Künstliche akustische Bandbreitenerweiterung: Spektrale und temporale Einhüllende

Artificial Bandwidth Extension: Spectral and Temporal Envelope

Thomas Schlien, Peter Vary
Institut für Kommunikationssysteme, RWTH Aachen University, 52074 Aachen, Deutschland,
Email: {schlien,vary}@iks.rwth-aachen.de

Abstract

Despite the tremendous technological progress of signal processors and signal processing algorithms, the acoustic bandwidth of voice calls is in most cases still limited to 3.4 kHz. Wideband telephony, that takes into account the frequency range between 50 Hz and 7 kHz, is introduced nowadays under the name "HD Voice". Unfortunately both users must have HD devices of the same provider and network, which is mostly not the case. For this reason, until the ubiquitous availability of "HD Voice", a transitional technology is desirable that promotes the acceptance and aides market penetration of HD devices. The artificial acoustic bandwidth extension can meet these requirements by extending the narrowband speech signal to "HD Voice" with the help of statistical models and methods based on the model of speech production.

Past studies always emphasized the importance of the estimation of the spectral envelope, which is considered in this work in conjunction with the examination of the temporal envelope and the temporal fine structure of speech.

It is shown that in particular the disturbing artifacts generated by common bandwidth extension algorithms cannot be removed by a perfect spectral envelope and therefore a new approach for generating the excitation signal is necessary.

Introduction

Most modern artificial bandwidth extension (ABWE) algorithms are based on the parametric source-filter model of speech production. They estimate a set of parameters to extend a narrowband to wideband signal. These are usually spectral components like Linear Prediction (LP) coefficients or Line Spectral Frequencies (LSF). They are estimated by statistical algorithms like the mapping of the entries of a narrowband codebook to a wideband shadow codebook [1], (piecewise) linear mapping [2, 3], artificial neural networks, [4, 5], Minimum Mean Square Error (MMSE) estimation based on Gaussian Mixture Models (GMMs) [6], Hidden Markov Models (HMMs) [7, 8], or deep neural networks [9].

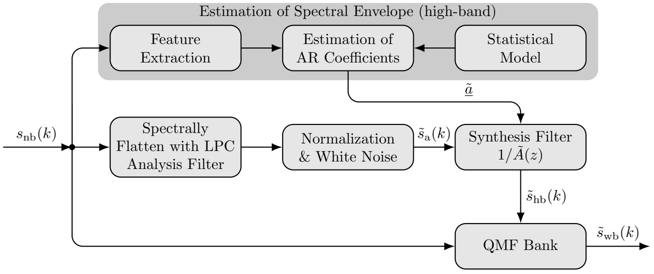
Estimation of Spectral Envelope (high-band)
Feature Extraction
Estimation of AR Coefficients
Statistical Model
$\underline{\tilde{a}}$
$s_{\mathrm{nb}}(k)$
Spectrally Flatten with LPC Analysis Filter
Normalization & White Noise
$\tilde{s}_{\mathrm{a}}(k)$
Synthesis Filter $1/\tilde{A}(z)$
$\tilde{s}_{\mathrm{hb}}(k)$
QMF Bank
$\tilde{s}_{\mathrm{wb}}(k)$

Figure 1: Block diagram of a typical artificial bandwidth extension algorithm

A typical block diagram of an artificial bandwidth extension algorithm [10, 11] is shown in Figure 1. For the generation of a high-band excitation $\tilde{s}_{\mathrm{a}}(k)$, the narrowband speech signal $s_{\mathrm{nb}}(k)$ with time index k is spectrally flattened with the help of an LPC analysis filter. The result is normalized and some additional white noise is added to compensate the stronger noisiness of the excitation in the high-band in comparison to the narrowband excitation.

To estimate the necessary parameters for the spectral envelope of the high-band, audio features are extracted from the narrowband speech signal $s_{\mathrm{nb}}(k)$. These features are, e.g., Mel Frequency Cepstral Coefficients (MFCC) and the Zero Crossing Rate (ZCR). With the help of a wideband speech pretrained statistical model, the LP coefficients $\underline{\tilde{a}}$ can be estimated. To obtain the estimated high-band signal $\tilde{s}_{\mathrm{hb}}(k)$ the LP coefficients $\underline{\tilde{a}}$ are applied to the high-band excitation $\tilde{s}_{\mathrm{a}}(k)$. With a Quadrature Mirror Filter (QMF) bank, $\tilde{s}_{\mathrm{hb}}(k)$ is interpolated and mixed with the narrowband signal $s_{\mathrm{nb}}(k)$ to get the estimated wideband signal $\tilde{s}_{\mathrm{wb}}(k)$. The sampling rate of the output signal $\tilde{s}_{\mathrm{wb}}(k)$ is twice the sampling rate of the input signal $s_{\mathrm{nb}}(k)$.

Importance of Spectral and Temporal Envelope

Artifacts are the main acceptance problem of ABWE algorithms. Depending on the algorithm they sound disharmonious on the one hand and like fast modulated noise on the other. To reduce or even expunge these artifacts new ways have to be found. Previous works, mentioned in the introduction, assume that the generation of the excitation signal is uncritical with the presented method while the estimation of the spectral envelope of the high-band is supposed to be very important. This, however, does not agree with the authors' experiences which show that the precision and number of parameters forming the spectral envelope only have little influence on the perceived quality and almost no influence on the generation of artifacts.

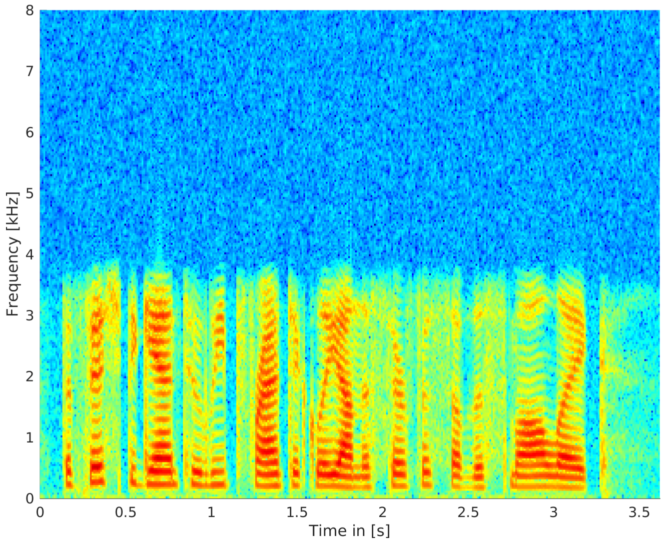

Figure 2: Spectrogram of original narrowband signal $s_{\mathrm{nb}}(k)$.

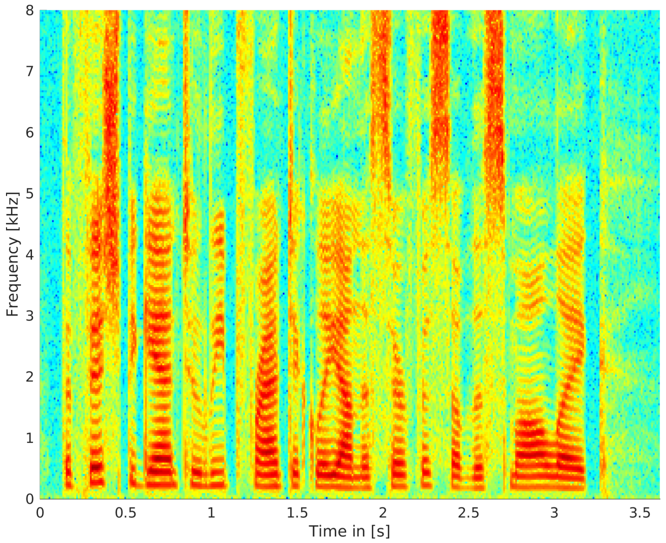

Figure 3: Spectrogram of original wideband signal $s_{\mathrm{wb}}(k)$.

Influence of High-band Spectral Envelope

To challenge the assumptions of the importance of the spectral envelope, oracle experiments were carried out.

The narrowband speech signal $s_{\mathrm{nb}}(k)$ ("The fish began to leap frantically on the surface of the small lake.") (Figure 2) and the original wideband signal $s_{\mathrm{wb}}(k)$ (Figure 3) are considered.

To show the influence of the spectral envelope two approaches were compared. On the one hand a signal was created using the original excitation in the high-band without applying a spectral envelope, i.e., the high-band is spectrally flat. This case is shown in Figure 4. On the other hand the high-band excitation signal was created by spectrally flatten the narrowband signal $s_{\mathrm{nb}}(k)$. Afterwards, the original spectral envelope was applied. Figure 5 shows the spectrogram of the resulting signal.

An informal listing test was conducted which revealed that in almost every case the listeners preferred the results of the first approach (Figure 4). There were almost no audible artifacts and the signal only suffers from a slightly degraded naturalness in comparison to the original wideband signal while in the second approach the signal (Figure 5) was disturbed by the well known artifacts.

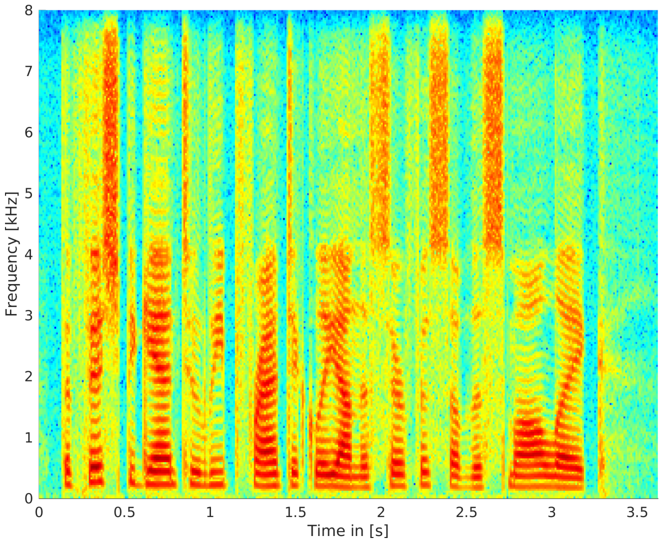

Figure 4: Spectrogram of wideband signal with spectral flat high-band.

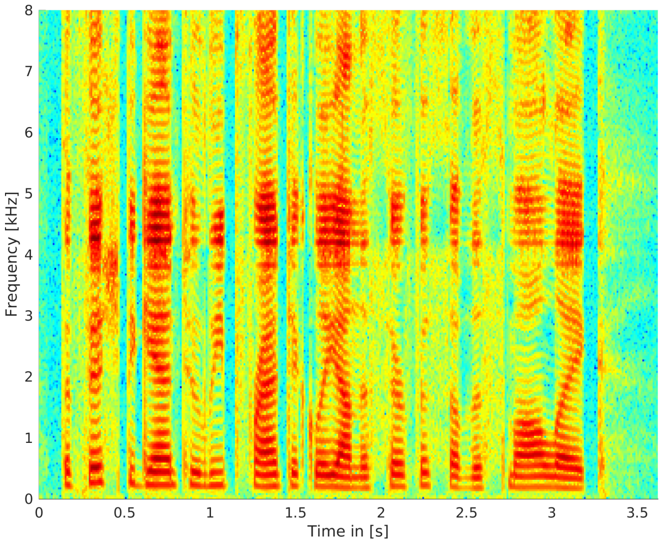

Figure 5: Spectrogram of wideband signal with copied narrowband excitation to the high-band and applying original spectral envelope.

Considerations for Excitation Signal Creation

To create an artifact-free artificial excitation signal, a new method has to be found. To achieve this goal, the authors had a closer look at the temporal envelope of the spectrally flattened high-band signal $s_{\mathrm{a}}(k)$. To get the temporal envelope $s_{\mathrm{te}}(k)$ (Figure 6) an Hilbert transform [12] is applied to the high-band.

$$s_{\mathrm{te}}(k) = \mathrm{abs}(\mathrm{Hilbert}(s_{\mathrm{a}}(k))) \tag{1}$$

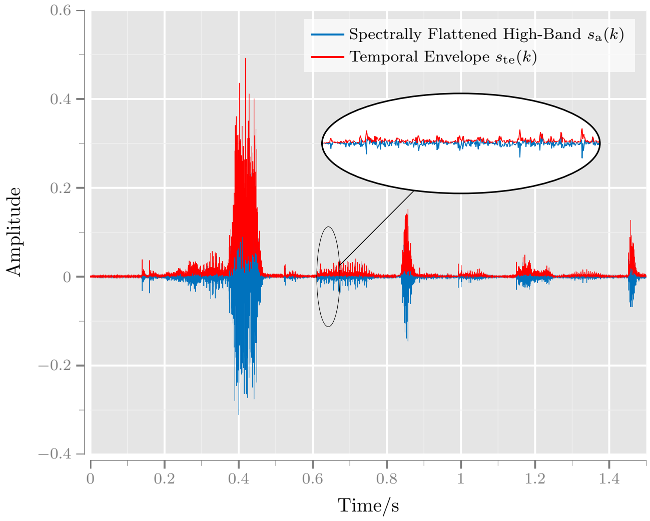

Figure 6: Hilbert time envelope $s_{te}(k)$ of spectral flattened high-band signal $s_a(k)$.

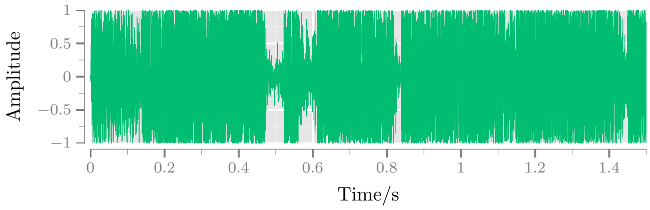

Figure 7: Temporal fine structure $s_{tf}(k)$ of whitened high-band signal $s_a(k)$.

The temporal fine structure can be observed by dividing the time signal $s_a(k)$ by the temporal envelope $s_{te}(k)$ (cf. Figure 7).

$$s_{tf}(k) = s_{hb}(k)/s_{te}(k) \tag{2}$$

It can be noticed that the temporal fine structure looks like white noise, at least in the active speech segments.

To test this assumption, the Hilbert transform time envelope $s_{te}(k)$ is multiplied with artificially created white noise. Informal listening tests showed that the difference between the white noise excited temporal envelope signal and the original signal with spectral flattened high-band is only hardly audible. Therefore it is possible to generate the high-band excitation signal by this approach.

The next logical step for the creation of an artificial excitation signal is to estimate the temporal envelope of the spectral flattened high-band. Since it is not very reasonable to estimate every sample of the temporal envelope, smoothing of the envelope in steps of 20, 40, 80, 160, and 320 samples, which equals to 2.5, 5, 10, 20, and 40 ms, is applied. While the smoothing over 20 till 80 samples is hardly audible, in the case of 160 and 320 samples the high-band sounds slightly more noisy. However, even these long smoothed versions sound more natural than the normal ABWE artifacts. Therefore, the estimation of a subsampled temporal envelope and its multiplication with white noise is a reasonable way to generate an artificial high-band excitation signal.

Summary

In this paper the importance of the spectral and temporal envelope in artificial bandwidth extension algorithms was investigated and analyzed. It has been shown that the spectral envelope is less important than expected and its further improvement does not reduce the disturbing artifacts. Concluding that the excitation seems to produce the typical ABWE artifacts, a new method to create an improved excitation signal has to be found. To achieve this goal, estimating the subsampled temporal envelope and multiplying it with white noise is a promising approach. Applying a smoothed spectral envelope afterwards might help to improve the naturalness and robustness of the estimated high-band signal.

References

[1] H. Carl and U. Heute, “Bandwidth enhancement of narrow-band speech signals,” in *Proceedings of European Signal Processing Conference (EUSIPCO)*, Edinburgh, Scotland, Sep. 1994, pp. 1178–1181.

[2] Y. Nakatoh, M. Tsushima, and T. Norimatsu, “Generation of broadband speech from narrowband speech using piecewise linear mapping,” in *Proceedings of EUROSPEECH*, Rhodes, Greece, Sep. 1997, pp. 1643–1646.

[3] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, “Speech enhancement via frequency bandwidth extension using line spectral frequencies,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, USA, 2001, pp. 665–668.

[4] J. Kontio, L. Laaksonen, and P. Alku, “Neural network-based artificial bandwidth expansion of speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 873–881, 2007.

[5] H. Pulakka and P. Alku, “Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2170–2183, 2011.

[6] K.-Y. Park and H. S. Kim, “Narrowband to wideband conversion of speech using GMM based transformation,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, 2000, pp. 1843–1846.

[7] P. Jax, “Enhancement of bandlimited speech signals: Algorithms and theoretical bounds,” Ph.D. dissertation, IND, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany, 2002, volume 15 in “Aachener Beiträge zu Digitalen Nachrichtensystemen (ABDN)”, Verlag Mainz, Aachen, Germany.

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

[9] K. Li and C.-H. Lee, “A deep neural network approach to speech bandwidth expansion,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015, pp. 4395–4399.

[10] T. Schlien, F. Heese, M. Schäfer, C. Antweiler, and P. Vary, “Audiosignalverarbeitung für Videokonferenzsysteme,” in *Workshop Audiosignal- und Sprachverarbeitung (WASP)*, ser. Lecture Notes in Informatics (LNI) - Proceedings, vol. Vol. P-220, Koblenz, Germany, Sep. 2013, pp. 2987–3001.

[11] F. Heese, B. Geiser, and P. Vary, “Intelligibility assessment of a system for artifical bandwidth extension of telephone speech,” in *Proceedings of German Annual Conference on Acoustics (DAGA)*. Darmstadt, Germany: DEGA, Mar. 2012, pp. 905–906.

[12] A. V. Oppenheim, R. W. Schafer, J. R. Buck *et al.*, *Discrete-time signal processing.* Prentice hall Englewood Cliffs, NJ, 1989, vol. 2.