Artificial Bandwidth Extension of Spectral Envelope along a Viterbi Path

Can Yağlı, M.A. Tuğtekin Turan, Engin Erzin*

Multimedia, Vision and Graphics Laboratory,
College of Engineering, Koç University, 34450 Sariyer, Istanbul, Turkey

Abstract

In this paper, we propose a hidden Markov model (HMM)-based wideband spectral envelope estimation method for the artificial bandwidth extension problem. The proposed HMM-based estimator decodes an optimal Viterbi path based on the temporal contour of the narrowband spectral envelope and then performs the minimum mean square error (MMSE) estimation of the wideband spectral envelope on this path. Experimental evaluations are performed to compare the proposed estimator to the state-of-the-art HMM and Gaussian mixture model based estimators using both objective and subjective evaluations. Objective evaluations are performed with the log-spectral distortion (LSD) and the wideband perceptual evaluation of speech quality (PESQ) metrics. Subjective evaluations are performed with the A/B pair comparison listening test. Both objective and subjective evaluations yield that the proposed wideband spectral envelope estimator consistently improves performances over the state-of-the-art estimators.

*Corresponding author.
E-mail addresses: canyagli@ku.edu.tr (C. Yagli), mturan@ku.edu.tr (M.A.T. Turan), eerzin@ku.edu.tr (E. Erzin).

Keywords: Artificial Bandwidth Extension, Source-Filter Separation, Line Spectral Frequency, Joint Temporal Analysis

1. Introduction

The artificial bandwidth extension (ABE) problem deals with the estimation of wideband speech signals from narrowband speech, which is historically used in public telephone networks with band-limited spectra in the range of 250 to 3400 Hz. Although intelligibility of narrowband speech is high, studies show that the perceived quality of narrowband speech is significantly degraded compared to wideband speech, which is often limited to the range 50-7000 Hz in frequency (Voran, 1997). The missing frequency bands between wideband and narrowband speech carry spectrally rich information for the speech signal and reduce the listening effort. Artificial bandwidth extension has been studied widely to upgrade the quality of the conventional narrowband speech (Cheng et al., 1994; Jax and Vary, 2003; Agiomyrgiannakis and Stylianou, 2007).

The recent studies on artificial bandwidth extension have exploited various approaches, among which the source-filter separation is one of the widely used techniques. This technique breaks down the ABE problem into two, namely the extension of the excitation (source) and the extension of the spectral envelope (filter). Jax (2004) presents the fact that in listening experiments with original spectral envelope and modified excitation in the extended frequency bands, the perceived quality is only slightly inferior to the quality of the original wideband speech. Hence, Jax and Vary (2003) pose the extension of the envelope as the principal problem of ABE. In the ABE

literature various techniques have been developed for the extension of the spectral envelope.

In an early study, Enbom and Kleijn (1999) use vector quantization to estimate the spectral envelope of the wideband signal. A codebook of wideband and narrowband spectral features, which are extracted from parallel recordings of wideband and narrowband speech, has been created. In the extension phase, a narrowband spectral feature vector of each narrowband speech frame is quantized over the narrowband part of the codebook entries with the minimum Euclidean distance. The wideband part of a quantized codebook entry is then used to determine the spectral envelope in the highband of the wideband speech.

Park and Kim (2000) and Agiomyrgiannakis and Stylianou (2007) successively present artificial bandwidth extension techniques using a statistical approach based on Gaussian mixture models (GMMs). As performed in the vector quantization approach of Enbom and Kleijn (1999), a set of joint wideband and narrowband spectral feature vectors is used to construct a GMM by the expectation maximization (EM) training. In the extension phase, a mapping function transforms a narrowband spectral feature vector to a wideband spectral feature vector with the minimum mean square error (MMSE) estimation for each mixture component of the GMM. Note that each mixture component in the GMM defines a joint cluster for the wideband and narrowband spectral vectors. A weighted sum of the transformed wideband spectral feature vectors over all mixture components then defines a soft mapping from narrowband to wideband spectra. The weights of the soft mapping are set as the posterior probabilities that the mixture components

of the GMM generated the given narrowband spectral feature vector.

Jax and Vary (2003) introduced a hidden Markov model (HMM)-based wideband spectral envelope estimator. Their HMM-based estimator can be modeled as a weighted sum of all estimations in all states with a soft mapping, which is defined by the emission probabilities of the narrowband observations and the state transition probabilities. An open challenge in the HMM-based estimation is to perform online, i.e., real-time, estimation along the most likely state sequence with minimum latency. In this paper, we address this challenge by adapting the HMM-based joint temporal analysis framework of Erzin (2009) and by improving our recent work in (Yagli and Erzin, 2011). Note that Yagli and Erzin (2011) presents an offline solution to the estimation of a wideband spectral envelope in an HMM-based framework. In this study we upgrade this offline solution to an online solution by defining an online Viterbi decoding algorithm with a number of look-ahead frames. The proposed HMM-based estimator decodes an online optimal Viterbi path, which is a state sequence generated with the highest likelihood for a given sequence of narrowband spectral features covering present, past and some future time frames. Note that future time frames, i.e., look-ahead frames, are needed to attain better state decoding for the present time frame. Then state dependent MMSE estimators predict wideband spectral features from narrowband spectral features along the decoded state sequence with a soft mapping. We present supporting experimental evaluations to investigate the optimal number of look-ahead frames in the proposed online solution for the wideband spectral envelope estimation problem. Furthermore, we compare the proposed online estimator to the state-of-the-art HMM and GMM-based

estimators using both objective and subjective evaluations.

The rest of the paper is organized as follows. In Section 2, we present the proposed HMM-based spectral envelope estimation method as well as brief definitions of the state-of-the-art GMM and HMM-based estimation methods. We evaluate the proposed HMM-based estimator compared to the state-of-the-art estimators in Section 3. The conclusion is given in Section 4.

2. Spectral Envelope Estimation

Let us consider narrowband and wideband spectral envelope representations respectively as an observable source $\mathcal{X}$ and as a hidden source $\mathcal{Y}$. We define the spectral envelope estimation problem as uncovering $\mathcal{Y}$ given $\mathcal{X}$. In this section we introduce a new wideband spectral envelope estimation framework. Fig. 1 presents a block diagram of the proposed framework. In this framework we model temporal correlations of the joint narrowband and wideband sources, which we define as source space $\mathcal{Z}$, using a hidden Markov model. Then the wideband spectral envelope is estimated along the Viterbi path of the narrowband spectral envelope with minimum mean square error (MMSE) estimators.

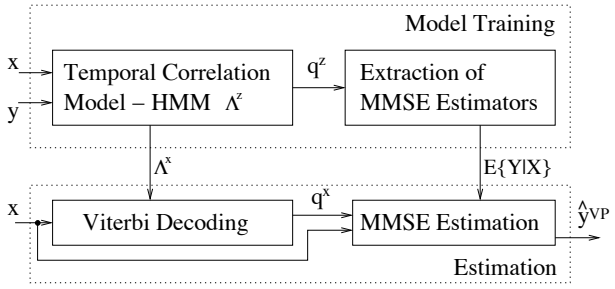


Figure 1: The block diagram of the spectral envelope estimation framework.

2.1. Temporal Correlation Model

Given a sequence of parallel narrowband and wideband spectral envelope representations, we train a hidden Markov model to capture the temporal correlations between these two sources. Let us define the elements of this sequence at time frame t as column vectors $\mathbf{x}_t$ and $\mathbf{y}_t$, respectively representing the narrowband and wideband spectral envelopes. In this study, the spectral envelope is represented with the line spectrum frequency (LSF) vector of the linear prediction filter. The narrowband and wideband spectral representations are respectively extracted as 10-th order and 16-th order linear prediction filters over a 20 ms time frame. Note that we assume frame-based synchrony between $\mathbf{x}_t$ and $\mathbf{y}_t$ that can be achieved on the parallel narrowband and wideband speech signal streams. The joint source vector of LSF features then is defined as $\mathbf{z}_t = [\mathbf{x}_t^T \ \mathbf{y}_t^T]^T$ in $\mathcal{Z}$.

An HMM representing observations from the joint source $\mathcal{Z}$ with N fully connected states can be modeled as $\Lambda^z = (\mathbf{A}_z, \mathbf{B}_z, \mathbf{\Pi}_z)$. The states of Λ^z represent clusters of correlated observations from the joint source, and these clusters are temporally connected to each other with a state transition probability matrix of Λ^z. The $N \times N$ state transition matrix $\mathbf{A}_z$ is defined by entries a_{ij} representing the state transition probability from state s_i to s_j,

$$\mathbf{A}_z : \{a_{ij} = P(q_t = s_j | q_{t-1} = s_i)\} \quad i, j = 1, ..., N, \tag{1}$$

where q_t represents the state at time t. The observation emission distribution $\mathbf{B}_z$ is modeled by the K-mixture Gaussian density functions for each state

s_i,

$$\begin{aligned} \mathbf{B}_z : \{b_i(\mathbf{z}_t) &= P(\mathbf{z}_t|q_t = s_i) \\ &= \sum_{k=1}^{K} \omega_{ik} \mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_{z,ik}, \boldsymbol{\Sigma}_{z,ik})\}, \end{aligned} \tag{2}$$

where ω_{ik} is the mixture weight for the k-th mixture at state s_i. The mean vector $\boldsymbol{\mu}_{z,ik}$ and the diagonal covariance matrix $\boldsymbol{\Sigma}_{z,ik}$ are defined as the cascades of mean vectors and diagonal covariance matrices of the narrowband and wideband sources,

$$\boldsymbol{\mu}_{z,ik} = [\boldsymbol{\mu}_{x,ik}^T \ \boldsymbol{\mu}_{y,ik}^T]^T, \tag{3}$$

$$\boldsymbol{\Sigma}_{z,ik} = diag(\boldsymbol{\Sigma}_{x,ik}, \boldsymbol{\Sigma}_{y,ik}). \tag{4}$$

The initial state probability vector $\boldsymbol{\Pi}_z$ is defined by entries π_i representing the probability of visiting state s_i as the first state q_1,

$$\boldsymbol{\Pi}_z : \{\pi_i = P(q_1 = s_i)\} \quad i = 1, ..., N. \tag{5}$$

The Λ^z model is trained using the Baum-Welch re-estimation algorithm with a joint source vector sequence, which is extracted from a parallel narrowband-wideband speech corpus.

2.2. Extraction of MMSE Estimators

Given the Viterbi state sequence $\mathbf{q}^z$ for the parallel training corpus, we can extract the set of observations, $\mathcal{Z}_{ik}$, for state s_i and the mixture component k, such that,

$$\begin{aligned} \mathcal{Z}_{ik} = \{\mathbf{z}_{t_1,ik}, ..., \mathbf{z}_{t_M,ik} | q_{t_m} = s_i, \\ k = \arg\min_l ||\mathbf{z}_{t_m,ik} - \boldsymbol{\mu}_{z,il}||\}, \end{aligned} \tag{6}$$

where observation $\mathbf{z}_{t_m,ik}$ at time frame t_m occurs at state s_i and has the minimum Euclidean distance to the mean of the k-th mixture component, $\boldsymbol{\mu}_{z,ik}$. The indexed sets of observations $\mathcal{Z}_{ik}$ define neighborhoods for the MMSE sense estimators. Within each neighborhood a full covariance matrix can be computed as

$$\begin{aligned} C_{z,ik} &= \frac{1}{M}\sum_{m=1}^{M}(\mathbf{z}_{t_m,ik}-\mu_{z,ik})(\mathbf{z}_{t_m,ik}-\mu_{z,ik})^T \\ &= \begin{bmatrix} C_{xx,ik} & C_{xy,ik} \\ C_{yx,ik} & C_{yy,ik} \end{bmatrix}. \end{aligned} \tag{7}$$

Recall that the HMM Λ^z includes diagonal covariance matrix representations, which have faster convergence in the expectation-maximization training. Furthermore mixtures of Gaussians are known to model underlying multivariate distributions successfully. However in the MMSE estimation, temporal correlations are captured with the full covariance matrix representations, which are defined in the temporal neighborhoods of mixtures.

Then having the full covariance matrix representations and given the observation source vector $\mathbf{x}_t$ at state s_i, the MMSE estimator is defined as

$$\begin{aligned} \hat{\mathbf{y}}_{t,i} &= E\{\mathbf{y}_t|\mathbf{x}_t, q_t = s_i\} \\ &= \sum_{k=1}^{K}\omega_{y|x,ik}(\mu_{y,ik} + C_{yx,ik}C_{yy,ik}^{-1}(\mathbf{x}_t - \mu_{x,ik})), \end{aligned} \tag{8}$$

where the weighting factor $\omega_{y|x,ik}$ defines the probability of the k-th mixture at state s_i given observation x_t and it is computed as the normalized Gaussian probability density function,

$$\omega_{y|x,ik} = \frac{\omega_{ik}\mathcal{N}(\mathbf{x}_t;\mu_{x,ik},C_{xx,ik})}{\sum_{l=1}^{K}\omega_{il}\mathcal{N}(\mathbf{x}_t;\mu_{x,il},C_{xx,il})}. \tag{9}$$

2.3. Viterbi Decoding

An HMM Λ^x representing observations from the narrowband source $\mathcal{X}$ can be extracted from the HMM Λ^z of the joint source $\mathcal{Z}$. The model Λ^x shares the same state transition $\mathbf{A}_z$ and initial state probability $\mathbf{\Pi}_z$ matrices with the joint model Λ^z, as they use the same underlying state machine. The observation emission distribution $\mathbf{B}_x$ of Λ^x can be extracted from the observation emission distribution $\mathbf{B}_z$ of Λ^z as

$$\mathbf{B}_x : \{b_i(\mathbf{x}_t) = \sum_{k=1}^{K} \omega_{ik} \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{x,ik}, \mathbf{\Sigma}_{x,ik})\} \quad i = 1, ..., N, \tag{10}$$

where ω_{ik} are the weights in (2), and the mean vector $\boldsymbol{\mu}_{x,ik}$ and the diagonal covariance $\mathbf{\Sigma}_{x,ik}$ are extracted respectively from (3) and (4). The resulting HMM for the narrowband source then can be represented as $\Lambda^x = (\mathbf{A}_z, \mathbf{B}_x, \mathbf{\Pi}_z)$.

Viterbi decoding can find the optimal state sequence $\{q_1, q_2, \dots\}$ given a sequence of narrowband observation vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$ and the model Λ^x. However we would like to put a constraint on the latency of this decoding, since common ABE applications require real-time processing with minimal latency. Hence we consider the following modified Viterbi decoding algorithm with T frames look-ahead:

(1) Set time frame $t = 1$.

(2) Initialization: Initialize an accumulated forward probability for the best path as $V_t(i)$ at present time frame t in state i:

$$V_t(i) = p_{ti} b_i(\mathbf{x}_t) \quad i = 1, ..., N, \tag{11}$$

where p_{ti} represents the initial state probability, π_i, at time frame $t = 1$, and an accumulated probability to be in state i at time frame t when $t > 1$:

$$p_{ti} = \begin{cases} \pi_i & t = 1, \\ V_{t-1}(j)a_{ji} & j = q_{t-1} \text{ and } t = 2, 3, ... \end{cases} \tag{12}$$

(3) Recursion: Compute the accumulated forward probability for the best path in state j with T frames look-ahead for $t' = t+1, t+2, ..., t+T$:

$$V_{t'}(j) = \max_i \{V_{t'-1}(i)a_{ij}\}b_j(x_{t'}) \quad j = 1, ..., N, \tag{13}$$

$$Q_{t'}(j) = \arg\max_i \{V_{t'-1}(i)a_{ij}\} \quad j = 1, ..., N, \tag{14}$$

where $Q_{t'}(j)$ holds the source state for the most likely transition to state j at time frame t', and $V_{t'}(j)$ holds the highest likelihood score to be in state j at time frame t'.

(4) Backtrace: Set the most likely state at time frame $t+T$:

$$q_{t+T} = \arg\max_i \{V_{t+T}(i)\}. \tag{15}$$

Then backtrace over the time frames to decode the state at time t,

$$q_{t'} = Q_{t'+1}(q_{t'+1}) \quad t' = t+T-1, t+T-2, ..., t. \tag{16}$$

(5) Iterate: Update $t = t+1$, go to (2)

2.4. MMSE Estimation of Wideband Spectra

The modified Viterbi algorithm with T frames look-ahead decodes the state q_t given the narrowband observation vectors $\mathbf{x}_t, \mathbf{x}_{t+1}, ..., \mathbf{x}_{t+T}$ and the

Λ^x model. The optimal linear estimator in the MMSE sense along the Viterbi path is the conditional expectation of the wideband source vector $\mathbf{y}_t$ given the state at time $t-1$, q_{t-1}, and the narrowband observations $\mathbf{x}_t, \mathbf{x}_{t+1}, ..., \mathbf{x}_{t+T}$. Note that this conditional expectation after decoding q_t is equivalent to the expectation of $\mathbf{y}_t$ given the current state q_t and the current observation $\mathbf{x}_t$. Hence the optimal linear estimator along the Viterbi path can be defined as

$$\hat{\mathbf{y}}_t^{VP} = E\{\mathbf{y}_t | q_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, ..., \mathbf{x}_{t+T}\} \quad (17)$$

$$= E\{\mathbf{y}_t | \mathbf{x}_t, q_t\} = \hat{\mathbf{y}}_{t,q_t} \quad (18)$$

where the resulting expectation can be computed as in (8) to be $\hat{\mathbf{y}}_{t,q_t}$.

2.5. Baseline Estimators

In this study we consider two state-of-the-art baseline estimation techniques for wideband spectral estimation. These are the GMM estimator of Agiomyrgiannakis and Stylianou (2007) and the HMM estimator of Jax and Vary (2003).

The GMM estimator of Agiomyrgiannakis and Stylianou (2007) is a soft mapping from observable source $\mathcal{X}$ to hidden source $\mathcal{Y}$ with an optimal linear transformation in the MMSE sense. It can be formulated as the MMSE estimator in (8) for a single state,

$$\hat{\mathbf{y}}_t^{GMM} = \sum_{l=1}^{L} p(\gamma_l | \mathbf{x}_t)[\boldsymbol{\mu}_{y,l} + \mathbf{C}_{yx,l}(\mathbf{C}_{xx,l})^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_{x,l})], \quad (19)$$

where γ_l is the l-th Gaussian mixture and L represents the total number of Gaussian mixtures. The vectors $\boldsymbol{\mu}_{x,l}$ and $\boldsymbol{\mu}_{y,l}$ are respectively the centroids for the l-th Gaussian for sources $\mathcal{X}$ and $\mathcal{Y}$, $\mathbf{C}_{xx,l}$ is the covariance matrix

of source $\mathcal{X}$ in the l-th Gaussian, and $\mathbf{C}_{yx,l}$ is the cross-covariance matrix of sources $\mathcal{X}$ and $\mathcal{Y}$ for the l-th Gaussian mixture. The probability of the l-th Gaussian mixture given the observation $\mathbf{x}_t$ is defined as the normalized Gaussian pdf as

$$p(\gamma_l|\mathbf{x}_t) = \frac{\mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{x,l}, \mathbf{C}_{xx,l})}{\sum_{m=1}^{L} \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{x,m}, \mathbf{C}_{xx,m})}. \tag{20}$$

The HMM estimator of Jax and Vary (2003) is a soft mapping from the MMSE estimators in (8) across all states with all possible state transitions,

$$\hat{\mathbf{y}}_t^{HMM} = \sum_{i=1}^{N} \hat{\mathbf{y}}_{t,i} P(q_t = s_i|\mathbf{x}_t), \tag{21}$$

where the conditional probability of the state s_i given the source $\mathbf{x}_t$ is the soft weighting factor and can be computed using the joint probability of state and source,

$$P(q_t = s_i|\mathbf{x}_t) = \frac{P(q_t = s_i, \mathbf{x}_t)}{\sum_{j=1}^{N} P(q_t = s_j, \mathbf{x}_t)}. \tag{22}$$

The joint probability of state and source can be defined as a forward variable, $\alpha_i(t) = P(q_t = s_i, \mathbf{x}_t)$, and calculated with the following iterative computation over time t:

$$\alpha_i(1) = \pi_i b_i(\mathbf{x}_1) \quad i = 1, ..., N, \tag{23}$$

$$\alpha_i(t) = \sum_{j=1}^{N} \alpha_j(t-1) a_{ji} b_i(\mathbf{x}_t) \quad i = 1, ..., N. \tag{24}$$

3. Experimental Evaluations

We use the source-filter separation framework for the ABE, where a speech signal is decomposed into excitation signal and spectral envelope.

We model the spectral envelope with the line spectrum frequency (LSF) representation of the linear prediction filter. We extract 10-th order and 16-th order linear prediction filters for each 20 ms time frame for the narrowband and wideband speech spectra representations, respectively. The excitation signal, when needed in evaluations, is kept as the original excitation signal of the wideband speech, since the main focus of this study is to evaluate the proposed spectral envelope estimation technique. Use of the original excitation signal eliminates possible degradations that are caused by excitation extension and helps us to evaluate the performance of the spectral envelope estimation techniques.

Experimental evaluations are performed on the TIMIT database. The training set contains 2079 sentences from 693 speakers. Independent from the training set, the test set contains 756 sentences from 252 speakers. In our evaluations we use two types of narrowband speech source: low-pass and band-pass filtered versions of the wideband speech. The low-pass filtered wideband speech is down-sampled to get the narrowband speech source with the missing high-band spectral content above 3900 Hz. An intermediate reference system (IRS) filter simulates the band-pass magnitude response of the telephone lines (ITU-T Recommendation P.48, 1993). The IRS filter delivers -10 dB attenuation at frequency limits 340 Hz and 3550 Hz. The IRS filtered wideband speech is down-sampled to get the band-pass filtered narrowband speech source with the missing low-band and high-band spectral contents.

3.1. Objective Evaluations

Evaluations of the spectral envelope estimation for the ABE are performed with two distinct objective metrics, the logarithmic spectral distortion (LSD) and the perceptual evaluation of wideband speech quality (PESQ) metrics. The LSD is a widely used metric for speech spectral envelope quality assessment. The LSD metric assesses the quality of the estimated spectral envelope with respect to the original wideband spectral envelope, and it is defined as

$$d_{LS} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(10\log\frac{|A_y(\omega)|^2}{|\hat{A}_y(\omega)|^2}\right)^2 d\omega}, \tag{25}$$

where $A_y(\omega)$ and $\hat{A}_y(\omega)$ represent the original and estimated wideband spectral envelopes, respectively. The ITU-T Standard PESQ (ITU-T Recommendation P.862.2, 2005) is employed as the second objective metric to evaluate the perceptual quality of the synthesized wideband speech signal, which is constructed using the estimated spectral envelope and the original excitation signal of the wideband speech.

3.1.1. Effect of the look-ahead duration

First, we investigate the effect of the look-ahead duration for the proposed HMM-based MMSE estimation along the Viterbi path. The LSD and PESQ performance evaluations of the proposed estimator as a function of the look-ahead duration together with performances of the baseline HMM estimator (Jax and Vary, 2003) for the low-pass and IRS filtered narrowband sources are plotted in Fig. 2. In both HMM-based estimators $N = 64$ states and $K = 16$ Gaussian mixture components per state are used. Although relative performance improvements are larger for the IRS filtered narrow-

band source, objective quality metrics have better scores for the low-pass filtered narrowband source. This is expected since the wideband spectral envelope estimation is harder for the IRS filtered narrowband source, which misses both the low-band and high-band spectral contents. The resulting ABE synthesis from the IRS filtered source has more degradations than the ABE synthesis of the low-pass filtered source.

On the other hand, we observe consistent performance improvements in all cases for the proposed estimator as the look-ahead duration increases. This indicates that temporal correlations in the look-ahead segment help to improve estimation of the wideband spectral envelope. An eventual performance saturation is observed after look-ahead durations of $T = 8$ and $T = 6$ frames for the low-pass and IRS filtered narrowband sources, respectively. This behaviour is expected, since the HMM structure models temporal correlations, and temporal correlations get weaker as frames become separated from each other temporally. Hence we observe that frames separated by more than 8 frames for the low-pass filtered and more than 6 frames for the IRS filtered narrowband sources do carry weak correlations. This indicates that temporal correlations last shorter for the IRS filtered narrowband source. A possible reason for this effect could be the missing slowly changing low-band spectral content of the IRS filtered source. Note that significant performance improvements of the proposed estimator occur with the first couple of look-ahead frames, and the proposed estimator achieves better performance scores than the baseline HMM estimator when the number of look-ahead frames is larger than one.

3.1.2. Effect of the estimators

We also investigate the mean LSD and PESQ performances of different estimators with the low-pass and IRS filtered narrowband sources in Fig. 3. In these investigations we fixed $K = 16$ Gaussian mixture components per state for the HMM-based estimators, and performance analysis is done as a function of the number of states, N. Performance analysis of the GMM estimator is done as a function of the number of mixture components, L.

Although the mean LSD performance of the proposed estimator without look-ahead, i.e., $T = 0$, is observed to be equivalent to the performance of the baseline HMM estimator with the low-pass filtered source, the proposed estimator performs significantly better with the IRS filtered source. We observe that even without look-ahead, the temporal correlation model of the proposed estimator brings significant performance improvements with the IRS filtered source. Furthermore the use of $T = 8$ look-ahead frames with the proposed estimator brings consistent mean LSD performance improvements with both types of the narrowband sources. As plotted in Fig. 3(a) for the low-pass filtered narrowband source, the mean LSD performance improvement of the proposed estimator with $T = 8$ look-ahead frames is comparable to the mean LSD improvement from the GMM estimator to the baseline HMM estimator.

Fig. 3(b) and 3(d) present the mean PESQ performance scores for the low-pass and IRS filtered sources, respectively. In Fig. 3(b) for the low-pass filtered narrowband source, mean PESQ performance improvements of the proposed estimator with $T = 8$ look-ahead frames are more limited. The mean PESQ performance scores for the IRS filtered source have stronger

improvements with the proposed estimators, and furthermore using $T = 8$ look-ahead frames introduces a consistent improvement over the other three estimators.

3.2. Subjective Evaluations

We performed two subjective A/B comparison listening tests, for the low-pass and IRS filtered sources, to measure the perceived quality of the proposed spectral envelope estimation with respect to baseline estimators. During the tests, the subjects are asked to indicate their preference for each of given A/B test pair sentences on a scale of (-2; -1; 0; 1; 2), where the scale corresponds to *strongly prefer A*, *prefer A*, *no preference*, *prefer B*, and *strongly prefer B*, respectively. The ABE synthesis with the proposed spectral envelope estimator is compared to wideband and narrowband speech as well as the ABE synthesis of the baseline HMM and GMM estimators. Among these five conditions we select ten comparison pairs including a null comparison, wideband with wideband, to assess the reliability of the listening test. We include three sentences for each comparison pair and the test is performed over 18 subjects. The average preference scores for all comparison sets are presented in Table 1 and 2, respectively, for the low-pass and IRS filtered sources. Note that the rows and the columns of these tables correspond to A and B of the A/B pairs, respectively. Also, the average preference scores that tend to favor B are given in bold to ease visual inspection.

The wideband speech is the most preferred and the narrowband speech is the least preferred condition. The ABE synthesis of the proposed spectral envelope estimation (VP with $N = 64$, $T = 8$ and $K = 16$) is preferred over the baseline HMM estimator (HMM with $N = 64$ and $K = 16$) with 0.118

Table 1: Average results of the subjective A/B pair comparison test for the low-pass filtered narrowband source

		B			
		WB	**GMM**	**HMM**	**NB**
A	WB	**0.019**	-0.815	-0.815	-1.510
	VP	**0.593**	-0.294	-0.118	-1.314
	GMM	**0.815**			-1.274
	HMM	**0.815**			-1.137

and 0.167 average scores for the low-pass and IRS filtered sources, respectively. Furthermore the proposed estimator is preferred more strongly over the GMM estimator (GMM with $L = 256$) with 0.294 and 0.185 average scores for the low-pass and IRS filtered sources, respectively. When we compare with the wideband speech quality, the ABE synthesis with the proposed estimator attains the minimum average scores 0.593 and 0.630 respectively for the low-pass and IRS filtered sources. Also note that compared with the narrowband speech, although the wideband speech and the ABE synthesis with the proposed estimator are respectively the most preferred conditions with 1.510 and 1.314 average scores for the low-pass filtered source, it's not the case with the IRS filtered source. Synthesized wideband speech samples, which demonstrate perceptual quality, are also available online (Yagli and Erzin, 2012).

4. Conclusions

In this paper, we present a new HMM-based wideband spectral envelope estimation method. The main contribution of the proposed method is to de-

Table 2: Average results of the subjective A/B pair comparison test for the IRS filtered narrowband source

		B			
		WB	**GMM**	**HMM**	**NB**
A	WB	**0.019**	-0.889	-0.907	-1.574
	VP	**0.630**	-0.185	-0.167	-1.167
	GMM	**0.889**			-1.315
	HMM	**0.907**			-1.500

code an optimal Viterbi path, which defines a temporally correlated contour, on the sequence of the narrowband spectral envelope. Then the wideband spectral envelope is estimated along this path with linear estimators in the MMSE sense. The state-of-the-art wideband spectral envelope estimation methods, which are investigated as baseline estimators, are in general a soft fusion of all linear estimators within the model. However, in the proposed HMM-based estimation method soft fusion is performed at the state level and the state sequence is extracted through Viterbi decoding.

We investigate the optimal duration of temporally correlated spectral envelope contours to maximize the performance of the proposed estimator with the low-pass and IRS filtered narrowband sources. We observe that a temporal duration of $T = 8$ look-ahead frames is optimal to attain the best LSD and PESQ performance scores with the low-pass filtered source. A faster convergence to the best objective scores is observed with the IRS filtered source. Hence we can conclude that the temporal correlations last shorter for the IRS filtered source. Two thirds of the performance improvements are attained with a temporal look-ahead duration of $T = 2$ frames for both

performance metrics with both narrowband sources. We also observe that the proposed HMM-based estimator consistently improves LSD and PESQ performances as the look-ahead duration and the number of states in the model increase. The proposed estimator attains higher objective and subjective performances than the baseline estimators, and stronger objective performance improvements are especially observed with the IRS filtered narrowband source.

The proposed HMM-based estimator, which is capable of modeling strong temporal correlations between two synchronous sources, can be further studied for missing feature extraction, voice conversion or non-acoustic sensors to speech conversion problems.

References

Agiomyrgiannakis, Y., Stylianou, Y., Feb. 2007. Conditional vector quantization for speech coding. IEEE Transactions on Audio, Speech, and Language Processing 15 (2), 377–386.

Cheng, Y. M., O'Shaughnessy, D., Mermelstein, P., Oct. 1994. Statistical recovery of wideband speech from narrowband speech. IEEE Transactions on Speech and Audio Processing 2 (4), 544–548.

Enbom, N., Kleijn, W., 1999. Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients. In: IEEE Workshop on Speech Coding Proceedings. pp. 171–173.

Erzin, E., Sept. 2009. Improving throat microphone speech recognition by

joint analysis of throat and acoustic microphone recordings. IEEE Transactions on Audio, Speech, and Language Processing 17 (7), 1316–1324.

ITU-T Recommendation P.48, 1993. Specification for an intermediate reference system.

ITU-T Recommendation P.862.2, November 2005. Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs.

Jax, P., 2004. Bandwidth extension for speech. In: Larsen, E., Aarts, R. M. (Eds.), Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design. Wiley, pp. 171–236.

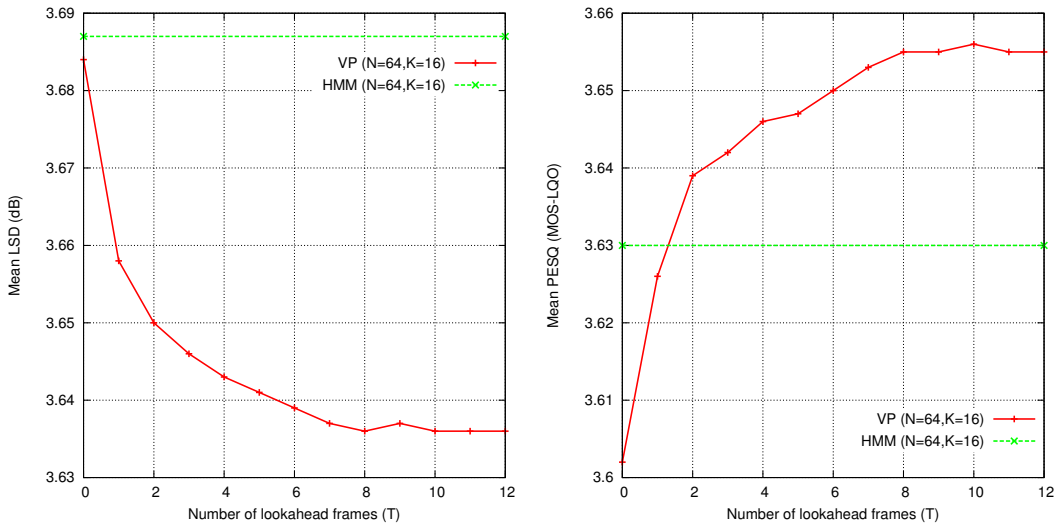
Jax, P., Vary, P., 2003. On artificial bandwidth extension of telephone speech. Signal Processing 83, 1707–1719.

Park, K., Kim, H., 2000. Narrowband to wideband conversion of speech using gmm based transformation. In: 2000 IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings (ICASSP '00). Vol. 3. pp. 1843–1846.

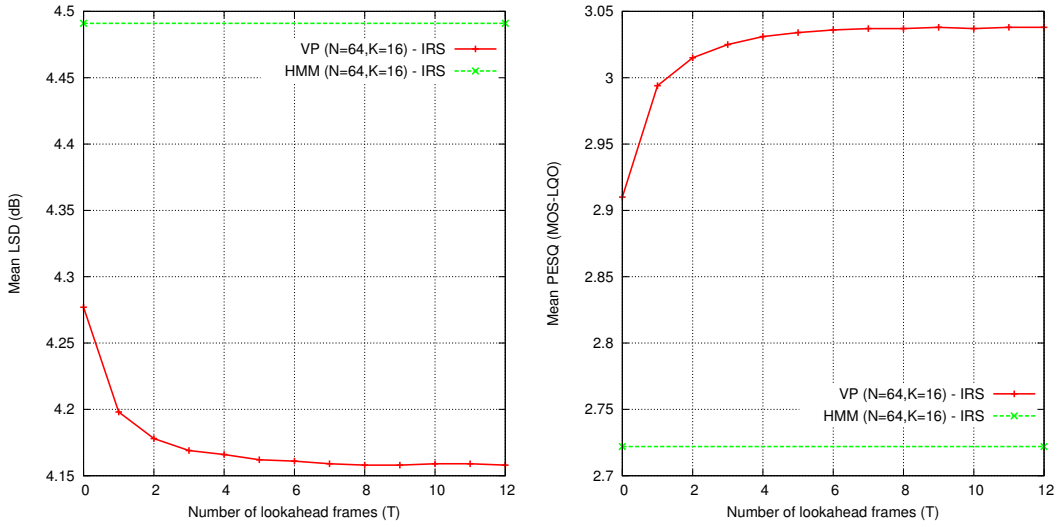
Voran, S., September 1997. Listener ratings of speech passbands. In: IEEE Workshop on Speech Coding Proceedings. pp. 81–82.

Yagli, C., Erzin, E., 2011. Artificial bandwidth extension of spectral envelope with temporal clustering. In: 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP '11). pp. 5096–5099.

Yagli, C., Erzin, E., Jan. 2012. The ABE speech synthesis samples. Internet: http://home.ku.edu.tr/~eerzin/abe_vp.



(a) (b)



(c) (d)

Figure 2: Objective performance evaluations as a function of the look-ahead duration T for the proposed estimator (VP) together with the performances of the HMM estimator (Jax and Vary, 2003): (a) Mean LSD and (b) Mean PESQ performances for the low-pass filtered narrowband source; (c) Mean LSD and (d) Mean PESQ performances for the IRS filtered narrowband source.

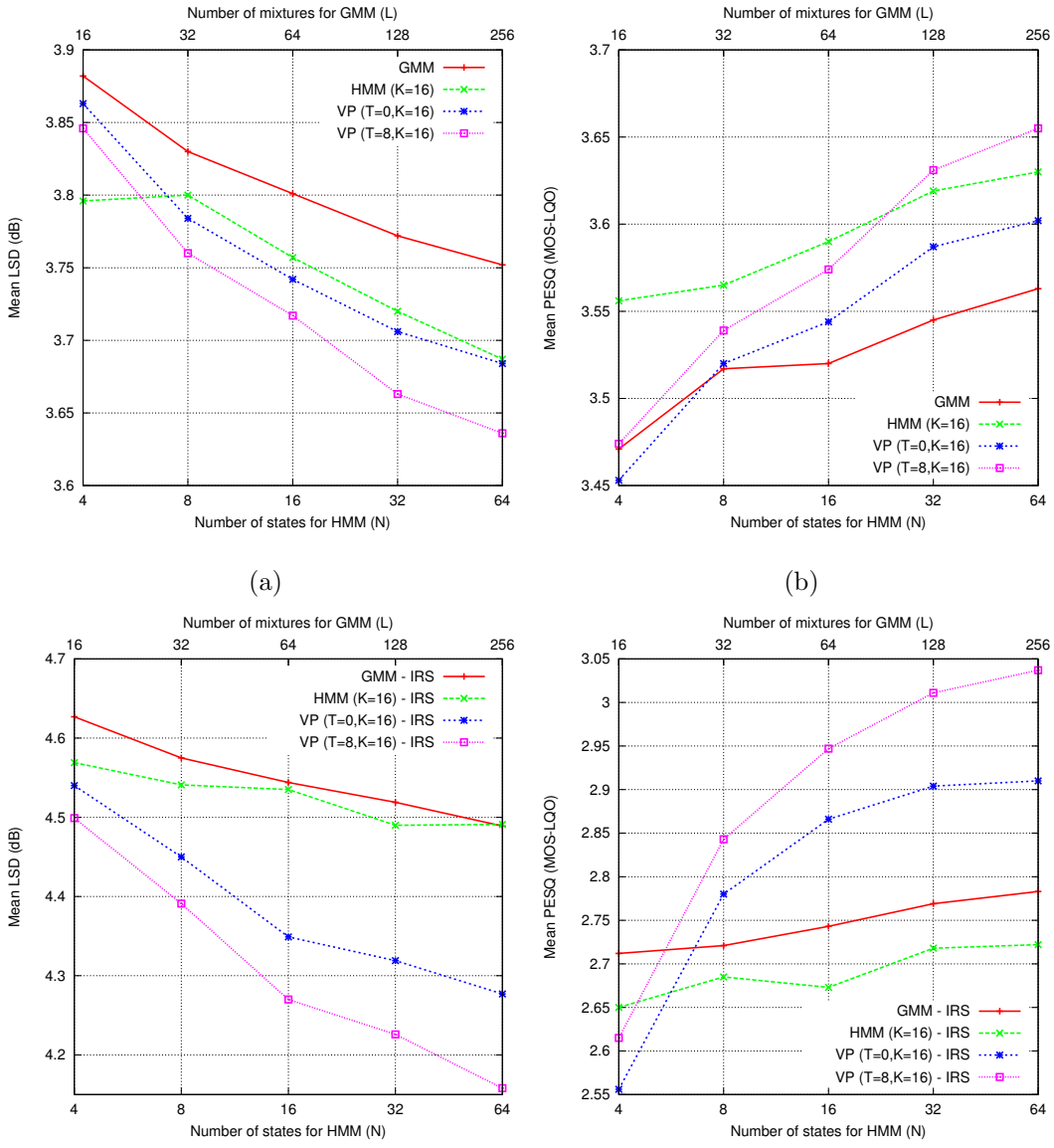


Figure 3: Objective performance evaluations as a function of number of states in the HMM-based and number of mixtures in the GMM-based estimators: (a) Mean LSD and (b) Mean PESQ performances for the low-pass filtered narrowband source; (c) Mean LSD and (d) Mean PESQ performances for the IRS filtered narrowband source.