Artificial Bandwidth Extension of Speech Signals

Peter Jax
Institute of Communication Systems and Data Processing (IND)
Aachen University (RWTH), Templergraben 55, 52056 Aachen, Germany
E-Mail: jax@ind.rwth-aachen.de

Abstract

In today's digital telephone networks the acoustic bandwidth of speech signals is still limited to the range of about 300 Hz to 3.4 kHz due to the evolution from analogue transmission systems. This is the reason for the limited quality and intelligibility of "telephone speech". The objective of this contribution is the artificial bandwidth extension of speech signals. The aim is to improve the speech quality at the receiving point without transmitting any additional side information about the original wideband speech signal across the telephone link.

1. Introduction

The limited acoustic bandwidth of today's public telephone networks originates from the former analogue transmission techniques. The limitation to a frequency range of about 0.3 to 3.4 kHz causes the typical sound of the *narrowband* telephone speech. In the transition to digital transmission the upper frequency limit of 3.4 kHz has been retained (passband up to 3.4 kHz, sampling frequency $f_\mathrm{s} = 8$ kHz), whereas the lower frequency limit may be somewhat below 300 Hz.

Listening experiments have shown that the acoustic bandwidth of speech signals contributes significantly to the perceived speech quality [1], which is measured in terms of the *mean opinion score* (MOS). In comparison to telephone speech, typical *wideband speech* with a frequency range of 50 Hz to 7 kHz yields a considerable gain of up to about 1.3 MOS points.

Although the sentence intelligibility of clean telephone speech is about 99%, the intelligibility of meaningless syllables is roughly 90%, only. As a result, we sometimes need a spelling alphabet to communicate words that cannot be understood from the context, such as unknown names. Improving the intelligibility of syllables makes the communication more comfortable and less strenuous in many cases, i.e., the *listening effort* can be reduced.

True digital wideband speech communication can be achieved by redesigning the transmission link, i.e., by introducing new speech codecs on both sides of the link. Actually, several wideband speech coding schemes have been developed for the increased acoustic bandwidth (50 Hz – 7 kHz). Already in the 1980s the G.722 codec was standardized for teleconferencing and ISDN telephony. As yet this codec has not found widespread introduction into ISDN. Recently, the so-called *adaptive multi-rate wideband* (AMR-WB) speech codec was developed and standardized for mobile radio systems such as GSM and UMTS. For the future the gradual introduction of wideband terminals can be expected. However, for a long transitional period mixed telephone networks with both narrowband and wideband terminals will exist due to economical reasons.

An approach to enhance the perceived acoustic bandwidth based on the information from the available narrowband speech is *artificial bandwidth extension* (BWE) [2, 3] at the receiving end. The problem of BWE is illustrated in Fig. 1: the original wideband (wb) signal s_wb is band-pass filtered prior to analogue-to-digital conversion and transmission over the telephone network. At the receiving terminal only the narrowband (nb) signal s_nb is available. By artificial bandwidth extension an estimate $\tilde{s}_\mathrm{wb}$ of the wideband speech is produced by adding some artificial low and/or high frequency signal components. Although true wideband speech quality cannot be obtained by artificial bandwidth extension, BWE represents a very attractive enhancement of any receiving wideband terminal as long as there are sending narrowband terminals in the network. In this paper the bandwidth extension of speech signals towards higher frequencies is addressed. The high frequency band will be called the *extension band* (eb) in the following.

2. Bandwidth Extension Algorithm

The key point of the bandwidth extension algorithm is to exploit implicit redundancy of the speech production process as proposed in [2, 3]. The linear source-filter model of speech, widely used in speech coding and recognition, consists of an *auto-regressive* (AR) filter (corresponding to the vocal tract) and a source producing a spectrally flat excitation (cf. Fig. 2). According to this model the algorithm for bandwidth extension is divided into two tasks, which are to a certain extent mutually independent [2]: (1) the *extension of the spectral envelope* of the speech signal and (2) the *extension of the excitation signal*.

In Fig. 3 a block diagram of our algorithm [4] is shown. It is assumed that the bandlimited input signal has already been interpolated to a sampling frequency f_s that is sufficient to represent the extended wideband speech signal (e.g. $f_\mathrm{s} = 16$ kHz). The signal is processed on a frame-by-frame basis with a frame-size of 20 ms.

The first step in the bandwidth extension algorithm is the *estimation of the spectral envelope* of the original wideband speech signal. In Fig. 3 this task is performed by the upper blocks. The resulting coefficients $\tilde{\mathbf{a}}_\mathrm{wb}$ describe the estimated wideband spectral envelope, i.e., the all-pole (vocal tract) filter of the source-filter model.

The core of the spectral envelope estimation within our BWE system is the estimation of the spectral envelope of the extension band (eb), represented by the cepstral vector $\tilde{\mathbf{y}}_\mathrm{eb}$. In Fig. 3 the cor-

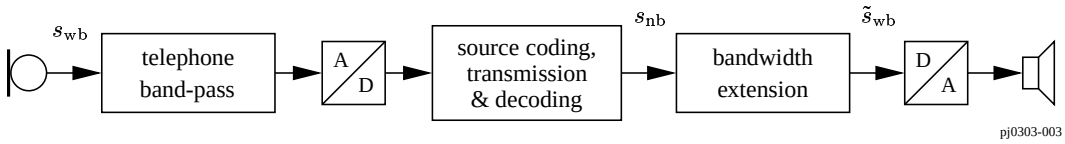

Figure 1: Bandwidth extension in digital speech transmission.

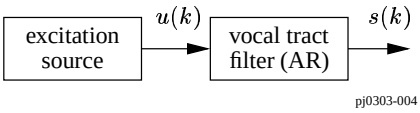

Figure 2: Source-filter model of speech production.

responding block is shaded in gray. The estimation is based on the observation of a feature vector $\mathbf{x}$ that is extracted from the narrowband speech signal $s_{\mathrm{nb}}(k)$. A technique is used that is related to pattern recognition and relies on a statistical hidden Markov model (HMM) of the speech generation process. The HMM is trained offline using a large data base of wideband speech signals. The estimate $\tilde{\mathbf{y}}_{\mathrm{eb}}$ is combined with the narrowband signal frame in a short-term power spectrum domain. The (wideband) AR coefficients $\tilde{\mathbf{a}}_{\mathrm{wb}}$ are obtained via the auto-correlation function using standard linear prediction analysis techniques.

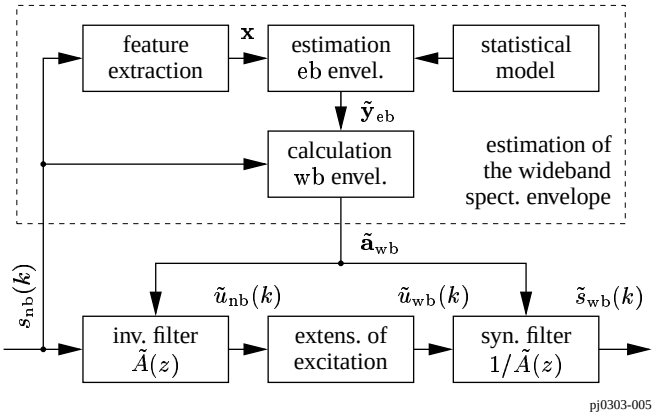

Figure 3: Block diagram of the proposed BWE algorithm.

The wideband spectral envelope is represented by the coefficients $\tilde{\mathbf{a}}_{\mathrm{wb}}$ of the all-pole vocal tract filter $1/\tilde{A}(z)$ of the (wideband) source model. By applying the corresponding FIR analysis filter $\tilde{A}(z)$ to the narrowband input signal $s_{\mathrm{nb}}(k)$, an estimate $\tilde{u}_{\mathrm{nb}}(k)$ of the narrowband excitation signal (prediction residual) is derived, since the analysis filter is the inverse of the vocal tract (synthesis) filter. Thus, the estimated excitation signal will be more or less spectrally flat within the frequency band of the narrowband speech signal $s_{\mathrm{nb}}(k)$.

The *extension of the excitation signal* converts the narrowband excitation signal $\tilde{u}_{\mathrm{nb}}(k)$ into an extended version $\tilde{u}_{\mathrm{wb}}(k)$ by exploiting the spectral flatness. The estimated wideband excitation signal $\tilde{u}_{\mathrm{wb}}(k)$ is fed into the wideband all-pole synthesis filter $1/\tilde{A}(z)$ to synthesize the enhanced output speech $\tilde{s}_{\mathrm{wb}}(k)$.

Our approach differs from most of the methods known from literature, e.g. [2, 3, 5, 6, 7], mainly in the following aspects:

1. The proposed system uses the same AR coefficients $\tilde{\mathbf{a}}_{\mathrm{wb}}$ for the analysis filter $\tilde{A}(z)$ and the synthesis filter $1/\tilde{A}(z)$. The transfer functions of these two filters are inverse to each other. Thus, transparency of the system with respect to the narrowband speech $s_{\mathrm{nb}}(k)$ is guaranteed, since the narrowband components of the excitation signal are not modified.

2. The estimation is based on a *hidden Markov model* (HMM) of the speech generation process. Thereby, our estimator can exploit information from preceding signal frames to improve the estimation quality.

3. The estimation of the spectral envelope of the extension band is based on the cepstral representation $\tilde{\mathbf{y}}_{\mathrm{eb}}$, resulting in a specific mean square error criterion which has subjective relevance.

For more details on the individual sub-systems of our BWE algorithm the interested reader is referred to [8, 4].

3. Conclusions

In this paper an algorithm for artificial bandwidth extension has been described that is based on a linear source-filter model of the speech signal. According to the two-stage structure of the source-filter model, the bandwidth extension algorithm is divided into two sub-systems that are mutually independent to a large extent [2]. The BWE algorithm proposed in the paper inherently guarantees transparency of the system with respect to the narrowband input signal.

The principal part of the algorithm is assigned to the human vocal tract: it analyses and extends the envelope of the frequency spectrum of the speech signal. The estimation of the power spectrum representing the spectral envelope within the high frequency band of the speech signal is performed in a weighted cepstral domain, aiming at the minimization of the mean square error (MSE). This minimization of the MSE corresponds to the explicit optimization of the perceptually relevant mean log spectral distortion measure.

The actual estimation procedure is based on a hidden Markov model of the signal source. By taking into account the statistics of the HMM state sequence, our MMSE estimation rule additionally considers the observations from previous signal frames. In listening tests, the additional utilization of the HMM yields a significant reduction of unnatural artifacts in the enhanced speech.

4. References

[1] W. Krebber, *Sprachübertragungsqualität von Fernsprech-Handapparaten*, Ph.D. thesis, RWTH Aachen, 1995, (in German).

[2] H. Carl, *Untersuchung verschiedener Methoden der Sprachkodierung und eine Anwendung zur Bandbreitenvergrößerung von Schmalband-Sprachsignalen*, Ph.D. thesis, Ruhr-Universität Bochum, Bochum, Germany, 1994, (in German).

[3] Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Trans. Speech and Audio Proc.*, vol. 2, no. 4, pp. 544–548, Oct. 1994.

[4] P. Jax and P. Vary, "On artificial bandwidth extension of speech signals," *accepted for publication in SIGNAL PROCESSING*.

[5] N. Enbom and W. B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," in *IEEE Speech Coding Worksh.*, Porvoo, Finland, Sept. 1999, pp. 171–173.

[6] J. Epps and W. H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," in *IEEE Speech Coding Worksh.*, Porvoo, Finland, Sept. 1999, pp. 174–176.

[7] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM-based transformation," in *Proc. of ICASSP*, Istanbul, Turkey, June 2000, vol. 3, pp. 1847–1850.

[8] P. Jax, *Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds*, Ph.D. thesis, Aachen University (RWTH), Aachen, Germany, vol. 15 of P. Vary (ed.) Aachener Beiträge zu digitalen Nachrichtensystemen, 2002.