Synchronous Overlap and Add of Spectra for Enhancement of Excitation in Artificial Bandwidth Extension of Speech

M. A. Tuğtekin Turan and Engin Erzin

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

mturan, eerzin@ku.edu.tr

Abstract

In this paper, a new approach that extends narrow-band excitation signals using synchronous overlap and add (SOLA) of spectra have been proposed. Although artificial bandwidth extension (ABE) of speech has been extensively studied, the role of excitation spectra has not been as widely studied as the spectral envelope extension. In this study ABE is investigated with the widely used source-filter framework, where speech signal is decomposed into excitation signal (source) and spectral envelope (filter). For the spectral envelope extension, our former work based on hidden Markov model has been used. For the excitation signal extension, we propose a SOLA of excitation spectra, where the high end of the excitation spectra is extended by preserving the harmonic structure. In experimental studies, we also apply two other well-known extension techniques for excitation signals. Then comparatively we evaluate the overall performance of proposed system using the PESQ metric. Our findings indicate that the proposed excitation extension method delivers significant quality improvements.

Index Terms: artificial bandwidth extension, speech enhancement, excitation extension, hidden Markov model.

1. Introduction

One of the main criterion that identifies the quality of speech is definitely bandwidth of incoming signal. Today, the upper frequency bound of conventional telephony speech is defined as 3400 Hz due to some historical reasons from analog communication era [1]. Although intelligibility of many phonetic groups is still around 90% within this frequency limit, fricative phones like /s/ and /f/ or affricates like /c/ and /ch/ have considerable information beyond this upper bound [2]. Speech signals are also somewhat susceptible to disturbance of both power transmission lines and electrical noises around the lower frequency bound, defined as 300 Hz. There is roughly 25 dB attenuation between 50 Hz and 60 Hz [3]. As a consequence of these effects, narrowband speech reveals slightly different auditory perception in comparison with wideband speech. Note that wideband speech communication is formally defined between 80 Hz and 8000 Hz [4]. It is possible to observe some problems in terms of intelligibility and naturalness due to the aforementioned bandwidth loss.

Artificial bandwidth extension (ABE) defines an enhancement mapping from narrowband speech to wideband speech. In this paper, we investigate ABE problem by focusing on the excitation signal extension problem along with the use of our former

work that applies hidden Markov model (HMM) for spectral envelope extension [5]. The proposed excitation extension scheme constructs missing frequency band of wideband excitation signal using synchronous overlap and add of the higher bands in excitation spectrum. In order to evaluate the proposed excitation extension scheme, we also define two widely-used methods as benchmarks in the experimental evaluations. The organization of the paper is the following: Section 2 introduces the ABE approach and related literature, then the benchmark methods and the proposed system are described in Section 3. Finally, experimental results are discussed using objective metrics in Section 4 with future research comments.

2. Artificial Bandwidth Extension

Existing studies on ABE problem mostly use source-filter analysis of speech production. The excitation signal (source) and the spectral envelope (filter) are defined as two independent channels of information. In general, wideband extension of spectral envelope has been studied more extensively in the literature. Statistical mapping schemes using machine learning or speech recognition are applied to construct extended spectral envelope. Parameters that shape spectral envelope are mostly chosen as linear prediction, cepstral or reflection coefficients. In some studies voiced/unvoiced and short-term power information are added to these feature sets [6]. Widely used techniques for the spectral envelope extension are codebook based linear prediction [7], linear or piece-wise linear mapping [8] and Bayesian estimation based Gaussian mixture model (GMM) [9, 10] or HMM transformations [11, 12]. Also, neural-network based mapping schemes have been applied to the ABE problem [13]. In this paper, we use the HMM-based wideband spectral envelope estimation method in [5]. This method 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.

The second information channel for the ABE problem is excitation extension. Excitation extension can be performed more efficiently than the envelope extension, as the excitation spectra is much more flat than the envelope spectra. Since human auditory system cannot easily notice variations of spectral flatness, reproduced frequency components are perceived well even if they do not satisfy spectral flatness entirely [14]. In an important work, Makhoul and Beouti introduced a high-frequency regeneration method for the excitation signal [15].

Their presented technique is based on spectral duplication of the baseband. Later, similar approaches have been proposed by other studies using spectral mirroring [12], folding [4] and translation [16]. On the other hand, recovering frequency harmonics via non-linear filtering or cosine generator have also been proposed in several studies [17].

3. Excitation Extension Methods

In this paper, we define two benchmark methods and introduce a new method based on synchronous overlap and add (SOLA) of spectra for the excitation extension problem. The first benchmark is the upsampling method, which performs zero padding between two consecutive samples. The second benchmark is a spectral shifting method, which is proposed by Andersen et al. and moves spectrum by preserving the relations between harmonics [18]. In this paper we propose a new method that extends narrowband excitation signals using SOLA of excitation spectra at high bands.

3.1. Upsampling

Upsampling is the most intuitive way of mapping narrowband excitation to wideband. Upsampling creates a spectral mirroring in the resulting wideband excitation spectra. Figure 1 shows the details of this method where the plot on the left side shows the signal in time domain while right side shows the same signal in frequency domain.

Figure 1: Sample time and frequency domains signals for the upsampling method. Left side is the signal in time domain and right side is the magnitude spectrum. Original signal is shown at the top, extended version by upsampling is shown at the bottom.

3.2. Spectral Shifting

The second benchmark method moves the spectrum such that the distance between the harmonics and the structure are preserved in the high-band. This method is expected to perform better than the upsampling method as the resulting excitation signal follows the spectral orientation low-bands at higher frequencies [18]. Using the modulation property of exponential signals, the spectral shifting procedure can be realized in time domain. The spectrum of a signal shifts by multiplying it with an exponential function.

The block diagram of this method is depicted in Figure 2. The letters (a), (b), (c) and (d) denotes the different intermediate states in the process. These steps can be seen extensively in the time domain as well as in the frequency domain on Figure 3.

A demonstration of the spectral shifting method is given in Figure 3. A sample signal is extended by zero padding and followed by low pass filtering. Now the signal is band-limited up

Figure 2: Block diagram of the spectral shifting method.

to 16 kHz in step (a). Next step is to multiply by the modulation function, $e^{j\omega_0 n}$, where the result of this multiplication is in (b). The lower frequencies are undesirable and they can be removed by a high-pass filter as in (c). The modulated signal and the original signal can be added to construct the wideband signal in (d).

Figure 3: Steps in the spectral shifting: (a) the original signal, (b) the signal is amplitude modulated, (c) the signal is high-pass filtered, and (d) the wideband extension is the sum of (a) and (c).

The factor G adjusts the attenuation of the artificial high frequency components. This factor has to be adjusted by subjective listening tests in such a way that the high frequency components are not annoying. The artificial bands can disturb listening effort because it is too periodic with respect to a real speech signal. The natural signal is often more blurred in higher frequencies because a small derivation of the pitch frequency result in a large effect in the higher frequencies.

3.3. The SOLA of Excitation Spectra

Both of the two benchmark methods reflect the harmonics around the half sample frequency of the narrowband signal.

Hence in the excitation extension of a voiced speech, which contains strong harmonics at low frequencies, creates strong harmonics at the high frequencies after the extension. Also in both methods there is a possibility of creating a discontinuity in the harmonic structure at the middle of the spectra. Hence the common drawbacks of the benchmark methods are observed as strong artificial harmonics at the high frequencies and possible harmonic structure discontinuities in the middle of the spectra. The proposed SOLA of excitation spectra targets to eliminate these two drawbacks. In the SOLA of spectra, the harmonic structure of the high-end spectra is extended by preserving the harmonic structure. A block diagram of the SOLA of spectra scheme is given in Figure 4 and it can be defined with the following steps:

- (i) Start with a spectra covering $[0 - 4]$ KHz band.
- (ii) Take the 2 KHz high-end magnitude spectra, i.e. $[2 - 4]$ KHz band, and correlate it starting from 2.5 KHz and up to find the maximum correlated band and frequency shift Δf .
- (iii) Perform overlap and add of the high-end magnitude spectra starting from $(2 + \Delta f)$ KHz using Hamming window. Keep the phase information from the shifted spectra.
- (iv) Repeat steps (ii) and (iii) until Δf accumulates to 4 KHz.
- (v) Compute a pitch search on the narrowband signal and extract a normalized correlation score for the pitch lag in $[0, 1]$ interval. This normalized correlation score represents the voicing information. Perform a lowpass filter with 8 KHz cut-off and an adaptive attenuation up to 10 dB as the normalized correlation score decreases to 0.3.

Figure 4: Block diagram of the SOLA of excitation spectra.

A sample realization of the SOLA of spectra is given in Figure 5. Note that if a harmonic structure exists in the high frequencies, it's propagated during the extension by preserving harmonic structure. Furthermore, if harmonic structure is weak then normalized correlation score introduces an attenuated low-pass filter to reduce excessive strong components in the high frequencies.

Figure 5: A sample realization of the SOLA of excitation spectra: (a) Narrow-band signal, (b-f) sliding high-end spectra with correlation maximization, (g) SOLA of excitation, (h) low-pass filtering with adaptive voicing attenuation.

4. Experimental Results

Experimental evaluations are performed over TIMIT database with 1176 sentences by 231 subjects. Narrowband samples are extracted from this database after down-sampling operation. The spectral envelope extension model in [5] has been trained using the training portion of the TIMIT database. Then the performance analysis of the ABE system has been executed on the testing portion of the TIMIT database. In performance evaluations, we use the PESQ, which is an ITU-T recommendation, as the objective quality metric [19].

Table 1: Average PESQ scores for all excitation extension methods

	PESQ
Upsampling	2.21
Spectral Shifting	2.52
SOLA of Spectra	2.83
Wideband Excitation	2.90

Table 1 presents average PESQ scores for all excitation extension methods. Note that bottom line presents the average PESQ for spectral envelope extension with the original wideband excitation signal. Hence the bottom line sets an upper bound for the performance of excitation enhancement. Note that the PESQ difference between upsampling extension and the upper bound condition is significantly high. Spectral shifting method introduces almost 0.3 PESQ improvement over the upsampling extension. Furthermore the proposed SOLA of excitation spectra brings an additional 0.3 PESQ improvement over the spectral shifting method and attains a high PESQ score with respect to the upper bound condition.

Figure 6 shows magnitude spectrum of the original and extended excitations where a 30 ms voiced segment is used. The spectrum at (a) belongs to the original 16 kHz wideband excitation. The spectrum after the upsampling method is shown at (b). The excitation spectra of the spectral shifting method is given at (c). The proposed SOLA of spectra method extracts the spectrum at (d). Note that the benchmark methods introduce strong harmonic structures to the high-end spectra and display discontinuities at 4 KHz. In the proposed scheme although some of the harmonic structure is preserved, it does not introduce any discontinuity and low-pass filter smooths high-end spectra with the calculated voicing score.

Figure 6: Magnitude spectrum of the original and extended excitations: (a) Original wideband spectrum, (b) spectrum after the upsampling, (c) spectrum after the spectral shifting, and (d) spectrum after the SOLA of spectra.

Figure 7 presents spectrogram view of original speech signal and all other schemes used in this paper. Similarly the benchmark methods restore high-band components intensively compared to the proposed method, which is shown at the bottom. However, in the benchmark methods frequency components of non-periodic unvoiced regions are directly copied or mirrored to the high-end spectra without analyzing periodicity of speech signals. This problem causes bursts, which mainly disturb listening effort. Speech samples from all the three ABE systems are available online at [20].

5. Conclusion

Although spectral envelope extension is widely studied in the ABE literature, a number of methods, which are presented in Section 1, exist for the extension of excitation via source-filter analysis framework. Conventional ABE systems rely on the flat spectral characteristic of the excitation signal. On the other hand, spectral envelope representation is largely considered as the dominant factor that represents general characteristics of human speech. Thus, ABE systems are widely dedicated to spectral envelope extension. However, experimental analysis

Figure 7: Spectrograms of the original and extended excitations: (a) Original wideband, (b) after the upsampling, (c) after the spectral shifting, and (d) after the SOLA of spectra.

of this study shows the importance of excitation enhancements. A finely tuned excitation extension does a much better job than the standard upsampling scheme, and the proposed SOLA of spectra scheme attains 2.83 average PESQ score, which is only 0.07 lower than the most informative upper bound score, which is 2.90 when the original wideband excitation has been used.

6. References

- [1] H. Pulakka, L. Laaksonen, V. Myllyla, S. Yrttiaho, and P. Alku, "Conversational evaluation of speech bandwidth extension using a mobile handset," *Signal Processing Letters, IEEE*, vol. 19, no. 4, pp. 203–206, 2012.
- [2] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1, 2004, pp. 709–713.
- [3] U. Kornagel, "Techniques for artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 86, no. 6, pp. 1296–1306, 2006.
- [4] J. A. Fuemmeler, R. C. Hardie, and W. R. Gardner, "Techniques for the regeneration of wideband speech from narrowband speech," *EURASIP Journal on Applied Signal Processing*, vol. 2001, no. 1, pp. 266–274, 2001.
- [5] C. Yağlı, M. A. T. Turan, and E. Erzin, "Artificial bandwidth extension of spectral envelope along a viterbi path," *Speech Communication*, vol. 55, no. 1, pp. 111–118, 2013.
- [6] B. Iser and G. Schmidt, "Bandwidth extension of telephony speech," in *Speech and Audio Processing in Adverse Environments*. Springer, 2008, pp. 135–184.
- [7] H. Carl and U. Heute, "Bandwidth enhancement of narrow-band speech signals," in *Proc. EUSIPCO*, vol. 2, 1994, pp. 1178–1181.
- [8] Y. Nakatoh, M. Tsushima, and T. Norimatsu, "Generation of broadband speech from narrowband speech using piecewise linear mapping," in *Proc. EUROSPEECH*, 1997.
- [9] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using gmm based transformation," in *Acoustics, Speech,*

- and *Signal Processing, IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1843–1846.
- [10] G.-B. Song and P. Martynovich, “A study of hmm-based bandwidth extension of speech signals,” *Signal Processing*, vol. 89, no. 10, pp. 2036–2044, 2009.
 - [11] K.-T. Kim, M.-K. Lee, and H.-G. Kang, “Speech bandwidth extension using temporal envelope modeling,” *Signal Processing Letters, IEEE*, vol. 15, pp. 429–432, 2008.
 - [12] P. Jax, “Bandwidth extension for speech,” *Audio bandwidth extension*, pp. 171–236, 2004.
 - [13] J. Kontio, L. Laaksonen, and P. Alku, “Neural network-based artificial bandwidth expansion of speech,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 873–881, 2007.
 - [14] H. Pulakka and P. Alku, “Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2170–2183, 2011.
 - [15] J. Makhoul and M. Berouti, “High-frequency regeneration in speech coding systems,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 4. IEEE, 1979, pp. 428–431.
 - [16] C.-F. Chan and W.-K. Hui, “Wideband re-synthesis of narrowband celp-coded speech using multiband excitation model,” in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 1. IEEE, 1996, pp. 322–325.
 - [17] Y. Qian and P. Kabal, “Dual-mode wideband speech recovery from narrowband speech,” in *INTERSPEECH*, 2003.
 - [18] B. Andersen, J. Dyreby, B. Jensen, F. H. Kjærskov, O. L. Mikkelsen, P. D. Nielsen, and H. Zimmermann. Bandwidth expansion of narrow band speech using linear prediction. [Online]. Available: <http://kom.aau.dk/group/04gr742/pdf/article.pdf>
 - [19] ITU-T, “Wide-band extension to recommendation p.862.2 for the assessment of wide-band telephone networks and speech codecs,” *International Telecommunication Union*, 2005.
 - [20] Speech samples of synchronous overlap and add of spectra for enhancement of excitation in artificial bandwidth extension of speech. [Online]. Available: <http://home.ku.edu.tr/~eerzin/is15Excitation>