.

4. CONCLUSIONS

We have proposed a bandwidth extension method for speech signals based on a joint dictionary training approach that recovers the missing spectral information in telephone channel speech. The WB and the corresponding NB spectrogram frames are used as features for training the dictionaries. The central idea behind the technique is to leverage the sparsity of the spectrogram frames in the trained dictionaries. We have demonstrated that the reconstruction quality can be further improved by imposing the consistency criterion, that is, by enforcing the low-frequency spectra of the reconstructed speech to exactly match that of the measured NB signal. Performance evaluation in terms of subjective listening test scores indicates the superiority of the proposed algorithm to a competing technique based on stochastic modeling proposed in [11]. The idea of joint dictionary learning could potentially find application to a more general class of problems, such as recovering missing patches in spectrograms, removing the artifacts introduced by denoising algorithms, post-processing of speech generated by *text-to-speech* converters to improve the naturalness, etc., and warrants further investigation.

5. REFERENCES

[1] S. Voran, "Listener rating of speech passbands," in Proc. *IEEE Workshop on Speech Coding*, pp. 81–82, Sep. 1997.

[2] I. Katsir, "Artificial bandwidth extension of bandlimited speech based on vocal tract shape estimation," *Master of Science in Electrical Engineering Thesis*, Israel Institute of Technology, Dec. 2011.

[3] M. Nilsson, H. Gustafsson, S. V. Andersen, and W. B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," in Proc. *IEEE Intl. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 525–528, 2002.

[4] P. Jax and P. Vary, "Feature selection for improved bandwidth of speech signals," in Proc. *IEEE Intl. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 697–700, 2004.

[5] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in Proc. *IEEE Intl. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 428–43, 1979.

[6] U. Kornagel, "Specral widening of excitation for telephone-band speech enhancement," in Proc. *Intl. Workshop on Acoust., Echo, and Noise Control.*, pp. 215–218, Sep. 2001.

[7] P. Jax and P. Vary, "On bandwidth extension of telephone speech," *Elsevier Signal Process.*, vol. 83, pp. 1707–1719, Aug. 2003.

[8] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Data-driven voice source waveform modeling," in Proc. *IEEE Intl. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 3965–3968, 2009.

[9] M. R. P. Thomas, J. Gudnason, P. A. Naylor, B. Gieser, and P. Vary, "Voice source estimation for artificial bandwidth extension of telephone speech," in Proc. *IEEE Intl. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 4794–4797, 2010.

[10] K. Park and H. S. Kim, "Narrowband to wideband conversion using GMM based transformation," in Proc. *IEEE Intl. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 1843–1846, 2000.

[11] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech," in Proc. *IEEE Intl. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 713–716, 2004.

[12] H. Pulakka, U. Remes, K. Palomaki, M. Kurimo, and P. Alku, "Speech bandwidth extension using Gaussian mixture model-based estimation of the highband mel spectrum," in Proc. *IEEE Intl. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 5100–5103, 2011.

[13] J. A. Fuemmeler, R.C. Hardie, and W. R. Gardner, "Techniques for the regeneration of wideband speech from narrowband speech," *EURASIP Journal on Applied Sig. Proc.*, pp. 1–9, 2001.

[14] Y. Qian and P. Kabal, "Wideband speech recovery from narrowband speech using classified codebook mapping," in Proc. 9^{th} *Australian International Conference on Speech Science and Technology.*, pp. 106–111, Dec. 2002.

[15] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," in Proc. *IEEE Intl. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 805–808, 2005.

[16] U. Kornagel, "Techniques for artificial bandwidth extension of telephone speech," *Elsevier Signal Process.*, vol. 86, no. 6, pp. 1296–1306, 2006.

[17] G. Chen and V. Parsa, "HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies," in Proc. *IEEE Intl. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 709–712, 2004.

[18] G. B. Song and P. Martynovich, "A study of HMM-based bandwidth extension of speech signals," *Elsevier Signal Process.*, vol. 89, no. 10, pp. 2036–2044, 2009.

[19] C. Yagli, M. A. T. Turan, and E. Erzin, "Artificial bandwidth extension of spectral envelope along a Viterbi path," *Journal of Speech Communication*, vol. 55, pp. 111–118, 2013.

[20] H. Pulakka and P. Alku, "Bandwidth extension of telephone speech using a neural network and a filterbank implementation for highband mel spectrum," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 7, pp. 2170–2183, Sep. 2011.

[21] I. Katsir, I. Cohen, and D. Malah, "Speech bandwidth extension based on speech phonetic content and speaker vocal tract shape estimation," *European Sig. Proc. Conf. (EUSIPCO)*, pp. 461–465, Sep. 2011.

[22] K. Kalgoankar and M. A. Clements, "Sparse probabilistic state mapping and its application to speech bandwidth expansion," in Proc. *IEEE Intl. Conf. on Acoust., Speech, and Sig. Proc.*, pp. 4005–4008, 2009.

[23] Y. M. Cheng, D. O'shaunghnessy, and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Trans. on Sig. Proc.*, vol. 2, no. 4, pp. 544–548, Oct. 1994.

[24] H. Pulakka, L. Laksonen, M. Vainio, J. Pohjalainen, and P. Alku, "Evaluation of an artificial speech bandwidth extension method in three languages," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 16, no. 6, pp. 1124–1136, Aug. 2009.

[25] K. T. Kim, M. K. Lee, and H. G. Kang, "Speech bandwidth extension using temporal envelope modeling," *Elsevier Signal Process.*, vol. 83, pp. 1707–1719, 2003.

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

[27] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 20, no. 6, pp. 1698−1712, Aug. 2012.

[28] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. on Info. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[29] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Sig. Proc.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.