SPEECH BANDWIDTH EXTENSION BASED ON SPEECH PHONETIC CONTENT AND SPEAKER VOCAL TRACT SHAPE ESTIMATION

Itai Katsir, Israel Cohen and David Malah

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

ABSTRACT

In this paper, we introduce a new speech bandwidth extension (BWE) algorithm which involves phonetic and speaker dependent estimation of the high-band part of the spectral envelope. Speech phoneme information is extracted by using a hidden Markov model. Speaker vocal tract shape information corresponding to the wideband signal is extracted by a codebook search. The proposed method allows better estimation of high-band formant frequencies, especially for voiced sounds, and better estimation of spectral envelope gain, especially for unvoiced sounds. Postprocessing of the estimated vocal tract shape allows artifacts reduction in cases of erroneous estimation of speech phoneme or vocal tract shape. We present experimental results that demonstrate improved wideband quality for different speech sounds in comparison to other BWE methods.

1. INTRODUCTION

Current public switched telephone networks (PSTN) limit the bandwidth of the speech signal to 0.3-3.4kHz. This narrowband (NB) limitation results in degradation of speech quality. One way to achieve high quality speech is by applying a wideband (WB) coding solution. WB coders expand the coded speech bandwidth to 0.05-7 kHz. Unfortunately, this solution requires an expensive network upgrade. A possible solution for the transition period to WB speech supporting networks, is to artificially extend 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.

Most BWE algorithms use the source-filter model of speech production. This model considers the speech signal as being produced by a spectrally flat excitation source that passes through an auto-regressive (AR) filter [2]. This model suggests separation of the BWE algorithm into two independent tasks of HB excitation and HB spectral envelope estimation [1].

The estimation of the HB spectral envelope and its gain is the most crucial stage for a high quality BWE algorithm. 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. Hence, most recent BWE algorithms use different techniques to map NB speech features to HB features that represent the HB spectral envelope and gain. These techniques include codebook mapping [3, 4], linear mapping [5, 6], Neural Networks [7] and statistical methods by Gaussian mixture models (GMM) and hidden Markov models (HMM) [8, 9]. These techniques still face problems with spectral envelope estimation of some speech sound classes, especially unvoiced sounds. They also show quality variations for different speakers and some hissing and whistling artifacts due to gain overestimation and discontinuities in the time evolution of the estimated spectral envelope.

One method to improve HB spectral envelope estimation is to incorporate speech-sound class information in the estimation process. This information is especially crucial for better estimation of unvoiced sounds, which are characterized by low NB energy while having high HB energy. Voiced and unvoiced sounds classification is used in [4] for codebook mapping. Voiced sounds, sibilants and stop-consonants classes are used in [7] for HB spectral shape estimation. Phonetic transcription is used in [10] for supervised training of an HMM statistical model.

Another way to improve the HB spectral envelope estimation is by making it robust to variation of speakers. Different speakers yield different formants locations even when representing the same speech linguistic content [2]. Using speaker related features such as vocal tract area function (VTAF) allows a better speaker-dependent estimation. The VTAF represents the vocal tract's physical shape as a function of the distance from the glottis. The concatenated tube model is used for VTAF shape representation [2, 11]. The algorithm in [5] uses this model to estimate the formants locations in the HB for voiced sounds. In [12, 13] the WB spectral envelope is obtained from an estimation of the WB VTAF. As speech is produced by a physical system modeled by the VTAF, the estimation of the WB VTAF is done in [12, 13] by interpolating the NB VTAF.

Gain estimation is possible as part of the HB spectral envelope estimation, like in [6], or by WB spectral envelope estimation and gain adjustment to the received NB spectral envelope, as in [3].

In this paper, we present a BWE approach using phonetic and speaker dependent information for HB spectral envelope estimation. The first step employs an HMM model to classify each speech frame to a specific phoneme type. The second step finds a speaker specific WB spectral envelope by WB VTAF shape estimation from the calculated NB VTAF shape. A new proposed postprocessing step, involving modification of the estimated WB VTAF, allows better gain adjustment and smoothing in time of the estimated spectral envelope.

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

2. PROPOSED BWE ALGORITHM

The proposed method for the estimation of the WB speech signal is described in this section. A general system overview

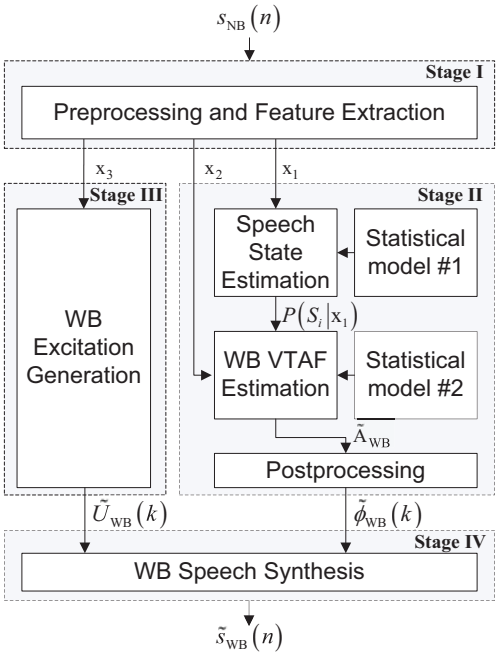

Figure 1: Block diagram of the proposed BWE algorithm.

and a block diagram are given first, followed by a detailed description of the algorithm stages.

2.1 Algorithm Overview

The general BWE algorithm scheme is described in Fig. 1. The system can be divided into four stages. Stage I carries out preprocessing and feature extraction. The input to this stage is the received NB speech signal $s_{\mathrm{NB}}(n)$ with sample index n. The NB speech signal is framed, upsampled and equalized in the low frequencies. Three sets of feature vectors are extracted from each preprocessed frame of the received signal: Frequency-based features, $\mathbf{x_1}$, for speech-state estimation; NB VTAF feature vector, $\mathbf{x_2}$, for WB VTAF estimation, and NB excitation, $\mathbf{x_3}$, for WB excitation generation.

In Stage II of the algorithm, the estimation of the WB spectral envelope $\tilde{\phi}_{\mathrm{WB}}(k)$, with frequency index k, is performed. It is calculated in a three-step process. In the first step, speech state estimation yields the probability of being in a specific speech-phoneme related state. The WB VTAF, $\tilde{\mathbf{A}}_{\mathbf{WB}}$, is then estimated, in the second step, from the calculated NB VTAF. Postprocessing of the estimated WB VTAF, in the third step, allows better gain adjustment and smoothing in time of the estimated WB spectral envelope.

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. In the last stage of the algorithm, Stage IV, the output WB speech signal $\tilde{s}_{\mathrm{WB}}(n)$ is synthesized in the frequency domain, without changing the received NB signal.

2.2 Detailed Algorithm Description

2.2.1 Stage I: Preprocessing and Feature Extraction

The received NB speech signal is segmented into frames of 20msec duration, with 10msec overlap between frames. The speech frame is upsampled to 16 kHz sampling rate and filtered through a LPF with 4 kHz cutoff frequency and 10dB boost at 300 Hz. The 10dB boost equalizes a typical telephone channel filter response [1], which attenuates the speech signal at and below 300 Hz. This equalization adds naturalness to the NB signal. The equalized frame is then windowed using a Hamming window.

Three sets of features are extracted from the upsampled and equalized speech frame. The purpose of the first feature vector, $\mathbf{x_1}$, is to allow good separation of different speech classes that give different HB spectral envelope shapes [14]. The feature vector $\mathbf{x_1} \in \mathbb{R}^{13}$ consists of the following features:

- *Mel Frequency Cepstral Coefficients* (MFCC) of nine subbands from 300 to 3400Hz. The MFCC are commonly used in speech recognition algorithms. They were shown to have high NB to HB speech mutual information and to provide good class separation [8].
- *Spectral centroid* of the NB power spectrum, which is generally high for unvoiced sounds [15].
- A *spectral flatness* measure [15] that indicates the tonality of the speech signal.
- *Spectral slope*, which is useful for discriminating voiced frames from sibilants and plosives [7].
- *Normalized frame energy* [9].

The second feature vector, $\mathbf{x_2}$, contains the area coefficients that represent the speaker's VTAF shape. The area coefficients are calculated from the reflection coefficients as described in [2]. Since the preprocessed NB frame is upsampled to 16 kHz, we use $N_A = 16$ area coefficients for NB VTAF calculation.

The last extracted feature vector, $\mathbf{x_3}$, is the NB excitation, which is calculated in the frequency domain by dividing the NB signal spectrum by the NB spectral envelope signal. A fast Fourier transform (FFT) of length 512 is used for frequency domain signals calculation from the upsampled frame.

2.2.2 Stage II: WB Spectral Envelope Estimation

The estimation of the WB spectral envelope is carried out in three steps. In the first step, the speech state which represents a specific speech phoneme is estimated using an HMM-based statistical model. The second step consists of estimating the WB VTAF shape by a codebook search, using the calculated NB VTAF shape. Postprocessing of the estimated WB VTAF is conducted in the last step, to reduce possible artifacts due to estimation errors in the previous steps.

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, where i is the state index, N_s is the number of states and m is the current frame time index. The following probability density functions (PDFs) were calculated:

- $p(S_i)$ - Initial probability of each state.
- $p\left(S_i(m) \middle| S_j(m-1)\right)$ - Transition probability of the Markov chain from state j to state i.
- $p(\mathbf{x_1}|S_i)$ - Observation probability for each state. This probability is approximated by GMM parameters with N_g mixtures, which are estimated for each state by the expectation-maximization (EM) algorithm [9].

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

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

where C_1 is a normalization factor to allow all the state probabilities to sum up to one [10]. Choosing the state with the highest a-posteriori probability yields a hard-decision for the current speech frame linguistic content.

Now, we wish to estimate a suitable WB spectral envelope for the estimated speech state. For this purpose we estimate the speaker's VTAF shape. As the VTAF shape models the physical speech production system, we wish to find the closest WB VTAF shape to the calculated NB VTAF shape. 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 using the Linde, Buzo, Gay training (LBG) algorithm [1]. We denote the calculated NB VTAF as $\mathbf{A_{NB}}$ and the CB entries corresponding to the estimated state S_i as $\mathbf{A}_{\mathbf{WB}}^{S_i}(j)$, $j = 1, \ldots, N_{\text{CB}}$. 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 $\mathbf{A}_{\mathbf{WB}}^{S_i}(j)$, $j = 1, \ldots, N_{\text{CB}}$:

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

In-order to reduce artifacts due to erroneous state estimation, we use N_{best} states with the highest a-posteriori probability $p_1, \ldots, p_{N_{\text{best}}}$ for WB VTAF estimation

$$\tilde{\mathbf{A}}_{\mathbf{WB}} = C_2 \cdot \left(p_1 \cdot \tilde{\mathbf{A}}_{\mathbf{WB}}^{S_{i_1}} + \ldots + p_{N_{\text{best}}} \cdot \tilde{\mathbf{A}}_{\mathbf{WB}}^{S_{i_{N_{\text{best}}}}} \right), \quad (3)$$

where C_2 is a normalization factor to constrain the highest N_{best} probabilities to sum up to one.

After this step we have an initial WB VTAF estimation, $\tilde{\mathbf{A}}_{\mathbf{WB}}^{\mathbf{0}}$. This estimated WB VTAF is further processed to allow better spectral envelope gain adjustment and smoothing in time. Better gain adjustment can be achieved by fitting the lower band of the estimated WB spectral envelope to the calculated NB spectral envelope. Better smoothness in time can be achieved by reducing time discontinuities of estimated WB spectral envelopes. The postprocessing step is fully described in the remainder of this sub-section. Fig. 2 presents the block diagram of the proposed postprocessing.

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 [16]. The iterative tuning process is conducted only in voiced speech frames, as those frames are characterized by strong NB formant frequencies.

The VTAF is perturbed by using a sensitivity function. 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} satisfies the following relationship:

$$\frac{\Delta f_{n_f}}{f_{n_f}} = \sum_{n_A=1}^{N_A} S_{n_f, n_A} \frac{\Delta A_{n_A}}{A_{n_A}}, \quad (4)$$

where Δf_{n_f} is the difference between the desired formant frequency and the current formant frequency, and ΔA_{n_A} is the perturbation size of the area number n_A. A vector form of (4) is:

$$\boldsymbol{\Delta}\hat{\mathbf{f}}_{[N_f x 1]} = \mathbf{S}_{[N_f x N_A]} \cdot \boldsymbol{\Delta}\hat{\mathbf{A}}_{[N_A x 1]}, \quad \boldsymbol{\Delta}\hat{\mathbf{f}} \triangleq \left[\frac{\Delta f_1}{f_1}, \ldots, \frac{\Delta f_{N_f}}{f_{N_f}}\right]^T, \ \boldsymbol{\Delta}\hat{\mathbf{A}} \triangleq \left[\frac{\Delta A_1}{A_1}, \ldots, \frac{\Delta A_{N_A}}{A_{N_A}}\right]^T. \quad (5)$$

The sensitivity function is calculated by measuring the formants frequencies deviation due to small area changes using (4).

The goal of each iteration is to minimize the difference between the calculated and estimated NB formant frequencies. The formant frequencies are obtained by spectral envelope peak picking. The VTAF perturbation is solved from (5) by:

$$\boldsymbol{\Delta}\hat{\mathbf{A}}_{[N_A x 1]} = \mathbf{S}^{\dagger}_{[N_A x N_f]} \cdot \boldsymbol{\Delta}\hat{\mathbf{f}}_{[N_f x 1]} \ , \quad (6)$$

where $\mathbf{S}^{\dagger}$ is the Moore-Penrose pseudo-inverse of $\mathbf{S}$. This solution minimizes the ℓ^2 norm of $\boldsymbol{\Delta}\hat{\mathbf{A}}$ when $N_f < N_A$. This criterion allows minimal area changes that give the desired formant frequencies changes. The pseudo-inverse of the sensitivity function matrix is calculated using the singular value decomposition (SVD) technique. Once $\boldsymbol{\Delta}\hat{\mathbf{A}}$ is calculated, the perturbation size for each VTAF area coefficient is $\Delta A_{n_A} = \Delta\hat{A}_{n_A} \cdot A_{n_A}$. A new estimate of the WB VTAF is obtained by:

$$\tilde{\mathbf{A}}_{\mathbf{WB}}^{l+1} = \tilde{\mathbf{A}}_{\mathbf{WB}}^{l} + \boldsymbol{\Delta}\tilde{\mathbf{A}}_{\mathbf{WB}}^{l} \ , \quad (7)$$

where l is the iteration number and $\boldsymbol{\Delta}\tilde{\mathbf{A}}_{\mathbf{WB}} = [\Delta A_1, \ldots, \Delta A_{N_A}]^T$.

The stopping condition for the iterative process is the reaching of an allowed deviation, $\boldsymbol{\Delta}\mathbf{f_d}$, 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 iteration. On average, 3.6 iterations were performed for each processed frame using $\boldsymbol{\Delta}\mathbf{f_d} = 50$ Hz.

When the described iterative process is finished, the result is a WB spectral envelope that its lower band is close to

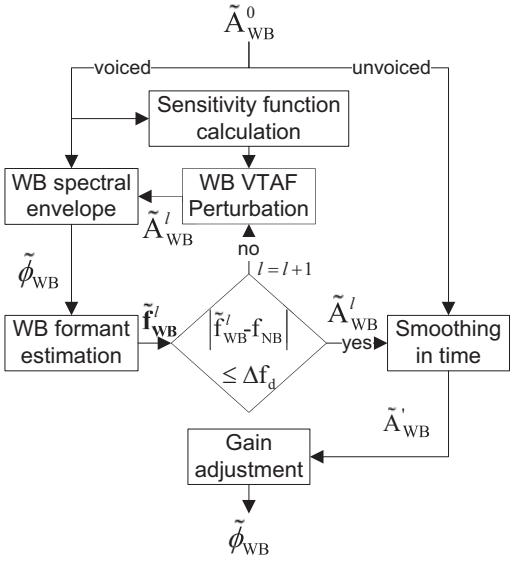

Figure 2: Block diagram of the proposed postprocessing step.

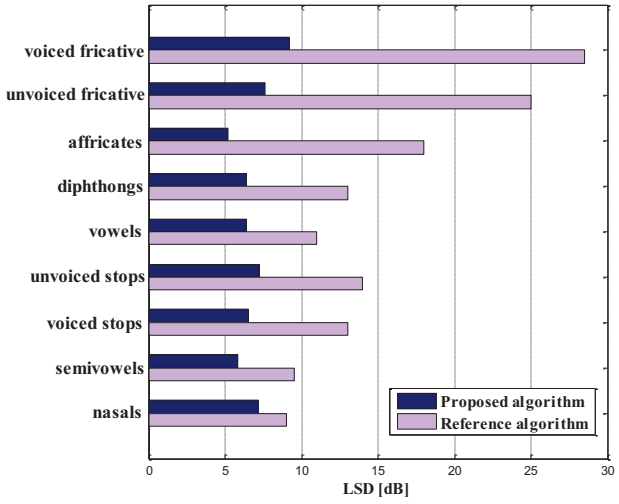

Figure 3: Average log spectral distortion for different phoneme categories.

the NB spectral envelope in terms of its NB formant locations. Now, the estimated WB VTAF shape should be further processed to reduce possible artifacts and further improve the speech quality. This is done by smoothing in time and gain adjustment of the final estimated WB VTAF.

Smoothing in time is performed on the estimated WB VTAF under the assumption of physical continuity of vocal tract shape in time. Smoothing is done recursively by:

$$\tilde{\mathbf{A}}'_{\mathbf{WB}}(m) = \beta \cdot \tilde{\mathbf{A}}'_{\mathbf{WB}}(m-1) + (1-\beta) \cdot \tilde{\mathbf{A}}_{\mathbf{WB}}(m), \quad (8)$$

where $\beta = 0.7$ for voiced frames and $\beta = 0.5$ for unvoiced frames.

Gain adjustment is performed by first converting the smoothed estimate of the WB VTAF to a WB spectral envelope, as described in [2]. The calculated WB spectral envelope can now be gain adjusted to match the energy of the input NB spectral envelope in its lower band [3].

2.2.3 Stage III: WB Excitation Generation

HB excitation generation is based on spectral copying of the NB excitation. The NB excitation in the transition band between 2.2-3.4 kHz is used repeatedly to fill the HB missing frequencies. This simple method allows keeping the original NB excitation signal untouched and filling all the missing HB frequencies with an excitation signal without any gap.

2.2.4 Stage IV: WB Speech Synthesis

The estimated final WB 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. The time-domain speech frame is calculated from the obtained BWE signal transform using the inverse fast Fourier transform (IFFT). Two sequential time frames are combined by the overlap-add method using a Hann synthesis window.

3. PERFORMANCE EVALUATION

To evaluate the algorithm performance, objective quality measurements were used. The proposed algorithm was implemented using the following parameters: number of states $N_s = 61$ (symbols in the TIMIT lexicon), number of Gaussian per state $N_g = 16$ (as in [9]), number of CB entries per state $N_{\text{CB}} = 16$, number of VTAF area coefficients $N_A = 16$ and number of states for VTAF estimation $N_{\text{best}} = 3$. 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 computed.

The first examined criterion was the Log Spectral Distance (LSD) measure in different phonetic categories. The LSD is calculated for the m^{th} frame by:

$$\text{LSD}_m = \sqrt{\frac{1}{k_{\text{high}} - k_{\text{low}} + 1} \sum_{k=k_{\text{low}}}^{k_{\text{high}}} \left[10 \log_{10} \frac{P_m(k)}{\tilde{P}_m(k)} \right]^2}, \quad (9)$$

where P_m is the power spectrum of the original WB frame, and $\tilde{P}_m$ is the power spectrum of the corresponding BWE frame. The distortion is calculated using the FFT bin indices from k_{low} to k_{high}, corresponding to the frequency range from 4 to 7 kHz. The analysis is performed in frames of 256 samples (16 ms) using Hamming windowing, with 50% overlap between successive frames, and an FFT of length 1024.

Our results, in terms of the average LSD over phonemes in a given class, are compared in Fig. 3 to the results obtained in [7]. The results show improved performance of the proposed algorithm for all phoneme classes. A major improvement is obtained for fricative sounds. The results demonstrate the effectiveness of phoneme dependent estimation of BWE speech frames.

The second evaluation criterion is the formant frequencies error between HB estimated formant frequencies and

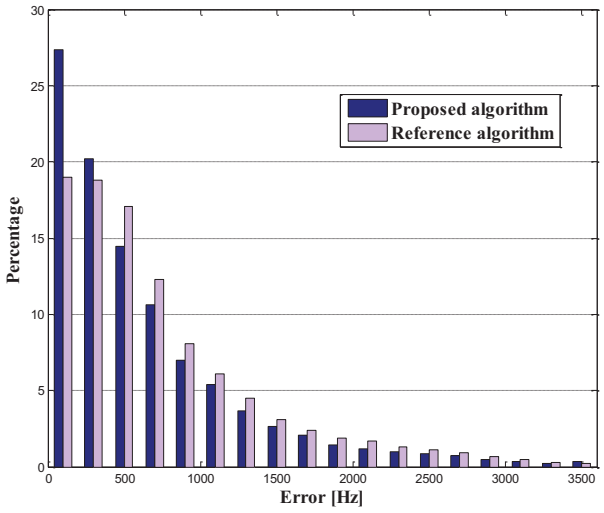

Figure 4: Histogram of estimated formants frequencies error.

their original counterparts. Formants locations for the same phoneme may be different for different speakers. For comparison reason with [5] a linear predictor of order 14 is used and the formant frequencies are derived by peak picking in the spectral envelope. This measure is calculated for voiced frames in all the test database using the TIMIT transcription.

Our results, are compared in Fig. 4 to the results obtained in [5]. Each histogram bin has a width of 200 Hz. The results demonstrate an improvement in formant frequencies estimation using the vocal tract shape modeling and tuning.

4. CONCLUSION

We have presented a new approach for speech BWE involving both phoneme dependent and speaker dependent estimation of the spectral envelope. A three-step estimation algorithm was developed to deal with common difficulties in spectral envelope estimation. These difficulties are the estimation of unvoiced sounds and the robustness to different speakers and to erroneous estimation. The phoneme estimation employs an HMM to estimate the phonetic content of a speech frame. The spectral envelope estimation relies on a CB searching to estimate the speaker's VTAF. Postprocessing of the initial estimated WB VTAF, by matching formant frequencies in the low band to those of the input NB speech, smoothing in time, and gain adjustment, improved the HB spectral envelope estimation.

The experimental results demonstrate the improved performance of the proposed algorithm compared to other methods. Informal listening tests show improved quality of the enhanced speech. The drawbacks of the proposed algorithm are twofold. First, the concatenated tube model is limited in modeling VTAF shape of unvoiced and nasal sounds. Second, the iterative postprocessing procedure and the online sensitivity function calculation require high computational complexity. Future work might include a different VTAF estimation technique for unvoiced and nasal sounds. Offline calculation of the sensitivity function, for each WB VTAF codeword, will reduce the computational complexity. The algorithm should also be evaluated using formal listening tests, under different background noise conditions.

REFERENCES

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

[2] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[3] U. Kornagel, "Techniques for artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 86, pp. 1296–1306, 2006.

[4] J. Epps and W. H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," in *Proc. IEEE Workshop on Speech Coding*, pp. 371–374, Porvoo, Finland, 1999.

[5] H. Gustafsson, U. A. Lindgren and I. Claesson, "Low-complexity feature-mapped speech bandwidth extension," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 2, pp. 577–588, 2006.

[6] T. Ramabadran and M. Jasiuk, "Artificial bandwidth extension of narrow-band speech signals via high-band energy estimation," in *Proc. EUSIPCO 2008*, Lausanne, Switzerland, August 25-29. 2008.

[7] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen and P. Alku, "Evaluation of an artificial speech bandwidth extension method in three languages," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 6, pp. 1124–1137, 2008.

[8] A. H. Nour-Eldin and P. Kabal, "Mel-frequency cepstral coefficient-based bandwidth extension of narrow-band speech," in *Proc. InterSpeech*, pp. 53–56, 2008.

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

[10] P. Bauer, and T. Fingscheidt, "A statistical framework for artificial bandwidth extension exploiting speech waveform and phonetic transcription," in *Proc. EUSIPCO 2009*, Glasgow, Scotland, pp. 1839–1843.

[11] H .Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. on Audio and Electroac.*, vol. 21, pp. 417–427, 1973.

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

[13] R. V. Cox, D. Malah and D. Kapilov, "Improving upon toll quality speech for VoIP," in *Proc. 38'th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, California, Nov. 7-10, 2004, pp. 405–409.

[14] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals," in *Proc. ICASSP 2004*, Montreal, Quebec, Canada, pp. 697–700.

[15] H. Pulakka, V. Myllyla, L. Laaksonen, P. Alku, "Bandwidth extension of telephone speech using a filter bank implementation for highband Mel spectrum," in *Proc. EUSIPCO 2010*, Aalborg, Denmark, pp. 979–983.

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