ARTIFICIAL BANDWIDTH EXTENSION OF SPECTRAL ENVELOPE WITH TEMPORAL CLUSTERING

Can Yağlı and Engin Erzin
College of Engineering, Koç University, Istanbul, Turkey
canyagli.eerzin@ku.edu.tr

ABSTRACT

We present a new wideband spectral envelope estimation framework for the artificial bandwidth extension problem. The proposed framework builds temporal clusters of the joint sub-phone patterns of the narrowband and wideband speech signals using a parallel branch HMM structure. The joint sub-phone patterns define temporally correlated neighborhoods, in which a linear prediction filter estimates spectral features of the corresponding wideband signal from the narrowband signal. The proposed framework is compared to a benchmark vector quantization based artificial bandwidth extension algorithm. Performance evaluations are performed with three distinct objective metrics and a subjective A/B test.

Index Terms — artificial bandwidth extension, linear prediction

1. INTRODUCTION

Historically public telephone networks operate with narrowband speech, which is bandlimited to (250, 3400) Hz in frequency. Even though public telephone exchanges are digital today, the low bandwidth limitation is still present due to the characteristics of the traditional analogue network and related standards. Although intelligibility of the narrowband speech is high, studies show that the perceived quality of the narrowband speech is significantly degraded compared to wideband speech, which is bandlimited to (50, 7000) Hz in frequency. The missing frequency band between wideband and narrowband speech carries spectrally rich information for the speech signal and reduces the listening effort.

Upgrading to wideband speech communication requires the thorough structure to be redesigned, which is an economical burden. In this manner, an alternative method that is compatible with the existing network but applicable on the telephone end is a desirable solution. For this purpose, artificial bandwidth extension (ABE) has been studied widely to upgrade quality of the conventional narrowband speech [1, 2].

The studies on artificial bandwidth extension have exploited various approaches. Source-filter separation is widely used in speech analysis, and majority of the previous work on artificial bandwidth extension employ this model as well. This model breaks down the problem into two, namely the extension of the excitation (source) and the extension of the spectral envelope (filter). As it is stated in [2], extension of the envelope has greater contribution to the perceived speech quality compared to the extension of the excitation. Therefore, more emphasis is given to the extension of the spectral envelope.

Enbom and Klein use vector quantization to estimate the spectral envelope of the wideband signal [3]. Wideband features and narrowband features are grouped together and trained offline, followed by the construction of the codebook using LBG algorithm. In application part, corresponding wideband features of the narrowband files are found using this very codebook.

Kim and Park accomplish the goal of artificial bandwidth extension using a statistical approach, by GMM [4]. Similarly, in the offline training phase, narrowband features and wideband features are fused, trained with GMM using expectation maximization and codebooks are constructed. In the online application phase, the corresponding codebook element decides on the wideband feature of the narrowband test signal.

Jax and Vary suggest the usage of HMM for wideband feature estimation [2]. They build a supervised HMM structure based on VQ clusters, and exploit the statistics of the HMM state sequence in the MMSE estimation of wideband spectra.

In this paper, we also employ an HMM structure as the statistical method to model the correlation between the narrowband and wideband signal. However the proposed HMM structure is trained in unsupervised manner. We subsequently integrate linear prediction filters that estimate wideband features from the entire narrowband spectrum within a temporal cluster. The proposed ABE system is evaluated with three distance metrics and compared to a benchmark system that uses vector quantization.

The proposed artificial bandwidth extension system is given in section 2. In Section 3, the experimentation, evaluation and results are given. Finally, Section 4 includes the discussions and future research directions.

2. ABE SYSTEM

Recently, we build a framework for joint temporal analysis of correlated sources [5]. In this study, we adapt and improve this joint temporal analysis framework to perform artificial bandwidth extension of spectral envelope. We employ the source-filter separation in the ABE system, which performs extension of the excitation signal and the spectral envelope.

Spectral envelope of the narrowband speech signal exhibits correlation to the wideband spectral envelope. In recent works on ABE, methods using vector quantization [3] and statistical mappings such as GMM [4] are used to address this correlation. In this study, we first build a framework for temporal clustering of narrowband and wideband speech signals. Then we construct a spectral mapping by predicting wideband spectral envelope from narrowband envelope within temporal clusters.

We model the spectral envelope with the line spectrum frequency (LSF) representation of linear prediction filter. We will refer narrowband and wideband LSF features as \( f^w = [f_1^w, f_2^w, \ldots, f_{10}^w] \) and \( f^n = [f_1^n, f_2^n, \ldots, f_{16}^n] \), respectively. In the temporal clustering, the LSF features together with the first and second derivatives...
are used and they are referred as \( F = [f, \Delta f, \Delta \Delta f] \). The temporal clustering is discussed in section 2.1. Based on this clustering model, estimation of wideband features from narrowband features is given in section 2.2. Finally, we describe the extension of the excitation signal in section 2.3.

2.1. Temporal Clustering

We use a hidden Markov model (HMM) based unsupervised multi-stream analysis framework to build a correlation model between narrowband and wideband speech [5, 6]. The HMM based unsupervised classifier is used to jointly segment temporal spectral features. The joint temporal feature patterns are used to form a correlation model between the narrowband and wideband speech.

The multi-stream unsupervised segmentation defines recurrent sub-phonetic patterns of the joint features. The narrowband and wideband joint feature stream, \( F^{nw} \), is used to train the HMM structure \( \Lambda^{nw} \) which capture recurrent phonetic segments. The HMM structure \( \Lambda^{nw} \), which is used for unsupervised temporal segmentation, is composed of \( B \) parallel HMMs, \( \{\lambda_1^{nw}, \lambda_2^{nw}, \ldots, \lambda_B^{nw}\} \), where each \( \lambda_b^{nw} \) is chosen to be single state HMM, \( s_b \), as shown in Fig. 1. Given the multimodal feature sequence, \( F^{nw} = \{F_1^{nw}, F_2^{nw}, \ldots, F_{K'}^{nw}\} \), \( F_k^{nw} \) denotes the joint feature vector at frame \( k \). The segmentation of the feature sequence is performed using Viterbi decoding to maximize the probability of model match. The Viterbi decoding yields a state sequence \( q^{nw} = \{q_1^{nw}, q_2^{nw}, \ldots, q_{K'}^{nw}\} \) associated with the feature sequence \( F^{nw} \). The correspondence between states \( q^{nw} \) and feature vectors \( F_k^{nw} \) forms vector quantization like clusters which also profit from temporal correlation due to the nature of the HMM structure.

![Fig. 1. Parallel HMM structure, where \( s_s \) and \( s_e \) are non-emitting states.](image)

2.2. Estimation of Wideband Spectra

In each state, we perform a linear prediction analysis to estimate the \( l \)-th wideband LSF feature from the highly correlated temporal and spatial neighborhoods of the narrowband LSF feature. Let us define the mean removed LSF feature vectors, which falls into state \( s_b \) in \( \Lambda^{nw} \) at frame \( k \) as \( \bar{f}_k^{lw} \) and \( \bar{f}_{k,l}^{lw} \). The source feature set can be formed from a temporal neighborhood of the narrowband LSF features as, \( \bar{x} = \{\bar{f}_{k-2}^{lw}, \ldots, \bar{f}_k^{lw}, \ldots, \bar{f}_{k+2}^{lw}\} \). Now we can state a linear estimator for the \( l \)-th component of the mean removed wideband LSF feature \( y = \bar{f}_{k,l}^{lw} \) as,

\[
y = \bar{x} \omega_l^T
\]

where row vector \( \omega_l \) represents the \( (2T+1) \)-th order linear prediction filter for the \( l \)-th wideband LSF component and \( [\cdot]^T \) is the vector transpose operator. The linear estimator, which minimizes the mean square error between \( y \) and \( \bar{y} \), yields to the well-known Yule-Walker equations,

\[
R_{yx} = R_{xx} \omega_l^T
\]

where \( R_{yx}, \) and \( R_{x,y} \) are the correlation of \( y, x_i, x_j \) signals, respectively \( R_{yx} = E\{yx_i\} \) and \( R_{x,y} = E\{x_i x_j\} \). Since the source feature vector \( x \) is high dimensional and the LSF feature components are mostly locally correlated, we propose a feature selection to reduce dimensionality with highly correlated source components. Hence we define a reduced dimensional source vector as, \( x' = [x_{i_1}, x_{i_2}, \ldots, x_{i_p}] \), such that \( \{i_1, \ldots, i_p\} \) are the indexes of the largest \( p \) correlations in \( R_{yx} \). The dimension reduction can be performed at the cost of registering indexes of the source components with the largest \( p \) correlations for each state \( s_b \) and for each target component \( \bar{f}_{k,l}^{lw} \). Then, the linear prediction analysis yields a solution on the reduced dimensional source vector \( x' \) for the LP filter as,

\[
\omega_{p}^T = R_{x,p}^{-1} R_{y,x'}
\]

where \( \omega_p \) becomes \( p \) dimensional linear predictor. Consecutively, the \( l \)-th wideband feature component at frame \( k \) can be estimated as,

\[
\tilde{f}_{k,l}^{lw} = x \omega_{p}^T + \mu_{p,l}
\]

where \( \mu_{p,l} \) is the \( l \)-th mean wideband feature component of state \( s_b \). Note that the \( p \)-th order LP filter \( \omega \) is extracted for each feature component \( l \) in each state \( s_b \).

The multi-stream parallel branch HMM, which is obtained in the temporal clustering, is split into narrowband, \( \Lambda^n \), and wideband, \( \Lambda^w \), models. These two models share the same state transition probabilities and they have split observation probability density functions representing \( f^n \) and \( f^w \) features. The observation probability density functions can be decidedly be divided into two for Gaussian densities with diagonal covariance. Given the narrowband model \( \Lambda^n \), state dependent linear estimators and narrowband speech, the flow of the wideband spectra estimation is described as follows:

i. The narrowband feature sequence, \( F^n \), is extracted from the narrowband speech.

ii. Temporal segmentation of the narrowband feature sequence \( F^n \) is performed using the HMM model \( \Lambda^n \) to extract temporal sub-phone patterns with a state sequence \( q^n \).

iii. The \( L \) linear predictors in state \( q_{i}^{nw} \) are used to extract the acoustic feature estimate \( \tilde{f}_{k}^{lw} \) as described in (4).

2.3. Extension of the excitation signal

The excitation signal is flat in spectrum. It contains weak harmonics at multiples of the pitch frequency for voiced sounds and is approximately white noise for unvoiced sounds. In [2], the authors note that the extension of the excitation signal has a minor role compared to the extension of the spectral envelope in improving the perceived speech quality. Therefore, we employ a simple, yet effective, spectral mirroring method for the extension of the excitation [3]. The wideband excitation is estimated from narrowband excitation as,

\[
e_w(m) = \begin{cases} 
e_n(k) & \text{if } m = 2k \\ 0 & \text{otherwise} \end{cases}
\]
3. EXPERIMENTATION AND EVALUATION

Experimental evaluations are performed on dialect, sa, and compact, sx, sentences of the TIMIT database. The training sets on SA and SX contain 924 sentences from 462 speakers and 1155 sentences from 231 speakers, respectively. Independent from the training sets, the test sets contain 336 sentences from 168 speakers and 420 sentences from 84 speakers, respectively for SA and SX collections. In our evaluations, we down-sample the speech signals in order to obtain narrowband sets, where the wideband and narrowband sets build a parallel corpus. Feature extraction is performed for each 20 ms frames.

The evaluation of the spectral envelope mapping and ABE system are performed with three distinct objective metrics. The ITU-T Standard WB-PESQ [7, 8] and Segmental SNR are employed to compare the synthesized wideband speech signals to the original wideband speech signals. The logarithmic spectral distortion (LSD) is used to evaluate the estimated spectral envelope with respect to the original wideband spectral envelope. In addition to the objective evaluations, a set of subjective evaluations are also performed.

3.1. Performance of the Benchmark System

In the benchmark system, we employ a VQ-based envelope extension algorithm. Codebooks are extracted using LBG algorithm and the best matching entry of the codebook for a given test signal is chosen to be the closest with respect to the euclidian distance. The extension of the excitation is run by up-sampling with zero insertion. We present the performance results of the VQ algorithm in Table 1. As expected, increasing the codebook size yields better estimation results. For both sets, the mean LSD between the estimated spectral envelope and the original wideband envelope drops from 5.6 dB to 4.6 dB as we increase the codebook size from 4 to 256. A similar improvement is apparent in terms of PESQ score as well. Segmental SNR does not imply a certain performance improvement or degradation with varying codebook size.

Table 1. Performance of the VQ based benchmark ABE system.

<table>
<thead>
<tr>
<th>CB</th>
<th>SegSNR(dB)</th>
<th>LSD(dB)</th>
<th>PESQ(MOS-LQO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>SA</td>
<td>SX</td>
<td>SA</td>
</tr>
<tr>
<td>4</td>
<td>8.503</td>
<td>8.377</td>
<td>5.680</td>
</tr>
<tr>
<td>8</td>
<td>7.898</td>
<td>8.134</td>
<td>5.294</td>
</tr>
<tr>
<td>16</td>
<td>7.871</td>
<td>8.251</td>
<td>5.019</td>
</tr>
<tr>
<td>32</td>
<td>7.765</td>
<td>8.227</td>
<td>4.829</td>
</tr>
<tr>
<td>64</td>
<td>7.780</td>
<td>8.314</td>
<td>4.700</td>
</tr>
<tr>
<td>128</td>
<td>7.751</td>
<td>8.301</td>
<td>4.647</td>
</tr>
<tr>
<td>256</td>
<td>7.760</td>
<td>8.279</td>
<td>4.592</td>
</tr>
</tbody>
</table>

3.2. Performance of the Proposed ABE System

We implement and test the system as explained in section 2 on both SA and SX collections. Table 2 and Table 3 respectively presents objective evaluations of the proposed ABE system for SA and SX collections. In these tables, the objective metrics, which are calculated without dimension reduction in the linear estimation are presented for each branch size in the temporal clustering and for two possible temporal neighborhoods $T = 0$ and $T = 1$. Note that without dimension reduction in linear estimation, estimator orders are $p = 10$ and $p = 30$ for temporal neighborhoods $T = 0$ and $T = 1$, respectively. The performance of the proposed ABE system is observed significantly better than the VQ-based benchmark system in LSD and PESQ metrics. We observed some improvements, especially with the LSD metric, when we increased the temporal neighborhood size to $T = 1$. Another common observation is that performance of the proposed ABE system improves significantly as number of branches in the temporal clustering model increases. As we increase $B$ from 4 to 256, LSD decreases from 3.9 dB level to 3.3 dB level both for the SA and SX collections. Similarly, PESQ scores improve as branch size $B$ increases and they are superior to those of the benchmark system. On the other hand, the presence of temporal neighborhood in linear estimation do not have a significant effect on the objective performance of the system in terms of PESQ score and Segmental SNR.

Table 2. Performance of the proposed method for the SA collection.

<table>
<thead>
<tr>
<th>$B$</th>
<th>SegSNR(dB)</th>
<th>LSD(dB)</th>
<th>PESQ(MOS-LQO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>$T=1$</td>
<td>$T=0$</td>
<td>$T=1$</td>
</tr>
<tr>
<td>4</td>
<td>7.131</td>
<td>7.085</td>
<td>3.844</td>
</tr>
<tr>
<td>8</td>
<td>7.494</td>
<td>7.509</td>
<td>3.708</td>
</tr>
<tr>
<td>16</td>
<td>7.393</td>
<td>7.312</td>
<td>3.525</td>
</tr>
<tr>
<td>32</td>
<td>7.481</td>
<td>7.664</td>
<td>3.439</td>
</tr>
<tr>
<td>64</td>
<td>7.323</td>
<td>7.502</td>
<td>3.383</td>
</tr>
<tr>
<td>128</td>
<td>7.344</td>
<td>7.553</td>
<td>3.342</td>
</tr>
<tr>
<td>256</td>
<td>7.343</td>
<td>7.603</td>
<td>3.314</td>
</tr>
</tbody>
</table>

Table 3. Performance of the proposed method for the SX collection.

<table>
<thead>
<tr>
<th>$B$</th>
<th>SegSNR(dB)</th>
<th>LSD(dB)</th>
<th>PESQ(MOS-LQO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>$T=1$</td>
<td>$T=0$</td>
<td>$T=1$</td>
</tr>
<tr>
<td>4</td>
<td>8.487</td>
<td>8.831</td>
<td>3.830</td>
</tr>
<tr>
<td>8</td>
<td>8.281</td>
<td>8.182</td>
<td>3.651</td>
</tr>
<tr>
<td>16</td>
<td>8.247</td>
<td>8.164</td>
<td>3.498</td>
</tr>
<tr>
<td>32</td>
<td>8.683</td>
<td>8.858</td>
<td>3.440</td>
</tr>
<tr>
<td>64</td>
<td>8.314</td>
<td>8.405</td>
<td>3.343</td>
</tr>
<tr>
<td>128</td>
<td>8.525</td>
<td>8.770</td>
<td>3.250</td>
</tr>
<tr>
<td>256</td>
<td>8.280</td>
<td>8.449</td>
<td>3.202</td>
</tr>
</tbody>
</table>

As discussed in section 2.2, we performed dimension reduction prior to extraction of linear estimators by selecting the highest $p$ correlations in $R_{iy}$ for each state $s_k$ and for each target component $f_{k,i}$. We evaluate the effect of the estimator order $p$ in terms of LSD metric. The LSD performance for varying estimator orders with temporal neighborhood $T = 0$ is given in Fig. 2. Note that, the best LSD with the benchmark system is around 4.6 dB at codebook size 256, which is significantly higher than the LSD distances of the proposed system. Choosing a smaller $p$ value reduces the order of the filter, hence diminishes memory concerns, at the cost of a slightly degraded LSD.

In order to better visualize the dimension reduction process, Fig. 3 illustrates the $R_{iy}$ correlation for a sample state in the training models. Higher correlations are darker regions. Columns represent the narrowband LSF features and rows represent wideband LSF features. When the temporal neighborhood is three frames, i.e. $T = 1$, collection of three rows represent temporal neighborhood, where the center row represents the current time frame. Note that, for low frequency wideband LSF features we observe a linear correlation with narrowband LSF features. Furthermore, the high frequency
Fig. 2. The LSD performance for varying $p$ and $B$ values over the SX collection.

Wideband LSF features are observed to correlate mostly with again a collection of high frequency narrowband LSF features.

Fig. 3. Sample $R_{xy}$ correlation matrix visualizations. Columns represent the narrowband LSF features and rows represent wideband LSF features. The intensity of the gray level represents higher correlation.

3.3. Subjective Test Evaluation

We have performed a subjective A/B comparison test to evaluate the proposed ABE system. During the test, the subjects are asked to indicate their preference for each given A/B test pair of 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 subjective A/B test includes 11 listeners, who compared 30 sentence pairs randomly chosen from our database. Of these 30, 3 pairs compared the proposed method with the narrowband version, 3 compared the proposed method with the benchmark system, 3 compared the proposed method with the original wideband version and 18 compared the proposed method with itself for varying $p$ and $T$ values. The final 3 pairs were identical.

Test results, given in Table 4, indicate that, even though not superior to original wideband speech, speech synthesized with the proposed method outperforms the narrowband speech and the speech synthesized with the benchmark system significantly. Moreover, changing the feature selection order $p$ or including temporal neighborhood in the estimation does not result in considerably audible differences.

4. DISCUSSION

We introduce a novel spectral envelope estimation for the artificial bandwidth extension problem. Our method introduces temporal clustering of spectral features and extracts linear estimators within temporal clusters to reconstruct wideband spectra from narrowband spectra. In experimental evaluations we present both objective and subjective results indicating that wideband spectral envelope reconstruction is significantly better than VQ based benchmark system, and the synthesized wideband speech is superior to the narrowband speech.

5. REFERENCES