Bandwidth Extension of Speech Signals: A Catalyst for the Introduction of Wideband Speech Coding?

Peter Jax and Peter Vary, RWTH Aachen University

ABSTRACT

The restricted audio quality of today's telephone networks is mainly due to the narrowband (NB) limitation to the frequency range from about 300 Hz to 3.4 kHz. Meanwhile, codecs for wideband (WB) telephony (50 Hz to 7 kHz) exist with significantly improved speech intelligibility and naturalness. However, the broad introduction of wideband speech coding will require strong efforts of both network operators and their customers because many elements of the networks (i.e., terminals and network nodes) have to be modified. An intermediate step to overcome the narrowband limitation can be achieved by applying artificial bandwidth extension (BWE) in the receiver. In this article we review the basic principles of bandwidth extension, and discuss several application scenarios in which both wideband coding and BWE complement each other. The introduction of BWE methods in terminals and networks may help to speed up the introduction of true wideband speech coding in the near future.

INTRODUCTION

The limited frequency range of about 300 Hz to 3.4 kHz of today's narrowband (NB) telephone networks leads to restricted audio quality compared to wideband (WB) telephony (50 Hz to 7 kHz). Wideband speech codecs have been standardized and are ready to be used, providing significant improvements in terms of speech intelligibility and naturalness. The conversion from NB to WB telephony requires investments by operators and customers. In the transition period NB and WB terminals will coexist for a long time, and compatibility of operation is a mandatory requirement. Therefore, each WB terminal has to be equipped with an NB codec to allow interoperability with any far-end NB terminal. The WB mode can only be used if the far-end terminal, the network, and the near-end terminal all have the improved WB capabilities.

A strong motivation for buying a WB terminal would be if the new telephone produces from the very beginning, at least at the near end, some WB speech, even if the far-end terminal as well as the network have not yet been converted to WB transmission. This situation is illustrated in Fig. 1. At the far end there is still a conventional NB telephone with analog-to-digital (A/D) conversion at a sampling rate of $f_s = 8$ kHz and an NB codec such as integrated services digital network (ISDN) A-law coding (International Telecommunication Union Telecommunication Standardization Sector [ITU-T] G.711), or Global System for Mobile Communications (GSM) enhanced full-rate encoding (European Telecommunications Standards Institute [ETSI] 06.60). At the receiving near-end terminal, in the first step, the NB speech signal s_{nb} is decoded using a conventional NB decoder. In a second step, artificial bandwidth extension (BWE) is applied to produce a WB signal $\hat{s}_{\mathrm{wb}}$ with a sample rate of $f_{\hat{s}} = 16$ kHz.

It cannot be expected that this BWE provides the same quality as true WB speech transmission, but it might significantly increase the acceptance of WB terminals. Hence, BWE might be an important catalyst for the conversion process from NB to WB telephony.

In this contribution we discuss several potential applications of BWE techniques, the most interesting being the bandwidth extension of telephone speech (frequency bandwidth 300–3400 Hz) to produce "wideband" speech (frequency bandwidth 50–7000 Hz).[1] We describe the principles of state-of-the-art BWE approaches, and further describe how some BWE techniques with side information are being used already as part of several speech codec standards.

FROM NARROWBAND TELEPHONY TO WIDEBAND TELEPHONY

As a matter of fact, the limited quality of NB telephone speech is widely accepted. However, in certain situations we clearly become aware of the impacts of bandwidth limitation. For example, the limited intelligibility of syllables becomes

[1] *In the literature in the area of speech coding, the term* wideband *has been established to denote a frequency bandwidth of 50 Hz to 7 kHz. The term* narrowband *typically implies a bandwidth from about 300 Hz to 3.4 kHz.*

0163-6804/06/$20.00 © 2006 IEEE

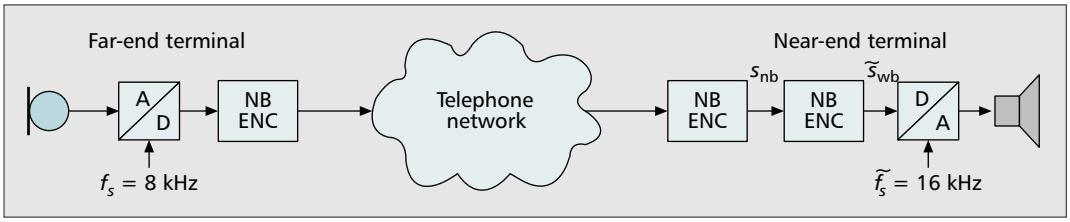

Figure 1. *Artificial bandwidth extension at the receiving terminal.*

apparent when we try to understand unknown words or names on the phone. In these cases we often need a spelling alphabet, especially to distinguish between certain unvoiced or plosive utterances, such as /s/ and /f/ or /p/ and /t/. Another drawback is that many speaker-specific characteristics are not retained transparently in the NB speech signal. Therefore, it is sometimes difficult to distinguish on the phone a mother from her daughter.

The bandwidth of WB transmission is comparable to that of amplitude modulated (AM) radio transmission, and it allows excellent speech intelligibility and very good speech quality. An example of unvoiced speech with significant frequency content beyond 3.4 kHz is given in Fig. 2, which shows a spectral comparison of the original speech with the corresponding NB and WB versions. A closer look at Fig. 2 reveals that NB speech may lack significant parts of the spectrum, and that the difference between WB speech and original speech is still noticeable.

The introduction of WB transmission in a telephone network requires at least new terminals with better electro-acoustic front-ends, improved A/D converters, and new speech codecs. In addition, signaling procedures are needed for detection and activation of WB capability. In cellular radio networks expensive modifications are necessary, since error protection (speech-codec-specific channel coding) is implemented in the base stations and not in the centralized switching centers.

Several WB speech codecs have been standardized in the past. In 1985 the first WB speech codec (G.722) was specified by CCITT (now ITU-T) for ISDN and teleconferencing with bit rates of 64, 56, and 48 kb/s. It is mainly applied in the context of radio broadcast stations by external reporters using special terminals and ISDN connections from outside to the studio. In 1999 a second WB codec (G.722.1) was introduced by ITU-T, which produces almost comparable speech quality at reduced bit rates of 32 and 24 kb/s. Most recently, the adaptive multirate WB (AMR-WB) speech codec has been specified by ETSI and 3GPP for code-division multiple access (CDMA) cellular networks such as Universal Mobile Telecommunications System (UMTS). The AMR-WB codec has also been adopted for fixed network applications by ITU-T (G.722.2). By the AMR-WB standard a family of wideband codecs (modes) with data rates between 6.6 and 23.85 kb/s is defined together with control mechanisms to adapt the codec mode to channel conditions. A further extension, the AMR-WB+ codec, supports general audio

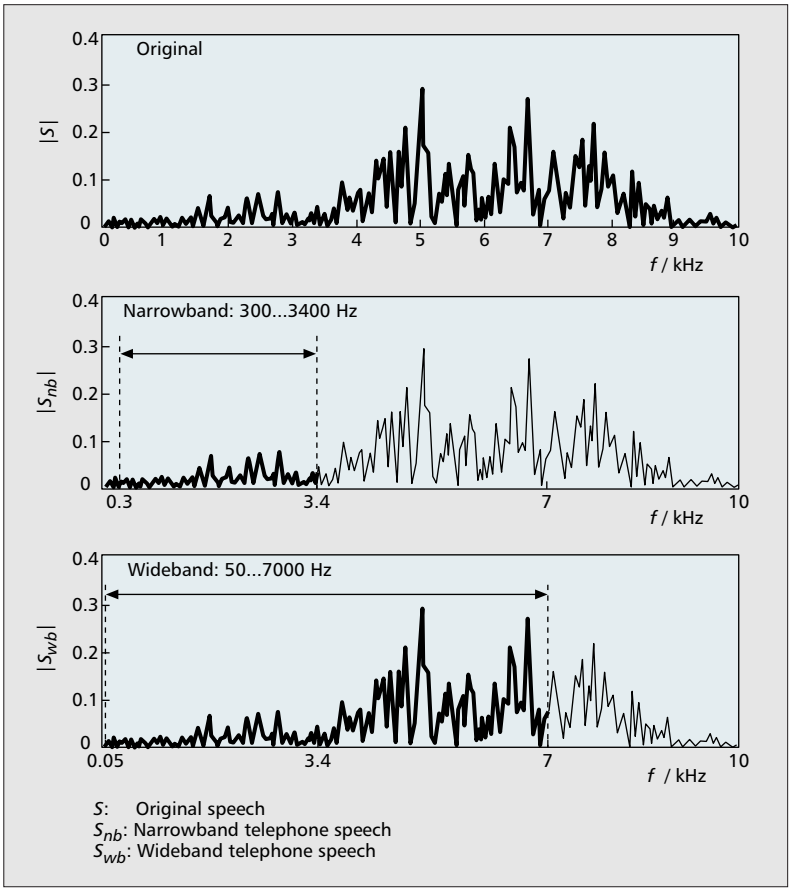

Figure 2. *Example short-term spectrum of an unvoiced utterance (linear scales).* S*: original speech;* S_{nb}*: narrowband telephone speech;* S_{wb}*: wideband telephone speech.*

in mono/stereo with frequency bandwidths up to more than 19 kHz and bit rates between 6 and 48 kb/s.

Even if cellular phones are replaced by new models much more often than fixed line telephones, there will be a long transitional period with NB and WB terminals in mixed use in both cellular and fixed networks. Different constellations of this transition period are illustrated in Fig. 3. There may be an NB terminal at the far end and NB transmission over the network, while the electro-acoustic front-end of the near-end terminal has already got WB capabilities (Fig. 3a). Due to the increased audio bandwidth of the near-end terminal (sampling rate 16 kHz), BWE can be applied to enhance the received speech signal. This produces more natural sounding speech, and the user can benefit from the improved WB capabilities of the terminal.

Many other setups are imaginable for which BWE in the network is reasonable, especially if a heterogeneous mixture of NB and WB terminals is involved. Examples include multi-party conference bridges, or mechanisms to prevent temporary switching from WB to NB.

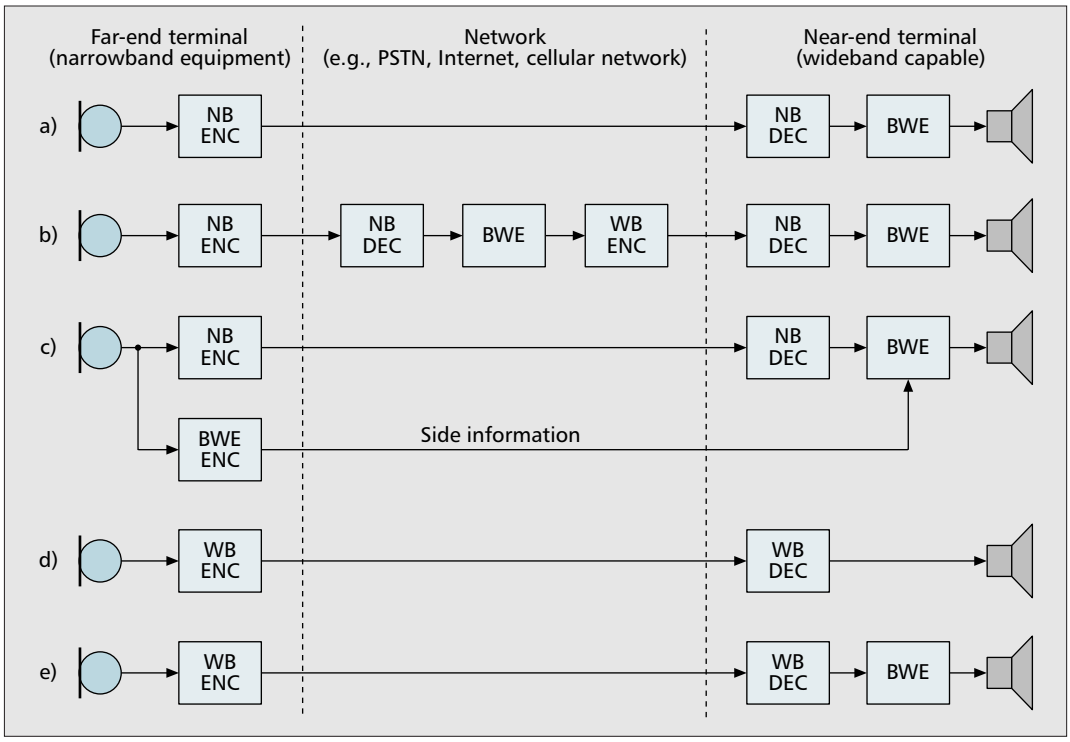

Figure 3. *Steps from narrowband to wideband telephony: a) narrowband transmission and bandwidth extension in the receiver; b) narrowband transmitter and bandwidth extension in the network; c) transmission with side information for bandwidth extension; d) Speech transmission using true wideband coding; and e) wideband transmission and bandwidth extension for "super-wideband" speech.*

This approach does not require any modification of the sending terminal and network. The implementation of BWE is particularly attractive for manufacturers with respect to the competition in the terminal market. For reasons of compatibility, the NB encoder has to be used in the WB terminal for the reverse direction.

Alternatively, the BWE can be placed within the core network, as illustrated in Fig. 3b. With this setup, the network operator can offer connections with improved quality at any time to any customer who is using a wb terminal, even if the far-end terminal provides only NB capabilities. During call setup the network can detect mixed connections between NB and WB terminals. Then it can route the connection via a transcoding unit located inside the core network. The transcoding unit consists of an NB decoder, BWE, and a WB encoder. The near-end terminal does not have to implement any BWE algorithms itself.

Many other setups are imaginable for which BWE in the network is reasonable, especially if a heterogeneous mixture of NB and WB terminals is involved. Examples include multiparty conference bridges, or mechanisms to prevent temporary switching from WB to NB (e.g., in case of intercell handovers in cellular networks).

A third solution is shown in Fig. 3c, which provides a significantly improved quality in comparison to the approaches of Fig. 3a and 3b. At the far end some side information is determined and communicated to the near-end terminal in parallel to the NB speech signal. The side information allows decoding of the WB speech signal on top of the already decoded NB speech. Accordingly, in certain cases, this approach can be interpreted as a variant of layered or embedded speech coding.[2]

A promising new approach is to embed the side information into the NB speech signal as a digital watermark message before encoding [1, 2]. The proper watermarking method makes this BWE system inherently backward-compatible without need for any signaling procedure: if the watermarked speech signal is presented to a human listener by a conventional NB receiver, he or she will not perceive any difference to the encoded original NB speech. If, on the other hand, the BWE receiver does not detect the embedded watermark in the NB speech, a stand-alone BWE approach (Fig. 3a) can still be activated. If both sides support BWE side information transmission, the receiver can produce WB speech with a very good quality, almost comparable to that of true WB codecs.

Finally, the true wideband connection requires, as shown in Fig. 3d, modifications of the transmitter, possibly the network, and the receiver by introducing new encoders and decoders. This solution can obviously provide the best speech quality.

Even if WB coding (50 Hz to 7 kHz) already has been implemented in the network, wideband extension beyond 7 kHz can be applied in addition to produce a *super-wideband* speech signal (e.g., with frequency components up to 15 kHz). This situation is depicted in Fig. 3e. It is obvious from Fig. 2 that the subjective speech quality can be further improved over the transmitted WB speech.

[2] *In layered speech coding the bitstream consists of several layers built on each other. At the receiver the base layer of the bitstream is sufficient to decode an acceptable speech signal. With each layer that is received in addition, the speech quality is improved successively.*

STANDALONE BANDWIDTH EXTENSION

To assess the prospects and limitations of BWE techniques it is necessary to understand the underlying principles. From Nyquist's theorem it is evident that it would be virtually impossible for arbitrary signals to perform nontrivial BWE directly and solely in the signal domain. Frequency components beyond half of the sampling frequency cannot be directly recovered. If a mathematical model of the signal generation process can be assumed, on the other hand, BWE becomes feasible indirectly via the parameters of this model. Knowing that both the NB and WB signals are governed by the same source model, we can estimate the source parameters from the NB signal, and then use these estimates to produce a corresponding WB speech signal.

Here, we restrict our view to speech signals. Therefore, we can make use of the well-known source-filter model of speech production. The modeling is motivated in Fig. 4. According to Fig. 4a, the human speech production process can be divided into two parts. A periodic, noise-like, or mixed excitation signal is produced by the vocal chords, or by constrictions of the vocal tract, respectively. Then the sound is shaped by the acoustic resonances of the vocal tract cavities. In analogy to the human physiology the mathematical source-filter model of speech production (Fig. 4b) consists of two parts, a signal generator producing a spectrally flat *excitation signal* u,[3] and a synthesis filter shaping the *spectral envelope* of the speech signal s. This source-filter model has been used extensively in many areas of speech signal processing, e.g., for synthesis, coding, recognition, and enhancement.

Almost all state-of-the-art approaches to bandwidth extension are build on this simple source-filter model. Following the two-stage structure of the model, the bandwidth extension is performed separately for the excitation signal u and for the spectral envelope $H(e^{j\Omega})$ of the speech signal [3]. These two constituents of the speech signal can be assumed to be mutually independent to a certain extent, such that more or less separate optimization of the two parts of the algorithm is possible. In Fig. 5 a generic block diagram of this concept is shown.

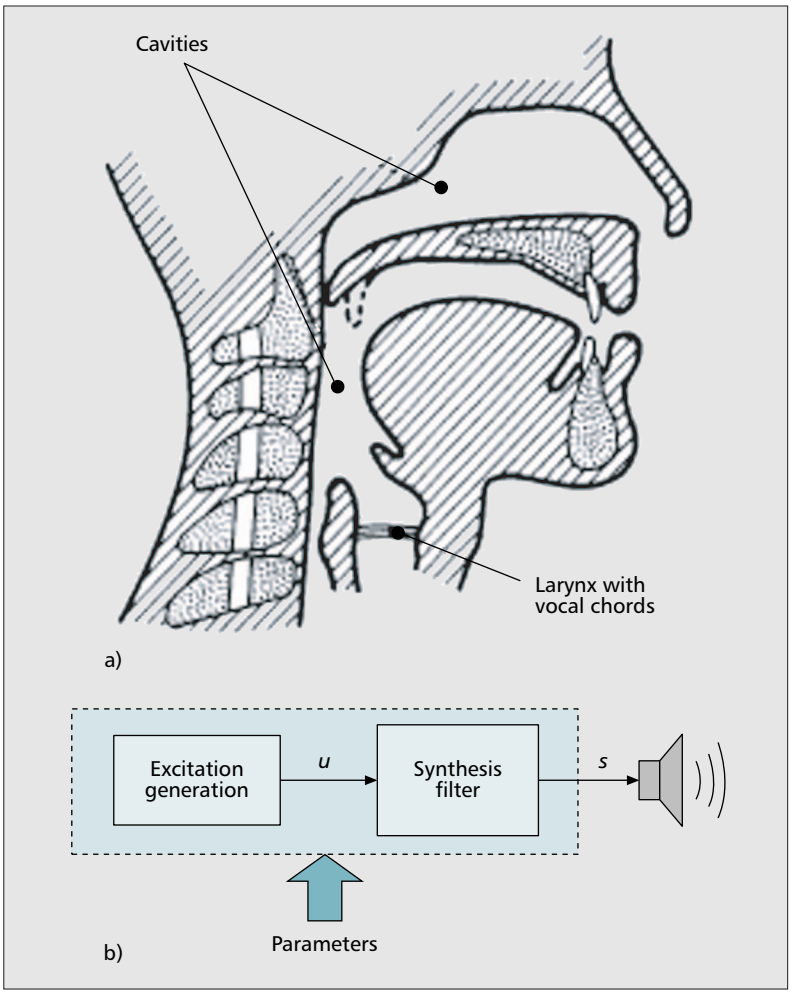

Figure 4. *Model of the speech production process: a) physiology of the human vocal tract; b) signal processing model.*

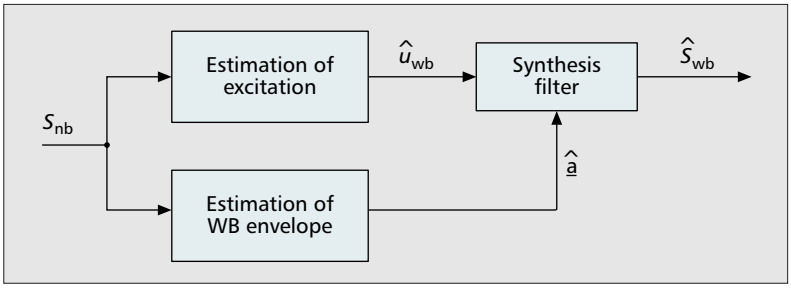

Figure 5. *Bandwidth extension with separate extension of the spectral envelope and excitation signal.*

ESTIMATION OF THE WIDEBAND SPECTRAL ENVELOPE

The bandwidth extension algorithm starts with the estimation of the spectral envelope of the wideband speech signal, see the lower signal path in Fig. 5. This block is shown in more detail in Fig. 6. In most adaptive BWE algorithms, statistical estimation methods are used which are to a certain extent similar to approaches from pattern recognition or speech recognition. The estimation scheme is based on a vector $\underline{\mathbf{x}}$ of features that is extracted from each frame of the narrowband input signal s_{nb}. Often, this feature vector is comprised of information on the spectral envelope of the narrowband speech signal (e.g., LSF or reflection coefficients, [3]) plus in addition certain features reflecting voiced/unvoiced attributes of the speech (e.g., short-term power, zero crossing rate, etc.) [4].

There are lots of different schemes in the literature for estimating the WB spectral envelope. The most important basic techniques include:

- Codebook mapping [3]
- Linear or piece-wise linear mapping [5]
- Bayesian estimation based on Gaussian mixture models (GMMs) [6] or hidden Markov models (HMMs) [4]

Within the estimation scheme, a priori knowledge on the joint behavior of the observation

[3] *Strictly speaking, the glottis signal is not spectrally flat due to the shape of the glottis pulses. However, the shape of the glottis pulse can be modeled by a glottis filter with a spectrally flat excitation* u. *In practice, the glottis filter is merged into the synthesis filter.*

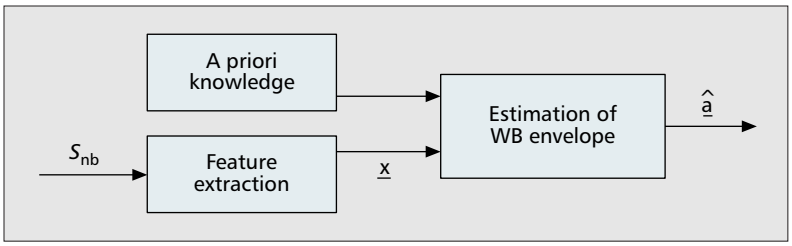

Figure 6. *Estimation of the spectral envelope.*

(feature vector) and the estimated quantity is needed. This a priori knowledge is contained in a statistical model, whose form depends on the employed estimation method. For example, in the case of codebook mapping, the statistical model comprises two LBG-trained vector quantizer codebooks for the LPC or LSF coefficients for both NB and WB speech. The statistical model has to be acquired and stored during an *offline training phase* using a database of representative WB speech signals.

The result of the estimation block is the WB spectral envelope of the speech frame, represented by the filter coefficient vector $\hat{\underline{a}}$ of the vocal tract synthesis filter from the source-filter model described above.

Extension of the Excitation Signal

The next step in the BWE system consists of substituting the missing frequency components in the excitation signal. Due to the assumed spectral flatness of the excitation signal u, and because the human ear is quite insensitive to variations of the spectral fine structure at high frequencies, the extension can be realized in a very efficient manner.

The basic functional principle of most algorithms can be described as in Fig. 7. After interpolation of the sampling rate from 8 to 16 kHz, the NB excitation $\hat{u}_{nb}$ is estimated by applying the interpolated signal $\tilde{s}_{nb}$ to the WB LPC analysis filter $1 - \hat{A}(z)$. The actual extension is performed in the block labeled HFR (for high frequency resynthesis, beyond 3.4 kHz) and LFR (for low frequency resynthesis, below 300 Hz). The techniques typically used for extension of the excitation signal are (see, e.g., [4, 7, 8] for more details):

- Mirroring, shifting or scaling of the baseband spectral components
- Generation of harmonics by nonlinear distortion and filtering
- Synthetic generation of the new frequency components

The extended frequency components are added to the estimated NB excitation. The output signal $\hat{u}_{wb}$ is the desired estimate of the WB excitation signal. Listening tests have shown that estimation of the WB spectral envelope has much more influence on the quality of the enhanced speech than extension of the excitation signal. Many of the listed techniques produce output signals with similar quality.

Performance and the State of the Art

Standalone BWE algorithms for speech have reached a stable baseline quality: the artificial WB output of a BWE system is in general preferred to NB telephone speech, even for a speaker- and language-independent setup.[4] The best results are obtained for systems trained for a specific language, or even for an individual speaker. In any case, the quality of the enhanced speech does not reach the quality of the original WB speech.

To date, BWE for speech has mostly been developed for clean input speech. The vast majority of the published approaches do not consider any adverse conditions such as additive background noise or distortion of the NB input signal. To improve acceptance in the wider range of possible applications, the robustness of BWE for speech schemes has to be increased. Important issues in this respect are robustness against additive background noises, and against input signals that differ from the model assumptions, like music. In such circumstances, at least the BWE system should be switched to a secure fallback solution.

Bandwidth Extension Techniques in Speech Coding

Artificial BWE is closely related to speech coding. In fact, some very special and effective variants of BWE techniques have been used as an integral part of various speech codecs for many years. Very prominent examples in this respect are the GSM full-rate codec and the more recent AMR-WB and AMR-WB+ codecs.

As motivated above, most of the BWE algorithms proposed in literature are based on the source-filter model of speech production. The extension of the source signal (excitation) and of the frequency response of the synthesis filter (spectral envelope) can be treated separately. The latter is much more challenging because the ear is rather insensitive with respect to coarse quantization or approximation of the excitation signal. Therefore, BWE can be implemented with great success if information on the complete (WB) spectral envelope is transmitted as side information, while the extension of the excitation is performed at the receiver without additional side information.

Baseband RELP-Codec

This idea has been used for coding of narrowband telephone speech for quite a long time to achieve bit rates below 16 kb/s with moderate computational complexity. The basic concept, which was originally proposed by Makhoul and Berouti [9] is called the baseband residual excited linear prediction (RELP) codec. The excitation signal is transmitted with a bandwidth even smaller than the standard telephone bandwidth by applying lowpass filtering and sample rate decimation by a factor of r. At the receiving end, the missing samples are replaced by zeros; thus, the baseband spectrum of the residual signal is repeated r times. Due to this spectral mirroring, this type of speech codec produces a slightly metallic sound, especially for female voices.

The transmission of the linear prediction

[4] *The results of many informal listening tests reported in research papers (e.g., [7, 8]) consistently indicate a preference for BWE-processed speech signals. However, to our knowledge, formal listening tests of standalone BWE algorithms have not been performed to date.*

coefficients may be considered the transmission of *side information* for the construction of the decoded signal in the extension band. This concept of the baseband RELP was later refined for different standardized speech codecs. A prominent example is the basic full-rate speech codec of the GSM system.

Split-Band CELP Wideband Speech Coding

More recently, BWE has been applied in the context of WB speech coding (e.g., in the 3GPP/ETSI AMR-WB codec). In this approach, code excited linear predictive (CELP) coding is applied to speech components up to 6.4 kHz, and artificial BWE is used to synthesize a supplementary signal for the narrow frequency range from 6.4 to 7 kHz. The extension is supported by transmitting different amounts of side information that controls the spectral envelope and level of noise excitation in the extension band. A more flexible version of this approach is used in the AMR-WB+ codec, which produces spectral components up to 16 kHz.

Somewhat related approaches have been introduced in the context of MPEG general audio coding as spectral band replication (SBR). Basic differences are that SBR does not rely on a signal model, and the extension starts with a signal that already has a cutoff frequency of, say, 8 kHz. The psycho-acoustic characteristics of the human ear can be exploited, especially the reduced resolution at higher frequencies. SBR has successfully been used to enhance the coding efficiency of MP3 (MP3pro) and Advanced Audio Coding (AAC-plus) [10].

Conclusions

Standalone artificial bandwidth extension approaches have the appeal of producing more natural sounding speech quality than conventional narrowband telephone connections. Besides improving quality perception, the enhanced speech signal has the benefit of reducing listening effort. Although the basic techniques are comparably young, BWE is on the threshold of practical implementation. Specialized BWE techniques with side information are already in use within several standardized speech codecs.

However, it has been shown in the literature that we cannot expect standalone BWE systems to produce the same speech quality as obtained by "true" wideband speech coding. Therefore, BWE should not be regarded as an alternative to wideband speech coding. We have outlined several application scenarios in this contribution in which both wideband coding and BWE complement each other. Thus, the introduction of BWE methods in terminals and networks may help to speed up the introduction of true wideband speech coding in the near future.

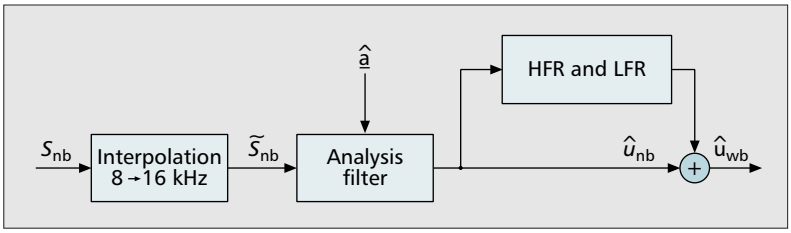

Figure 7. *Extension of the excitation signal.*

References

[1] H. Ding, "Wideband Audio over Narrowband Low-Resolution Media," *Proc. ICASSP*, vol. 1, Montreal, Canada, May 2004, pp. 489–92.

[2] B. Geiser, P. Jax, and P. Vary, "Artificial Bandwidth Extension of Speech Supported by Watermark-Transmitted Side Information," *Proc. INTERSPEECH*, Lisbon, Portugal, Sept. 2005.

[3] H. Carl and U. Heute, "Bandwidth Enhancement of Narrow- Band Speech Signals," *Proc. EUSIPCO*, vol. 2, Edinburgh, Scotland, Sept. 1994, pp. 1178–81.

[4] P. Jax, "Bandwidth Extension for Speech," Chapter 6, *Audio Bandwidth Extension*, Larsen and Aarts, Eds., Wiley, Nov. 2004.

[5] Y. Nakatoh, M. Tsushima, and T. Norimatsu, "Generation of Broadband Speech from Narrowband Speech using Piecewise Linear Mapping," *Proc. EUROSPEECH*, vol. 3, Rhodos, Greece, Sept. 1997, pp. 1643–46.

[6] K.-Y. Park and H. S. Kim, "Narrowband toWideband Conversion of Speech using GMM-based Transformation," *Proc. ICASSP*, vol. 3, Istanbul, Turkey, June 2000, pp. 1847–50.

[7] J. A. Fuemmeler, R. C. Hardie, and W. R. Gardner, "Techniques for the Regeneration of Wideband Speech from Narrowband Speech," *EURASIP J. Applied Sig. Proc.*, vol. 2001, no. 4, Dec. 2001, pp. 266–74.

[8] C.-F. Chan and W.-K. Hui, "Wideband Re-Synthesis of Narrowband CELP Coded Speech Using Multiband Excitation Model," *Proc. ICSLP*, vol. 1, Philadelphia, PA, Oct. 1996, pp. 322–25.

[9] J. Makhoul and M. Berouti, "High-Frequency Regeneration in Speech Coding Systems," *Proc. ICASSP*, Washington, DC, Apr. 1979, pp. 428–31.

[10] M. Dietz *et al.*, "Spectral Band Replication: A Novel Approach in Audio Coding," *Proc. 112th AES Convention*, Paper 5553, Munich, Germany, Apr. 2002.

Biographies

Peter Jax (Peter.Jax@thomson.net) received a Dipl.-Ing. degree in electrical engineering in 1997 and a Dr.-Ing. degree in 2003, both from RWTH Aachen University, Germany. Between 1997 and 2005 he worked as research assistant and senior researcher at the Institute of Communication Systems and Data Processing of RWTH Aachen University. Since 2005 he has been head of the Digital Audio Processing laboratory in Thomson Corporate Research, Hannover, Germany. His research interests include speech enhancement, speech and audio compression, coding theory, and statistical estimation theory.

Peter Vary (peter.vary@ind.rwth-aachen.de) received a Dipl.-Ing. degree in electrical engineering in 1972 from the University of Darmstadt, Germany. In 1978 he received a Ph.D. degree from the University of Erlangen-Nuremberg, and in 1980 he joined Philips Communication Industries (PKI), Nuremberg, Germany. He became head of the Digital Signal Processing Group, which made substantial contributions to the development of GSM. Since 1988 he has been a professor at Aachen University of Technology, Germany, and head of the Institute of Communication Systems and Data Processing. His main research interests are speech coding, joint source-channel coding, error concealment, and speech enhancement including noise suppression, acoustic echo cancellation, and artificial wideband extension.