Beyond Wideband Telephony — Bandwidth Extension for Super-Wideband Speech

Bernd Geiser and Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany

Email: {geiser|vary}@ind.rwth-aachen.de

Abstract

Driven by the market success of high-quality Voice over IP technology, the introduction of wideband telephony with an acoustic bandwidth of at least 7 kHz is meanwhile also foreseen for "traditional" digital voice communication services such as ISDN, DECT, or UMTS. While wideband speech addresses the basic requirement of intelligibility (even for meaningless syllables), the perceived "naturalness" and the experienced "quality" of speech can be further enhanced by providing an even larger acoustic bandwidth. Thus, the next logical step towards true "Hi-Fi Telephony" could be the rendering of "super-wideband" (SWB) speech signals with an acoustic bandwidth of at least 14 kHz.

In this contribution, we review previous and current standardization activities with this focus. Moreover, a method for artificial bandwidth extension (BWE) of wideband speech signals towards "super-wideband" is presented and evaluated. It is shown that improved naturalness and speech quality can be attained by a purely receiver-based modification of wideband terminals.

Wideband vs. Super-Wideband Speech

Typically, *wideband* (WB) speech is defined by its acoustic frequency range of $0.05 - 7$ kHz, whereas *super-wideband* speech provides a roughly doubled bandwidth of, e.g., $0.05-14$ kHz. The lower cutoff frequency of 50 Hz is usually considered sufficient for a natural reproduction of speech signals. An analysis[1] reveals that on average only about 1.5 % of the energy of super-wideband speech signals is located in the $7-14$ kHz *extension band* (EB). This average is only exceeded in less than 25 % of all active frames, which indicates that there must be strong outliers in the EB to SWB energy ratio $\sigma^2_{\mathrm{EB}}/\sigma^2_{\mathrm{SWB}}$. Such outliers are actually found in fricative and plosive speech sounds as illustrated in Fig. 1. For particularly strong outliers, the EB energy is even larger than the WB energy. This is the case for about 6 % of all active frames. Here, the largest benefits over WB signals can be expected. In addition to simple energy considerations, there is also evidence that *temporal* signal characteristics gain *perceptual* importance with an increasing frequency, cf. [1, 2]. For the EB range of $7-14$ kHz, detailed temporal characteristics may be even more important than the exact reproduction of the *spectral* envelope information.

[1] Results are based on approx. $1.6 \cdot 10^4$ *active* 20 ms speech frames, sampled at 32 kHz and low-pass filtered with $f_c = 14$ kHz.

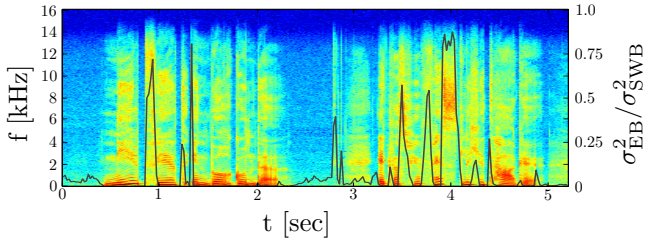

Figure 1: Exemplary spectrogram of the super-wideband speech sample *"Nilpferd in Burgunder, etwas für festliche Tage"*, sampled at $f_s = 32$ kHz and low-pass filtered with $f_c = 14$ kHz; the gray curve illustrates the relative energy contribution of the frequencies between 7 and 14 kHz.

Super-Wideband Speech Coding

Coding of *wideband* speech signals has been extensively investigated for many years, e.g., [3]. A number of standardized speech coding algorithms (such as the AMR-WB codec) are the result of these efforts. In fact, the introduction of wideband services is envisaged for the near future. But still, as motivated above, further quality improvement can be expected by doubling the transmitted frequency range, i.e., by transmitting SWB speech. Meanwhile a couple of SWB speech and audio codecs, that are suitable for conversational services and conferencing, exist. Two examples for such coders are Annex C of ITU-T G.722.1 [4] and the MPEG-4 low delay AAC codec [5] (which is actually a "full band" codec). The algorithmic delay of these codecs is compatible with conversational requirements and at least the former also offers low computational complexity. Apart from such "monolithic" solutions, a few proposals for SWB extensions of standardized codecs in an embedded coding framework [6] exist. The resulting "bandwidth scalable" codecs typically employ rather coarse *parametric* signal models to describe the high frequency components. Analog to WB extensions such as [7], such "bandwidth extension (BWE) with side information" is usually sufficient to synthesize SWB frequencies with an adequate quality. In [8] and [9], the additional frequency components are parametrically encoded in the spectral (MDCT) domain, for instance. With this approach, the total bit rate for the encoding of the $7-15$ kHz frequency components is 6.6 kbit/s. A *standardized* solution is found in the AMR-WB+ codec [10] which also uses parametric BWE techniques. Finally, a new SWB extension is currently being studied within ITU-T. This algorithm shall be applicable to *two* wideband codecs: ITU-T G.729.1 [11] and the new variable rate codec G.EV-VBR [12].

Bandwidth Extension for SWB Speech

Here, we investigate the performance of SWB BWE *without* side information, i.e., we *estimate* SWB parameters based on the WB speech signal. This approach might become useful once wideband speech transmission is deployed in communication networks [13]. Therefore, similar to our approach in [14], we have *integrated* the parameter estimation into a wideband speech codec, namely into the ITU-T G.729.1 coder.

First, a *feature set* that gives a relevant description of the wideband speech signal is required. For each frame, a feature vector $\mathbf{x}_\mathrm{f}$ is composed that includes temporal and spectral information about the low ($0-4\,\mathrm{kHz}$) and high ($4-7\,\mathrm{kHz}$) frequency bands, in particular the line spectral pairs (LSPs) from the G.729 core codec, a low band temporal envelope as well as the G.729.1 TDBWE parameter set [7] which describes the high band. Secondly, a SWB BWE algorithm is needed. Our implementation shapes artificial noise in the $7-14\,\mathrm{kHz}$ range according to 10 temporal subframe and 11 spectral subband gain factors per 20 ms frame. These gain factors constitute the SWB parameter set $\mathbf{y}$. The task is then to estimate $\mathbf{y}$ from the knowledge of the feature vector $\mathbf{x}_\mathrm{f}$.

In principle, all previously proposed WB BWE estimation schemes such as [15] or [16] are applicable. Here, we use a very simple piecewise linear mapping (PLM) approach similar to [17]. This estimation scheme is computationally very efficient, but yet quite effective for the estimation of SWB parameters. The (unweighted) PLM estimation rule for the SWB parameters is given by

$$\hat{\mathbf{y}} = \mathbf{H}_i^\mathrm{T} \cdot \mathbf{x}_\mathrm{f},$$

where $i \in \{1,\ldots,M\}$ represents a classification of the narrowband features via vector quantization. We have chosen a quite low number of classes ($M = 4$). The matrices $\mathbf{H}_i$ are computed from training data as follows:

$$\mathbf{H}_i = \mathbf{X}_i^+ \cdot \mathbf{Y}_i = \left(\mathbf{X}_i^\mathrm{T}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^\mathrm{T} \cdot \mathbf{Y}_i,$$

where $\mathbf{X}_i$ and $\mathbf{Y}_i$ are matrices of training vectors that belong to the i-th feature class.

Using the PLM method, a convincing SWB speech quality could be obtained; the results will be demonstrated by audio examples. Objectively, a spectral distortion of 4.46 dB has been measured by computing the spectral distance between 8th order AR models of the original and estimated subband signals. To give a comparison, an experimental *quantization* of $\mathbf{y}$ with approx. 3 kbit/s resulted in near-transparent SWB speech. The spectral distortion was evaluated to 3.61 dB in this case.

Conclusion and Outlook

We have reviewed work related to coding of SWB speech signals. It can be observed that *BWE with side information* is a suitable choice to synthesize the SWB frequency components. We have also investigated the application of *BWE without side information* using an exemplary implementation. Thereby, a convincing SWB speech quality could be achieved. Still, improved performance can be expected by using advanced estimation schemes such as [16]. Further, a formal investigation of mutual information between $\mathbf{x}_\mathrm{f}$ and $\mathbf{y}$ would be beneficial to assess the theoretical performance bounds for SWB BWE.

References

[1] N. F. Viemeister, "Temporal modulation transfer functions based upon modulation thresholds," *Journal of the Acoustical Society of America*, vol. 6, no. 5, pp. 1364–1380, Nov. 1979.

[2] K. T. Kim, J.-Y. Choi, and H. G. Kang, "Perceptual relevance of the temporal envelope to the speech signal in the 4–7 kHz band," *Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. EL88–EL94, Sept. 2007.

[3] J. Schnitzler and P. Vary, "Trends and perspectives in wideband speech coding," *Signal Processing*, vol. 80, no. 11, pp. 2267–2281, Nov. 2000.

[4] M. Xie, D. Lindbergh, and P. Chu, "ITU-T G.722.1 Annex C: A new low-complexity 14 kHz audio coding standard," in *Proc. of IEEE ICASSP*, Toulouse, France, May 2006.

[5] E. Allamanche, R. Geiger, J. Herre, and T. Sporer, "MPEG-4 low delay audio coding based on the AAC codec," in *106th Conv. of the AES*, Munich, Germany, May 1999.

[6] B. Geiser, S. Ragot, and H. Taddei, "Embedded Speech Coding: From G.711 to G.729.1," in *Advances in Digital Speech Transmission*, R. Martin, U. Heute, and C. Antweiler, Eds. Chichester, UK: Wiley, Jan. 2008, ch. 8, pp. 201–247.

[7] B. Geiser, P. Jax, P. Vary, H. Taddei, S. Schandl, M. Gartner, C. Guillaumé, and S. Ragot, "Bandwidth extension for hierarchical speech and audio coding in ITU-T Rec. G.729.1," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2496–2509, Nov. 2007.

[8] M. Oshikiri, H. Ehara, and K. Yoshida, "A scalable coder designed for 10-kHz bandwidth speech," in *Proc. of IEEE Speech Coding Workshop*, Tsukuba, Ibaraki, Japan, Oct. 2002, pp. 111–113.

[9] ——, "Efficient spectrum coding for super-wideband speech and its application to 7/10/15 kHz bandwidth scalable coders," in *Proc. of IEEE ICASSP*, vol. 1, Montreal, QC, Canada, May 2004, pp. 481–484.

[10] J. Makinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, "AMR-WB+: a new audio coding standard for 3rd generation mobile audio services," in *Proc. of IEEE ICASSP*, Philadelphia, PA, USA, Mar. 2005.

[11] S. Ragot *et al.*, "ITU-T G.729.1: An 8-32 kbit/s scalable coder interoperable with G.729 for wideband telephony and Voice over IP," in *Proc. of IEEE ICASSP*, Honolulu, Hawai'i, USA, Apr. 2007.

[12] M. Jelínek, T. Vaillancourt, A. E. Ertan, J. Stachurski, A. Rämö, L. Laaksonen, J. Gibbs, and S. Bruhn, "ITU-T G.EV-VBR Baseline Codec," in *Proc. of IEEE ICASSP*, Las Vegas, NV, USA, Mar. 2008.

[13] P. Jax and P. Vary, "Bandwidth extension of speech signals: A catalyst for the introduction of wideband speech coding?" *IEEE Communications Magazine*, vol. 44, no. 5, pp. 106–111, May 2006.

[14] B. Geiser, H. Taddei, and P. Vary, "Artificial bandwidth extension without side information for ITU-T G.729.1," in *Proc. of European Conf. on Speech Communication and Technology (INTERSPEECH)*, Antwerp, Belgium, Aug. 2007.

[15] H. Carl and U. Heute, "Bandwidth enhancement of narrowband speech signals," in *Proc. of EUSIPCO*, vol. 2, Edinburgh, Scotland, Sept. 1994, pp. 1178–1181.

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

[17] Y. Nakatoh, M. Tsushima, and T. Norimatsu, "Generation of broadband speech from narrowband speech using piecewise linear mapping," in *Proc. of European Conf. on Speech Communication and Technology (EUROSPEECH)*, vol. 3, Rhodos, Greece, Sept. 1997, pp. 1643–1646.