EVALUATION OF A SPEECH BANDWIDTH EXTENSION ALGORITHM BASED ON VOCAL TRACT SHAPE ESTIMATION

Itai Katsir, David Malah and Israel Cohen

Department of Electrical Engineering, Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
email: {kaziri@tx,malah@ee,icohen@ee}.technion.ac.il

ABSTRACT

In this paper, we evaluate a speech bandwidth extension (BWE) algorithm which involves phonetic and speaker dependent estimation of the high-band part of the spectral envelope. The BWE algorithm extracts speech phoneme information by using a hidden Markov model. Speaker vocal tract shape information corresponding to the wideband signal is extracted by a codebook search. Postprocessing of the estimated vocal tract shape using iterative tuning allows artifacts reduction in cases of erroneous estimation of speech phoneme or vocal tract shape. We present objective measurements results demonstrating the benefit of the iterative tuning. Subjective listening tests illustrate improved wideband quality in comparison to the input narrowband speech. The algorithm complexity is also analyzed.

Index Terms— Bandwidth extension, speech processing, vocal tract area function, sensitivity function, MUSHRA.

1. INTRODUCTION

Current public switched telephone networks (PSTN) limit the bandwidth of the speech signal to 0.3-3.4 kHz. This narrowband (NB) limitation results in degradation of speech quality. Wideband (WB) speech signal with bandwidth limitation of 0.05-7 kHz achieves high quality speech. A BWE algorithm estimates the WB speech signal by artificially extending the NB speech signal to high-band (HB) frequencies from 3.4 kHz to 7 kHz [1]. This technique is transparent to the transmitting network, as it is implemented only at the receiving end. The estimation of the HB spectral envelope and its gain is the most crucial stage for a high quality BWE algorithm [2, 3]. The HB extension of the spectral envelope aims to enhance speech quality, as well as intelligibility. The HB spectral envelope gain may affect the level of artifacts, interpreted as quality degradation.

Recently, we have presented a BWE approach using phonetic and speaker dependent information for HB spectral envelope estimation [3]. The first estimation step employs a hidden Markov model (HMM) to classify each speech frame to a specific phoneme type. The second step finds a speaker specific WB spectral envelope by WB vocal tract area function (VTAF) shape estimation from the calculated NB VTAF shape. A postprocessing step, involving iterative modification of the estimated WB VTAF, allows better gain adjustment and smoothing in time of the estimated spectral envelope. A preliminary evaluation of the algorithm was presented in [3]. It included the log spectral distance (LSD) of the estimated HB power spectrum and the estimation error of the HB estimated formant frequencies.

This research was supported by the Israel Science Foundation (grant no. 1130/11).

In this paper, we further evaluate the performance of the BWE approach presented in [3]. A spectral distortion measure (SDM) [1] is used to evaluate the importance of the iterative postprocessing step for quality improvement. A formal subjective listening test indicates the significance of the algorithm in enhancing the input NB speech. Complexity analysis of the algorithm allows to evaluate the feasibility of the algorithm implementation in real-time applications.

The paper is organized as follows. In Section 2, we summarize the BWE algorithm proposed in [2, 3]. In Section 3, we present the experimental evaluation results. Finally, in Section 4, we draw our conclusions.

2. BWE ALGORITHM OVERVIEW

In this section, we describe the method for estimating the WB speech signal from the input NB signal. The general BWE algorithm scheme is presented in Fig. 1. The system can be divided into four stages which are described in the following subsections.

2.1. Preprocessing and Feature Extraction

Stage I carries out preprocessing and feature extraction. The received NB speech signal, $s_{\mathrm{NB}}(n)$ with sample index n, is upsampled to 16 kHz sampling rate and filtered through a low pass filter with 4 kHz cutoff frequency and 10 dB boost at 300 Hz. This equalization adds naturalness to the NB signal. Three sets of features are extracted from the upsampled and equalized speech frame. The first feature vector, $\mathbf{x_1}$, contains spectral information including Mel-frequency cepstral coefficients. Its purpose is to allow good separation of different speech classes that give different HB spectral envelope shapes. The second feature vector, $\mathbf{x_2}$, contains the area coefficients that represent the speaker's VTAF shape, which is used for WB VTAF estimation. The last extracted feature vector, $\mathbf{x_3}$, is the NB excitation, which is used for WB excitation generation.

2.2. WB Spectral Envelope Estimation

In Stage II of the algorithm, the estimation of the WB spectral envelope $\hat{\phi}_{\mathrm{WB}}(k)$, with frequency index k, is performed. It is calculated in a three-step process.

2.2.1. Speech State Estimation

In the first step, the speech state which represents a specific speech phoneme is estimated using an HMM-based statistical model. The HMM statistical model was trained offline using the TIMIT transcription. Each frame was associated with a state $S_i(m)$, $i = 1, \ldots, N_s$, which represents a speech phoneme (one state for each

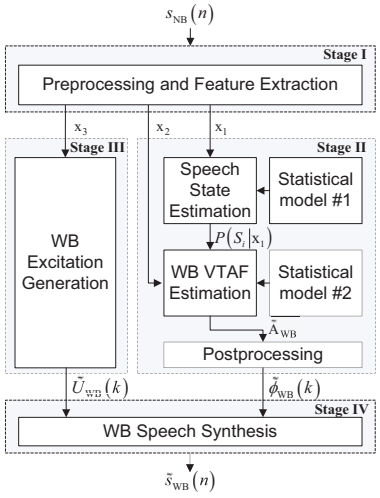

Fig. 1. Block diagram of the proposed BWE algorithm.

phoneme), where i is the state index, N_s is the number of states and m is the current frame time index. From the training database, the state S_i and the feature vector $\mathbf{x_1}$ of each speech frame were extracted. The following probability density functions (PDFs) were calculated:

- $p(S_i)$ - Initial probability of each state.
- $p(S_i|S_j)$ - Transition probability from state j to state i.
- $p(\mathbf{x_1}|S_i)$ - Observation probability for each state, approximated by Gaussian mixture model (GMM) parameters with N_g mixtures.

In the application phase, the state probabilities for an input speech frame are extracted from the a-posteriori PDF. We denote the observation sequence of the first feature vector $\mathbf{x_1}$ up to the current frame as $\mathbf{X_1}(m) = \{\mathbf{x_1}(1), \mathbf{x_1}(2), \ldots, \mathbf{x_1}(m)\}$. The conditional probability $p(S_i(m)|\mathbf{X}_1(m))$ expresses the a-posteriori probability. It is recursively calculated for each state by

$$p(S_i(m)|\mathbf{X}_1(m)) = C_1 \cdot p(\mathbf{x}_1(m)|S_i(m)) \cdot \sum_{j=1}^{N_s} p(S_i(m)|S_j(m-1))\, p(S_j(m-1)|\mathbf{X}_1(m-1)) \quad , \tag{1}$$

where C_1 is a normalization factor. The calculated state probabilities are used for WB VTAF estimation in the following step.

2.2.2. *Wideband VTAF Estimation*

The WB VTAF, $\tilde{\mathbf{A}}_{\mathbf{WB}}$, is estimated in the second step, from the calculated NB VTAF. We use a second statistical model that incorporates a set of WB VTAF codebooks (CBs). For each of the N_s states, we have a CB with N_{CB} entries. The CBs were trained offline with real WB VTAF data, extracted from the TIMIT train database. We denote the calculated NB VTAF as $\mathbf{A_{NB}}$ and the CB entries corresponding to the estimated state S_i as $\mathbf{A}_{\mathbf{WB}}^{S_i}(j)$, $j = 1, \ldots, N_{\mathrm{CB}}$. The optimal WB VTAF $\tilde{\mathbf{A}}_{\mathbf{WB}}^{S_i}$ for the estimated state in frame m is picked by minimizing the Euclidean distance between $\mathbf{A_{NB}}$ and $\mathbf{A}_{\mathbf{WB}}^{S_i}(j)$, $j = 1, \ldots, N_{\mathrm{CB}}$:

$$\tilde{\mathbf{A}}_{\mathbf{WB}}^{S_i} = \mathbf{A}_{\mathbf{WB}}^{S_i}(j^{opt}),$$
$$j^{opt} = \arg\min_{j=1}^{N_{\mathrm{CB}}} \left\| log(\mathbf{A_{NB}}(m)) - log\left(\mathbf{A}_{\mathbf{WB}}^{S_i}(j)\right) \right\|_2^2 . \tag{2}$$

In-order to reduce artifacts due to erroneous state estimation, we use N_{best} states with the highest a-posteriori probability $p_1, \ldots, p_{N_{\mathrm{best}}}$ for WB VTAF estimation. This is done by a weighted average of the corresponding values from (2).

2.2.3. *Postprocessing*

Postprocessing of the estimated WB VTAF, in the third step, allows better gain adjustment and smoothing in time of the estimated WB spectral envelope. Better gain adjustment is achieved by fitting the lower band of the estimated WB spectral envelope to the calculated NB spectral envelope. Better smoothness in time is achieved by reducing time discontinuities of the estimated WB spectral envelopes.

We denote the formant frequencies of the NB and the estimated WB spectral envelopes by $\mathbf{f_{NB}}$ and $\tilde{\mathbf{f}}_{\mathbf{WB}}$, respectively. The shape fitting of the estimated WB spectral envelope is conducted by tuning the lower subset of $\tilde{\mathbf{f}}_{\mathbf{WB}}$ to $\mathbf{f_{NB}}$. The tuning is done iteratively by perturbing the WB VTAF area coefficients [2, 3]. The VTAF is perturbed by using a sensitivity function [6]. The sensitivity function relates small changes in VTAF to changes in formant frequencies. We denote the VTAF values by A_{n_A}, $n_A = 1, \ldots, N_A$, where N_A is the number of area coefficients. The spectral envelope formant frequencies are denoted by f_{n_f}, $n_f = 1, \ldots, N_f$, where N_f is the number of formant frequencies. The sensitivity function S_{n_f, n_A} relates a small change in f_{n_f} to incremental changes in the area coefficients, via:

$$\frac{\Delta f_{n_f}}{f_{n_f}} = \sum_{n_A=1}^{N_A} S_{n_f, n_A} \frac{\Delta A_{n_A}}{A_{n_A}} \quad . \tag{3}$$

Here we set Δf_{n_f} to be the difference between the desired formant frequency and the current formant frequency. Thus, ΔA_{n_A} is the needed perturbation in the value of area coefficient number n_A.

The goal of each iteration is to minimize the difference between the calculated and the estimated NB formant frequencies. The stopping condition for the iterative process is the reaching of an allowed deviation, $\mathbf{\Delta f_d}$, between $\mathbf{f_{NB}}$ and the corresponding lower subset of $\tilde{\mathbf{f}}_{\mathbf{WB}}$. No improvement in the frequencies deviation may imply a convergence problem and a large estimation error of the spectral shape. Hence, the estimated WB VTAF is updated only when the average frequencies deviations in the current iteration is smaller than that of the previous update. On average, 3.6 iterations were performed for each processed frame using $\mathbf{\Delta f_d} = 50$ Hz. About 30% of the frames were processed using only one iteration. About 45% of the frames needed two to four iterations.

Next, smoothing in time is performed on the estimated WB VTAF under the assumption of physical continuity of vocal tract shape in time. Gain adjustment is performed by first converting the smoothed estimate of the WB VTAF to a WB spectral envelope. The calculated WB spectral envelope can now be gain-adjusted to match the energy of the input NB spectral envelope in its lower band.

2.3. WB Excitation Generation

In Stage III of the algorithm, the WB excitation, $\tilde{U}_{\mathrm{WB}}(k)$ is generated. The HB excitation is generated using a simple spectral copy of the calculated NB excitation.

2.4. WB Speech Synthesis

In the last stage of the algorithm, Stage IV, the output WB speech signal $\tilde{s}_{\mathrm{WB}}(n)$ is synthesized. The estimated final HB spectral envelope is used to shape the generated excitation in the frequency

domain. This provides a HB speech component that is then concatenated in the frequency domain to the original NB signal to create the estimated WB signal.

3. PERFORMANCE EVALUATION

To evaluate the algorithm performance, objective and subjective quality measurements were used. The BWE algorithm was implemented using Matlab®. The following parameters were used: number of states $N_s = 61$ (symbols in the TIMIT lexicon), number of Gaussians per state $N_g = 16$ (as in [4]), number of CB entries per state $N_{\mathrm{CB}} = 16$, number of VTAF area coefficients $N_A = 16$ (as in [5]), and number of states for VTAF estimation $N_{\mathrm{best}} = 5$. The TIMIT WB training database, including 4620 sentences, was used for training both the HMM and the CB statistical models. The TIMIT WB test database, including 1680 sentences, was used as an input to the proposed algorithm after being preprocessed by a telephone channel filter and down-sampled to 8 kHz. From the BWE processed signals and their original WB counterparts the following quality measurements were performed.

3.1. Objective Evaluation

In order to evaluate the significance of the iterative postprocessing step for quality improvement, the spectral distortion measure (SDM) [1] was used. This measure is a nonsymmetric weighted LSD. The distortion is generally calculated by using a decaying exponential for increasing frequencies and by giving higher penalty for spectral over-estimation than for under-estimation. The SDM measure for the m^{th} frame is calculated by:

$$\mathrm{SDM}_m = \frac{1}{k_{\mathrm{high}} - k_{\mathrm{low}} + 1} \sum_{k=k_{\mathrm{low}}}^{k_{\mathrm{high}}} \xi_m(k) \ , \tag{4}$$

where the distortion is calculated using the fast Fourier transform (FFT) bin indices from k_{low} to k_{high} and $\xi_m(k)$ is calculated as:

$$\xi_m(k) = \begin{cases} \Delta_m(k) \cdot \exp\{\alpha \Delta_m(k) - \beta k\} & \text{, if } \Delta_m(k) \geq 0 \\ \ln(-\Delta_m(k) + 1) \cdot \exp\{-\beta k\} & \text{, else} \end{cases}$$

$$\Delta_m(k) = 10 \log_{10} \frac{\tilde{\phi}_m(k)}{\phi_m(k)} \ ,$$

where α and β are the weighting factors, ϕ_m is the spectral envelope of the original WB frame, and $\tilde{\phi}_m$ is the spectral envelope of the corresponding BWE frame.

The motivation for using this measurement is the fact that high-frequency distortion is less significant for human perception. Another important issue that this measure deals with is giving more weight to the estimated spectrum above the magnitude of the original one. Overestimated HB energy leads to undesirable audible artifacts that in the opposite case (of underestimation) does not cause any artifacts [7].

The SDM measure was computed between the original WB spectral envelopes and the estimated spectral envelopes, with and without the iterative postprocessing step. This measure was taken only for voiced estimated frames using the HMM phoneme estimation output. The SDM parameters were set to $\alpha = 0.1$ and $\beta = 5$. The distortion was calculated using the FFT bin indices from k_{low} to k_{high}, corresponding to the frequency range from 3.4 to 8 kHz.

The mean SDM and LSD of the entire test database are presented in Table 1. It can be seen that the iterative postprocessing step improves the quality of the estimated spectral envelope and hence reduce the SDM. It is also noticeable from the SDM and LSD results that the achieved improvement of the iterative process is clearly seen in the SDM results and barely seen in the LSD results. This means that the postprocessing step reduces the spectral envelope over-estimation and reduces distortion in the low part of the HB frequencies which are more significant for human perception.

Table 1. Average SDM and LSD of estimated spectral envelope with and without the iterative postprocessing step.

Measured	SDM [dB]	LSD [dB]
Without iterative process	13.64	9.98
With iterative process	9.89	9.91

3.2. Subjective Evaluation

The chosen subjective measure in this research is the multistimulus test with hidden reference and anchors (MUSHRA) test [8]. In this test a person grades the processed speech test sentences in comparison to a reference sentence. The test sentences include all the sentences under test, an anchor sentence which should have the lowest quality compared to the reference sentence, and a hidden reference sentence similar to the reference sentence. This hidden reference is used for post-screening of subjects that gave a low grade to the hidden reference. All the test sentences can be replayed by the listener at will. The main advantage of the MUSHRA test over the mean opinion score (MOS) test is that it is easier to perform, as it requires fewer participants to obtain statistically significant results [2]. This test also allows finer measurements of small differences because of the 0-100 score scale.

The MUSHRA test was performed by 11 listeners. The test included 6 different experiments, each with a different English sentence, 3 by male and 3 by female. Every experiment included multiple conditions of the sentence: a WB reference speech signal, a NB anchor speech signal, the proposed BWE speech signal and a reference BWE speech signal. The reference BWE speech signal was based on the algorithm from [4] with some unpublished improvements made by Bernd Geiser until 2010. In the test, the listeners compared multiple conditions of a sample at the same time, and could repeat the samples. The listeners could also repeat the reference when they wanted. The test produced results for the conditions between 0 and 100, with 100 being same quality as the reference speech signal.

The results of the MUSHRA test are presented in Fig. 2. The obtained results indicate that the proposed BWE algorithm improves the received NB signal. It also exhibits some improvement over the reference algorithm results.

3.3. Complexity Evaluation

The algorithm was also examined for its complexity. The goal of this examination is to detect the most complex stages in the algorithm. This examination was performed by measuring the Matlab processing time of each major algorithm processing block, running on an Intel CPU at 2.66 GHz clock speed. The results were averaged over the entire speech frames of the TIMIT test database. The distinct blocks are:

- Preprocessing and feature extraction as described in Subsection 2.1.
- State estimation as described in Subsection 2.2.1.
- WB VTAF estimation as described in Subsection 2.2.2.

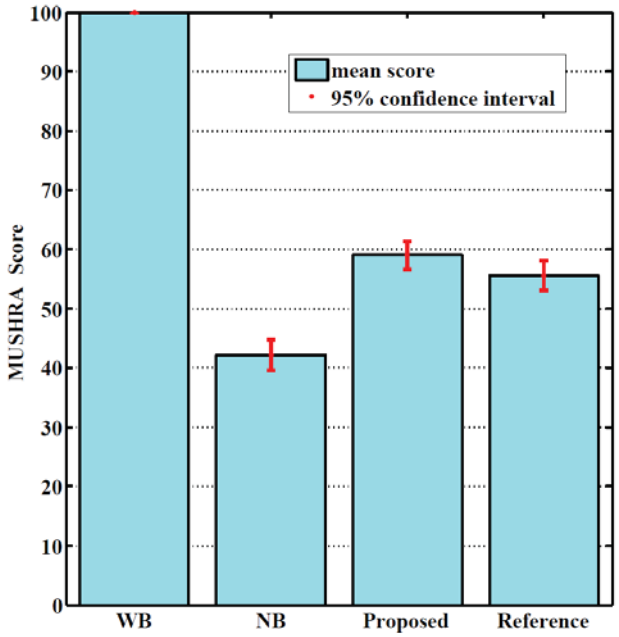

Fig. 2. MUSHRA subjective measure score.

- An iterative tuning process as described in Subsection 2.2.3.
- Gain adjustment as described in Subsection 2.2.3.
- WB excitation generation as described in Subsection 2.3.
- The WB speech synthesis as described in Subsection 2.4.

The obtained results are presented in Table 2. The results indicate the average processing time of a 20 msec speech frame. The results reveal that the HMM-based state estimation step in the WB spectral envelope estimation stage consumes the most processing time. This is mainly because of the GMM probability values calculation. The postprocessing step also exhibits high computation load due to the online sensitivity function calculation, which consumes about 85% of this step processing time.

4. CONCLUSION

We have presented a performance evaluation of the BWE algorithm proposed in [3]. The algorithm was evaluated by objective and subjective measurements. The SDM between the estimated HB spectral envelope and the original HB spectral envelope shows improved results of about 4 dB for frames which were postprocessed using the iterative tuning compared to those that were not. These results illustrate the effectiveness of the iterative tuning process in reducing estimation artifacts and especially in reducing gain over-estimation. The MUSHRA subjective listening tests show improved quality of the enhanced speech. An improvement of more than 15 points compared to the input NB speech illustrates the advantage of using the proposed BWE algorithm when using telephone networks that limit speech bandwidth to the NB frequency range. The proposed BWE algorithm could be used to improve the call quality in the period of transition to WB supported networks. Complexity evaluation of the algorithm main building blocks showed the high complexity of the phoneme estimation step and the iterative processing block of the postprocessing step, both in the WB spectral envelope estimation stage. The high complexity of the state probabilities calculation at the state estimation step and the online sensitivity function calculation at the postprocessing step, is one of the main algorithm drawbacks.

Table 2. Average processing time of a 20 msec speech frame of main BWE algorithm processing blocks.

Algorithm Processing Block	Computation Time [msec]
Preprocessing and feature extraction	1.27
State estimation	19.39
WB VTAF estimation	0.59
Postprocessing (iterative process)	7.69
Postprocessing (gain adjustment)	0.36
WB excitation generation	0.04
WB speech synthesis	0.57
Total	29.91

Future work might include an offline calculation of the sensitivity function for each WB VTAF codeword. This will reduce the online computational complexity. Using normalized formant frequency deviation for the iterative process, might reduce the needed number of iterations and as a results the processing time, while maintaining the same postprocessing quality. Using the postprocessing iterative procedure for better refinement and control of estimated spectral envelope by HB formants tuning to past estimated HB formants, could improve the smoothing in time of the estimated HB spectral envelope and further improve the speech quality. The algorithm should also be evaluated using formal listening tests under different background noise conditions and with different languages. This evaluation would determine the algorithm robustness to noisy environments and multiple languages.

5. REFERENCES

[1] B. Iser, W. Minker and G. Schmidt, *Bandwidth extension of speech signals*. Lecture Notes in Electrical Engineering, vol. 13, Springer, 2008.

[2] I. Katsir, "Artificial Bandwidth Extension of Band-Limited Speech Based on Vocal-Tract Shape Estimation," M.Sc. thesis, Technion, Israel Institute of Technology, Dec, 2011, http://sipl.technion.ac.il/siglib/FP/itai_katsir.pdf

[3] I. Katsir, I. Cohen, and D. Malah, "Speech bandwidth extension based on speech phonetic content and speaker vocal tract shape estimation," in *Proc. EUSIPCO 2011*, Barcelona, Spain, Aug. 2011, pp. 461–465.

[4] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, August 2003.

[5] D. Malah, "Method of bandwidth extension for narrow-band speech," Patent number: US 6988066 B2, Jan 2006.

[6] B. Story, "Technique for "tuning" vocal tract area functions based on acoustic sensitivity functions," *J. Acoust. Soc. Amer*, vol. 119, pp. 715–718, 2006.

[7] M. Nilsson and W. B. Kleijn, "Avoiding over-estimation in bandwidth extension of telephony speech," in *Proc. ICASSP 2001*, Salt Lake City, UT, USA, May 2001, pp. 869–872.

[8] ITU-R. Method for the subjective assessment of intermediate quality level of coding systems. Recommendation ITU-R BS.1534-1, 2003, http://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-1-200301-I!!PDF-E.pdf