LANGUAGE INFORMED BANDWIDTH EXPANSION

Jinyu Han[*]
EECS department
Northwestern University

Gautham J. Mysore
Advanced Technology Labs
Adobe Systems Inc.

Bryan Pardo
EECS department
Northwestern University

ABSTRACT

High-level knowledge of language helps the human auditory system understand speech with missing information such as missing frequency bands. The automatic speech recognition community has shown that the use of this knowledge in the form of language models is crucial to obtaining high quality recognition results. In this paper, we apply this idea to the bandwidth expansion problem to automatically estimate missing frequency bands of speech. Specifically, we use language models to constrain the recently proposed non-negative hidden Markov model for this application. We compare the proposed method to a bandwidth expansion algorithm based on non-negative spectrogram factorization and show improved results on two standard signal quality metrics.

***Index Terms*—** Non-negative Hidden Markov Model, Language Model, Bandwidth Expansion

1. INTRODUCTION

Audio Bandwidth Expansion (BWE) refers to methods that increase the frequency bandwidth of narrowband audio signals. Such frequency expansion is desirable if at some point the bandwidth of the signal has been reduced, as can happen during signal recording, transmission, storage, or reproduction.

A typical application of BWE is telephone speech enhancement [1]. The degradation of speech quality is caused by the bandlimiting filters with a passband from approximately 300 Hz to 3400 Hz, due to the use of analogue frequency-division multiplex transmission. Other applications include bass enhancement on small loudspeakers and high-quality reproduction of historical recordings.

Most BWE methods are based on the source-filter model of speech production [1]. Such methods generate an excitation signal and modify it with an estimated spectral envelope that simulates the characteristics of the vocal tract. The main focus has been on the spectral envelope estimation. Classical techniques for spectral envelope estimation include Gaussian mixture models (GMM) [2], hidden Markov models (HMM) [3], and neural networks [4]. However, these methods need to be trained on parallel wideband and narrowband corpora to learn a specific mapping between narrowband features and wideband spectral envelopes. Thus, a system trained on telephony and wideband speech cannot be readily applied to expand the bandwidth of a low-quality loudspeaker.

Another way to estimate the missing frequency bands is based on directly modeling the audio signal by learning a dictionary of spectral vectors that explains the audio spectrogram. By directly modeling the audio spectrogram, BWE can be framed as a missing data imputation problem. Such methods only need to be trained once on wideband corpora. Once the system is trained, it can be used to expand any missing frequencies of narrowband signals, despite never having been trained on the mapping between the narrowband and wideband corpus. To the best of our knowledge, the only existing work based on directly modeling the audio is [5] using non-negative spectrogram factorization.

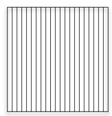
In this paper, we show that the performance of BWE can be improved by introducing speech recognition machinery. Specifically, if it is known that the given speech conforms to certain syntactic constraints, this high level information could be useful to constrain the model. In automatic speech recognition (ASR), such constraints are typically enforced in the form of a language model (constrained sequences of words) [6]. It has more recently been applied to source separation [7]. However, we are not aware of any existing BWE methods that explicitly explore syntactic knowledge about speech. Note that there has recently been an approach [3] that used language information to improve the performance of source-filter models for BWE. However, this approach requires an a-priori transcription of the given speech. In contrast, our technique does not require any information about the content of the specific instance of speech but rather uses syntactical constraints in the form of a language model.

2. MODEL OF AUDIO

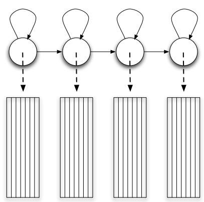
Non-negative spectrogram factorization refers to a class of techniques that include non-negative matrix factorization (NMF) [8] and its probabilistic counterparts such as probabilistic latent component analysis (PLCA) [9]. In this section,

[*]This work was supported in part by National Science Foundation award 0812314.

we first describe this with respect to PLCA because that is the technique used in [5]. We then describe the non-negative hidden Markov model (N-HMM) [10] and explain how it overcomes some of the limitations of non-negative spectrogram factorization.



(a) Probabilistic Latent Component Analysis



(b) Non-negative Hidden Markov Model with left-to-right transition model

Fig. 1. Comparison of non-negative models. Each column here represents a spectral component. PLCA uses a single large dictionary to explain a sound source, whereas the N-HMM uses multiple small dictionaries and a Markov chain.

PLCA models each time frame of a given audio spectrogram as a linear combination of spectral components. The model is as follows:

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

where $P_t(f)$ is approximately equal to the normalized spectrogram at time t, $P(f|z)$ are spectral components (analogous to dictionary elements), and $P_t(z)$ is a distribution of mixture weights at time frame t. All distributions are discrete. Given a spectrogram, the parameters of PLCA can be estimated using the expectation–maximization (EM) algorithm [9].

PLCA (Fig. 1(a)) uses a single dictionary of spectral components to model a given sound source. Specifically, each time frame of the spectrogram is explained by a linear combination of spectral components from the dictionary. For BWE, the dictionary learned by PLCA on wideband audio is used to reconstruct the missing frequencies of the narrowband audio.

Audio is non-stationary as the statistics of its spectrum change over time. However, there is a structure in this non-stationarity in the form of temporal dynamics. The dictionary learned by PLCA ignores these important aspects of audio: non-stationarity and temporal dynamics. To overcome these issues, the N-HMM [10] was recently proposed (Fig. 1(b)). This model uses multiple dictionaries such that each time frame of the spectrogram is explained by any one of the several dictionaries (accounting for non-stationarity). Additionally it uses a Markov chain to explain the transitions between its dictionaries (accounting for temporal dynamics).

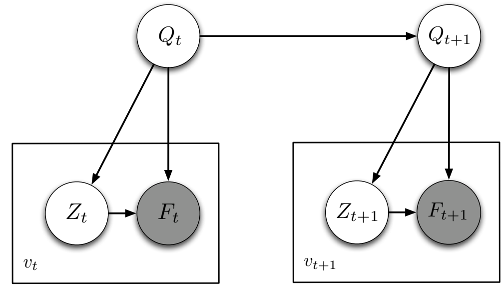


Fig. 2. Graphical Model of the N-HMM. {Q, Z, F} is a set of random variables, {q, z, f} their realization and t is the index of time. Shaded variable indicates observed data. v_t represents the number of draws at time t.

Fig. 2 shows the graphical model of the N-HMM. Each state q in the N-HMM corresponds to a dictionary. Each dictionary contains a number of spectral components indexed by z. Therefore, spectral component z of dictionary q is represented by $P(f|z,q)$. The observation model at time t, which corresponds to a linear combination of the spectral components from dictionary q, is given by:

$$P_t(f_t|q_t) = \sum_{z_t} P_t(z_t|q_t) P(f_t|z_t, q_t), \tag{2}$$

where $P_t(z_t|q_t)$ is a distribution of mixture weights at time t. The transitions between states are modeled with a Markov chain, given by $P(q_{t+1}|q_t)$. All distributions are discrete. Given a spectrogram, the N-HMM model parameters can be estimated using an EM algorithm [7].

We can then reconstruct each time frame as follows:

$$P_t(f) = \sum_{q_t} P_t(f_t|q_t) \gamma_t(q_t), \tag{3}$$

where $\gamma_t(q_t)$ is the posterior distribution over the states, conditioned on all the observations over all time frames. We compute $\gamma_t(q_t)$ using the forward-backward algorithm [6] as in HMMs when performing the EM iterations. Note that in practice $\gamma_t(q_t)$ tends to have a probability of nearly 1 for one of the dictionaries and 0 for all other dictionaries so there is usually effectively only one active dictionary per time frame.

3. SYSTEM OVERIVEW

A block diagram of the proposed system is shown in Fig. 3. The goal is to learn an N-HMM for each speaker from training data of that speaker and syntactic knowledge common to all speakers (in the form of a language model). We construct each speaker-level N-HMM in two steps. We first learn a N-HMM for each word in the vocabulary, detailed in Sec. 4. We then build a language model by concatenating all the word models together according to the word transitions specified by the language model, as elaborated in Sec. 5. Given the narrowband speech, the learned speaker-level

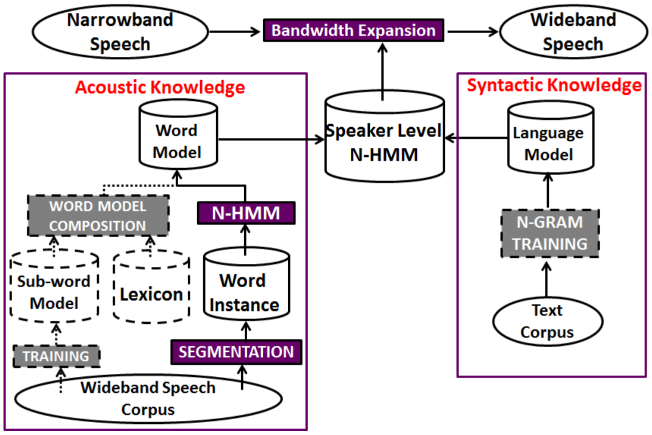


Fig. 3. Block diagram of the proposed system. Our current implementation includes modules with solid lines. Modules with dashed lines indicate possible extensions in order to make the system more feasible for large vocabulary BWE.

N-HMM can be utilized to perform bandwidth expansion by estimating the missing frequencies as an audio spectrogram imputation problem. This is described in Sec. 6.

Learning a word model for each word in a vocabulary is suitable for small vocabularies. However, it is not likely to be feasible for larger vocabularies. In this paper we are simply establishing that the use of a language model does improve BWE, rather than selecting the most scalable modeling strategy for large vocabulary situations. This work can be extended to use subword models such as phonelike units (PLUs) [6], which have been quite successful in ASR. In Fig. 3, we illustrate these extensions using dashed lines.

4. WORD MODELS

For each word in our vocabulary, we learn the parameters of an N-HMM from multiple instances (recordings) of that word as routinely done with HMMs in small vocabulary speech recognition [6]. The N-HMM parameters are learned using the EM algorithm [7].

Let $V^{(k)}$, $k = 1 \cdots N$, be the k^{th} spectrogram instance of a given word. We compute the E step of EM algorithm separately for each instance. The procedure is the same as in [7]. This gives us the marginalized posterior distributions $P_t^{(k)}(z_t, q_t|f_t, \bar{\mathbf{f}})$ and $P_t^{(k)}(q_t, q_{t+1}|\bar{\mathbf{f}})$ for each word instance k. Here, $\bar{\mathbf{f}}$ denotes the observed magnitude spectrum across all time frames, which is the entire spectrogram $V^{(k)}$.

We use these marginalized posterior distributions in the M step of the EM algorithm. Specifically, we compute a separate weights distribution for each word instance k as follows:

$$P_t^{(k)}(z_t|q_t) = \frac{\sum_{f_t} V_{ft}^{(k)} P_t^{(k)}(z_t, q_t|f_t, \bar{\mathbf{f}})}{\sum_{z_t} \sum_{f_t} V_{ft}^{(k)} P_t^{(k)}(z_t, q_t|f_t, \bar{\mathbf{f}})}, \qquad (4)$$

where $V_{ft}^{(k)}$ is the magnitude (at time t and frequency f) of spectrogram $V^{(k)}$ of word instance k.

However, we estimate a single set of dictionaries of spectral components and a single transition matrix using the marginalized posterior distributions of all instances of a given word as follows:

$$P(f|z,q) = \frac{\sum_k \sum_t V_{ft}^{(k)} P_t^{(k)}(z,q|f,\bar{\mathbf{f}})}{\sum_f \sum_k \sum_t V_{ft}^{(k)} P_t^{(k)}(z,q|f,\bar{\mathbf{f}})} \qquad (5)$$

$$P(q_{t+1}|q_t) = \frac{\sum_k \sum_{t=1}^{T-1} P_t^{(k)}(q_t, q_{t+1}|\bar{\mathbf{f}})}{\sum_{q_{t+1}} \sum_k \sum_{t=1}^{T-1} P_t^{(k)}(q_t, q_{t+1}|\bar{\mathbf{f}})} \qquad (6)$$

The remaining parameters are estimated as described in [10]

Once we learn the set of dictionaries and transition matrix for each word of a given speaker, we need to combine them into a single speaker dependent N-HMM.

5. SPEAKER LEVEL MODEL

The goal of the language model is to provide an estimate of the probability of a word sequence W for a given task. If we assume that W is a specified sequence of words, i.e.,

$$W = w_1 w_2 ... w_Q, \qquad (7)$$

$P(W)$ can be computed as:

$$\begin{aligned} P(W) =& P(w_1 w_2 ... w_Q) \\ =& P(w_1)P(w_2|w_1)P(w_3|w_1 w_2)... \\ & P(w_Q|w_1 w_2 ... w_{Q-1}). \end{aligned} \qquad (8)$$

In practice, N-gram ($N = 2$ or 3) word models are used to approximate the term $P(w_j|w_1...w_{j-1})$ as:

$$P(w_j|w_1...w_{j-1}) \approx P(w_j|w_{j-N+1}...w_{j-1}) \qquad (9)$$

i.e., based only on the preceding $N-1$ words.

The conditional probabilities $P(w_j|w_{j-N+1}...w_{j-1})$ can be estimated by the relative frequency approach:

$$\hat{P}(w_j|w_{j-N+1}...w_{j-1}) = \frac{R(w_j, w_{j-1}, ..., w_{j-N+1})}{R(w_{j-1}, ..., w_{j-N+1})}, \qquad (10)$$

where $R(\cdot)$ is the number of occurrences of the string in its argument in the given training corpus.

In an N-HMM, we learn a Markov chain that explains the temporal dynamics between the dictionaries. Each dictionary corresponds to a state in the N-HMM. Since we use an HMM structure, we can readily use the idea of language model to constrain the Markov chain to explain a valid grammar.

Once we learn an N-HMM for each word of a given speaker, we combine them into a single speaker dependent N-HMM according to the language model. We do this by constructing a large transition matrix that consists of each individual word transition matrix. The transition matrix of each individual word stays the same as specified in Eq. 6. However, the language model dictates the transitions between words. In this paper, the syntax to which every sentence in the corpus conforms to is provided in [11]. However, when this is not the case, one can learn the language model as described above.

6. ESTIMATION OF INCOMPLETE DATA

So far, we have shown how to learn a speaker-level N-HMM that combines the acoustic knowledge of each word, and syntactic knowledge, in the form of language model, from wideband speech. With respect to wideband speech, we can consider narrowband speech as incomplete data since certain frequency bands are missing. We generally know the frequency range of narrowband speech. We therefore know which frequency bands are missing and consequently which entries of the spectrogram of narrowband speech are missing. Our objective is to estimate these entries. Intuitively, once we have a speaker-level N-HMM, we estimate the mixture weights for spectral component of each dictionary, as well as the expected values for the missing entries of the spectrogram.

We denote the observed regions of a spectrogram V as V^o and the missing regions as $V^m = V \setminus V^o$. Within a magnitude spectrum V_t at time t, we represent the set of observed entries as V_t^o and the missing entries as V_t^m. $\mathscr{F}_t^o$ will refer to the set of frequencies for which the values of V_t are known, i.e. the set of frequencies in V_t^o. $\mathscr{F}_t^m$ will similarly refer to the set of frequencies for which the values of V_t are missing, i.e. the set of frequencies in V_t^m. $V_t^o(f)$ and $V_t^m(f)$ will refer to specific frequency entries of V_t^o and V_t^m respectively. For narrowband telephone speech, we set $\mathscr{F}_t^o = \{f|300 \leq f \leq 3400\}$ and $\mathscr{F}_t^m = \{f|f < 300 \; or \; f > 3400\}$ for all t.

Our method is an N-HMM based imputation technique that works for the estimation of missing frequencies in the spectrogram, as described in our previous work [12].

In this method, we perform N-HMM parameter estimation on the narrowband spectrogram. However, the only parameters that we estimate are the mixture weights. We keep the dictionaries and transition matrix from the speaker level N-HMM fixed. One issue is that the dictionaries are learned on wideband speech (Sec. 4) but we are trying to fit them to narrowband speech. We therefore only consider the frequencies of the dictionaries that are present in the narrowband spectrogram: $\mathscr{F}_t^o$, for the purposes of mixture weights estimation. However, once we estimate the mixture weights, we reconstruct the wideband spectrogram using all of the frequencies of the dictionaries.

The resulting value $P_t(f)$ in Eq. 3. (the counterpart for PLCA is Eq. 1) can be viewed as an estimate of the relative magnitude of of the frequencies at time t. However, we need estimates of the absolute magnitudes of the missing frequencies so that they are consistent with the observed frequencies. We therefore need to estimate a scaling factor for $P_t(f)$. In order to do this, we sum the values of the uncorrupted frequencies in the original audio to get $n_t^o = \sum_{f \in \mathscr{F}_t^o} V_t^o(f)$. We then sum the values of $P_t(f)$ for $f \in \mathscr{F}_t^o$ to get $p_t^o = \sum_{f \in \mathscr{F}_t^o} P_t(f)$. The expected magnitude at time t is obtained by dividing n_t^o by p_t^o, which gives us a scaling factor. The expected value of any missing term $V_t^m(f)$ can then be estimated by:

$$E[V_t^m(f)] = \frac{n_t^o}{p_t^o} P_t(f) \tag{11}$$

The audio BWE process can be summarized as follows:

1. Learn an N-HMM word model for each word in the training data set using the EM algorithm, as described in Sec. 4 from the wideband speech corpus. We now have a set of dictionaries, each of which corresponds roughly to a phoneme in the training data. We call these *wideband dictionaries*.

2. Combine the word models into one single speaker dependent N-HMM model as described in Sec. 5.

3. Given the narrowband speech, construct the *narrowband dictionaries* by considering only the frequencies of the *wideband dictionaries* that are present in the narrowband spectrogram – $f \in \mathscr{F}^o$

4. Perform N-HMM parameter estimation on the narrowband spectrogram. Specifically, learn the mixture weights $P_t(z_t|q_t)$ and keep all of the other parameters fixed.

5. Calculate $P_t(f)$ as shown in Eq. 3 using the *wideband dictionaries* and the mixture weights estimated in step 4.

6. Reconstruct the corrupted audio spectrogram as follows:

$$\bar{V}_t(f) = \begin{cases} V_t(f) & \text{if } f \in \mathscr{F}_t^o \\ E[V_t^m(f)] & \text{if } f \in \mathscr{F}_t^m \end{cases} \tag{12}$$

7. Convert the estimated spectrogram to the time domain.

This paper does not address the problem of missing phase recovery. Instead we use the recovered spectrogram with the original phase to re-synthesize the time domain signal. We found this to be more perceptually pleasing than a standard phase recovery method [13].

7. EXPERIMENTAL RESULTS

We performed experiments on a subset of the speech separation challenge training data set [11]. We selected 10 speakers (5 male and 5 female), with 500 sentences per speaker. We

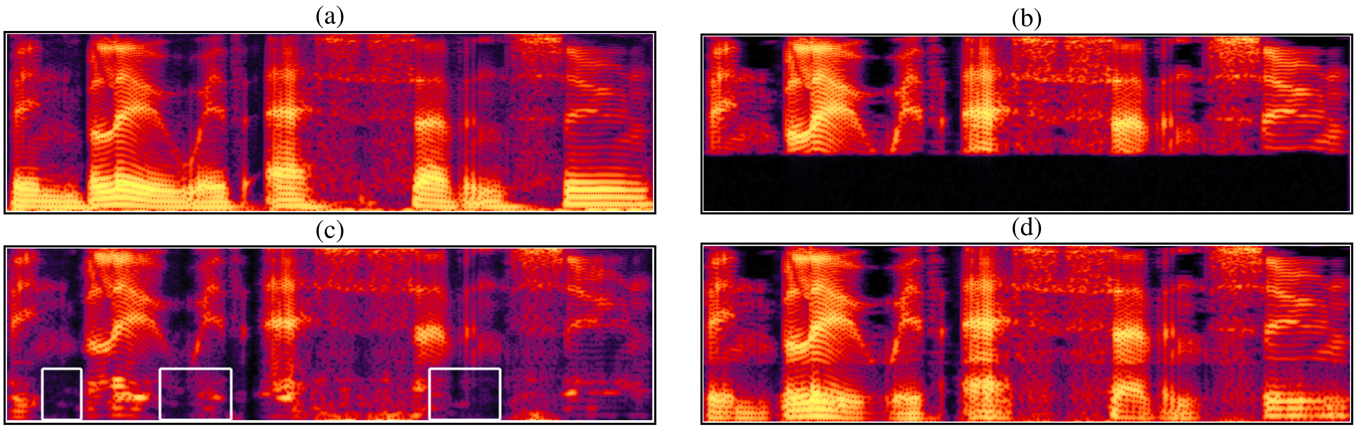


Fig. 4. Example of speech BWE. The x-axis is time and y-axis frequency. a) Original speech; b) Narrowband speech. c) Result using the PLCA; Regions marked with white-edge boxes are regions in which PLCA performed poorly. d) Result using the proposed method.

learned N-HMMs for each speaker using 450 of the 500 sentences, and used the the remaining 50 sentences as the test set.

We segmented the training sentences into words in order to learn individual word models as described in Sec. 4. We used one state per phoneme. This is less than what is typically used in speech recognition because we did not want to excessively constrain the model.We then combined the word models of a given speaker into a single N-HMM according to the language model, as described in Sec. 5 .

We performed speech BWE using the language-model constrained N-HMM on the 50 sentences per speaker in the test set, totaling 500 sentences. As a comparison, we performed BWE using PLCA [5] with a scaling factor calculated in Eq. 11 . When using PLCA, we used the same training and test sets that we used with the proposed model. However, we simply concatenated all of the training data of a given speaker and learned a single dictionary for that speaker, which is customary when using non-negative spectrogram factorizations.

We considered two different conditions. The first one is to expand the bandwidth of telephony speech signals, referred to as *Con-A*. The input narrowband signal has a bandwidth of 300 to 3400 Hz. In the second condition, referred to as *Con-B*, we removed all the frequencies below 1000 Hz. In *Con-B*, the speech is considerably more corrupted than the telephony speech since speech usually has strong energy distributed in the low frequencies. For both categories, we reconstructed wideband signals with frequencies up to 8000 Hz.

Signal-to-Noise-Ratio (SNR)[1] and overall rating [14] for speech enhancement (OVRL) are used to measure the narrowband speech and the outputs of both methods. In this context, SNR measures the "signal to difference" ratio between original and reconstructed signals. The higher the number, the closer the reconstructed signal is to the original one. OVRL is the predicted overall quality of speech using the scale of the Mean Opinion Score (1=bad, 2=poor, 3=fair, 4=good, 5=excellent).

We first illustrate the proposed method with an example in Fig. 4 . The original audio is a 2.2-second clip of speech of a male speaker saying, "bin blue with s seven soon". We removed the lower 1000 Hz of the spectrogram. The lower 4000 Hz are plotted in log-scale. Compared to PLCA, the proposed method provides a higher-quality reconstruction as can be clearly seen in the low frequencies. PLCA tends to be problematic with the reconstruction of low-end energy. We have marked with white-edge boxes the regions in Fi. 4(c) where PLCA performed poorly. The proposed method, on the other hand, has recovered most of the lower harmonics quite accurately. Sound Examples are available at `music.cs.northwestern.edu/research.php?project=imputation#Example_MLSP` to show the perceptual quality of the reconstructed signals.

The averaged performance across all 10 speakers is reported in Tab. 1 . The score for each speaker is averaged over all 50 sentences for that speaker. As shown, both methods produce results that have significantly better audio quality than the given narrowband speech. The proposed method, however, outperforms PLCA in both conditions using both metrics. The improvements of both metrics in both conditions are statistically significant between the proposed method and PLCA by *student t-test* with p-values smaller than 0.01.

In *Con-A*, PLCA has improved the speech quality (in terms of OVRL metric) of the input narrowband signals from "bad" to "between fair and good". The proposed method has further improved the rating to "above good". In *Con-B*, the OVRL metric of the corrupted speech signal is improved from "bad" to "above poor" by PLCA, and further to "between fair and good" by the proposed method. The improvement is clearly more apparent in *Con-B* than in *Con-A*. The reason

[1] $SNR = 10log_{10}\frac{\sum_t s(t)^2}{\sum_t(\bar{s}(t)-s(t))^2}$ where $s(t)$ and $s(\bar{t})$ are the original and the reconstructed signals respectively.

Con-A	Input	PLCA	Proposed
SNR (dB)	4.20	7.58	10.87
OVRL	1.15	3.58	4.26

Con-B	Input	PLCA	Proposed
SNR (dB)	0.17	1.43	5.90
OVRL	1.00	2.26	3.41

Table 1. Performance of audio BWE using the proposed method and PLCA.

is that *Con-B* is more heavily corrupted, so the spectral information alone is not enough to get reasonable results and the temporal information from the language model is able to boost the performance.

8. CONCLUSIONS AND FUTURE WORK

We presented a method to perform audio BWE using language models in the N-HMM framework. We have shown that the use of language models to constrain non-negative models has led to improved speech BWE performance when compared to a non-negative spectrogram factorization method. The main contribution of this paper is to show that the use of speech recognition machinery for the BWE problem is promising. In the proposed system, the acoustic knowledge of the word models and the syntactic knowledge in the form of a language model are incorporated to improve the results of BWE. The methodology was shown with respects to speech and language models, but it can be used in other contexts in which high-level structure information is available. One such example is incorporating music theory into the N-HMM framework for BWE of musical signals.

The current system can be extended in several ways to more complex language models as used in speech recognition. As discussed in Sec. 4, our system can be extended to use sub-word models, in order for it to be feasible for large-vocabulary speech BWE. Our current algorithm is an offline method since we used the forward-backward algorithm. In order for it to work online, we can simply use the forward algorithm [6].

9. REFERENCES

[1] P. Jax, *Enhancement of Bandlimited Speech Signals: Angorithms and Theoretical Bounds*, Ph.D. dissertation, Rheinisch-Westfalische Technische Hochschule Aachen, 2002.

[2] K-Y Park and H.S. Kim, “Narrowband to wideband conversion of speech using gmm based transformation,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2000.

[3] P. Bauer and T. Fingscheidt, “A statistical framework for artificial bandwidth extension exploiting speech waveform and phonetic transcription,” in *European Signal Processing Conference*, 2009.

[4] H. Pulakka and P. Alku, “Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum,” *IEEE Trans. Audio, Speech, & Language*, vol. 19, no. 7, pp. 2170–2183, 2011.

[5] P. Smaragdis, B. Raj, and M. Shashanka, “Example-driven bandwidth expansion,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007.

[6] L. Rabiner and B-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[7] G.J. Mysore and P. Smaragdis, “A non-negative approach to language informed speech separation,” in *International Conference on Latent Variable Analysis and Signal Separation*, 2012.

[8] D. Lee and S. Seung, “Learning the parts of objects by non-negative matrix actorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[9] P. Smaragdis, M. Shashanka, and B. Raj, “Probabilistic latent variable model for acoustic modeling,” in *Advances in models for acoustic processing workshop, NIPS*. 2006.

[10] G.J. Mysore, P. Smaragdis, and B. Raj, “Non-negative hidden markov modeling of audio with application to source separation,” in *International Conference on Latent Variable Analysis and Signal Separation*, 2010.

[11] M. Cooke, J.R. Hershey, and S.J. Rennie, “Monaural speech separation and recognition challenge,” *Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[12] J. Han, G.J. Mysore, and B. Pardo, “Audio imputation using the non-negative hidden markov model,” in *International Conference on Latent Variable Analysis and Signal Separation*, 2012.

[13] S. Nawab, T. Quatieri, and J. Lim, “Signal reconstruction from short-time fourier transform magnitude,” *IEEE Trans. Acoustics, Speech, & Signal Processing*, vol. 31, pp. 986–998, 1983.

[14] Y. Hu and P.C. Loizou, “Evaluation of objective measures for speech enhancement,” in *The Ninth International Conference on Spoken Language Processing*, 2006.