ROBUST WIDEBAND ENHANCEMENT OF SPEECH BY COMBINED CODING AND ARTIFICIAL BANDWIDTH EXTENSION

Bernd Geiser, Peter Jax, and Peter Vary

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

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Templergraben 55, Aachen, Germany

ABSTRACT

Techniques for *artificial bandwidth extension* (BWE) of speech aim at the reconstruction of the complete *wideband spectrum* (i.e., an acoustic bandwidth of for instance $50\,\mathrm{Hz} - 7\,\mathrm{kHz}$) from the knowledge of the narrowband speech signal, which is often limited in bandwidth by a "telephone bandpass" ($300\,\mathrm{Hz}{-}3.4\,\mathrm{kHz}$). This (still blind) reconstruction can be achieved by the estimation of parameters of a source model for speech production given the knowledge of the narrowband signal. The performance of this estimation can be shown to be theoretically bounded due to insufficient mutual information between the lower and the upper subbands (refer to [1] and [2]). Nevertheless, a significantly better wideband speech quality can be accomplished with the additional transmission of side information which can then be taken into account in the respective estimation rules. This leads to a new estimation scheme which tolerates side information transmission over channels of rather low capacity and is closely related to "softbit source decoding" [3].

1. INTRODUCTION

The challenging and crucial task in artificial bandwidth extension (BWE) of narrowband (telephone) speech signals is the estimation of the spectral envelope of the missing subband, e.g., $3.4\,\mathrm{kHz} - 7\,\mathrm{kHz}$, since the human perception is most sensitive regarding variations of the coarse spectral shape of speech signals. The spectral fine structure is, although by no means irrelevant, not as important and can consequently be reproduced by, e.g., spectral replicas of the corresponding narrowband signal or by explicit signal generation.

Since BWE can only deliver a wideband signal of limited quality [1], it is reasonable to provide some additional information which can support the envelope estimation and the reproduction of the spectral fine structure for the missing subband. Our paper concentrates on side information to aid the envelope estimation, because of its vital importance to BWE[1]. Therefore we will, first, briefly review our stand-alone BWE algorithm as described in [1] and [2] and then introduce *BWE with side information* with an appropriate estimation scheme. The revised estimation rules will be discussed and interpreted as a form of *error concealment*. Finally, we will present a practical example application for BWE with side information and briefly state some simulation results for this case.

2. ARTIFICIAL BANDWIDTH EXTENSION

The signal flow chart of our BWE algorithm is depicted in Figure 1. After an *interpolation* of the narrowband signal to wideband sample rate, a *feature vector* $\mathbf{x}_\mathrm{f}$ is computed. Then, by means of a pre-trained statistical *hidden Markov model* (HMM), an estimate for the wideband spectral envelope is determined in terms of *linear prediction* (LP) coefficients $\mathbf{a}_{\mathrm{wb,BWE}}$. These wideband coefficients are used for *analysis filtering* of the interpolated narrowband signal. After the *extension of the* resulting *excitation* (see [1], [2]), the inverse *synthesis filter* is applied. The choice of an excitation extension which does not alter the narrowband part leads to a BWE system which is *transparent* w.r.t. the *narrowband components*.

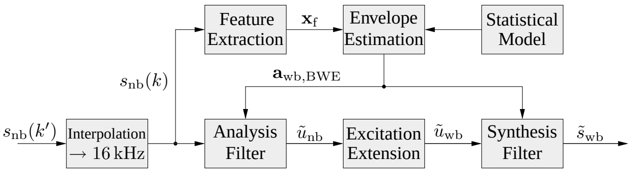

Figure 1: Stand-alone BWE algorithm.

[1] Nevertheless it is possible to transmit more side information which additionally helps to regenerate the spectral fine structure, such as a description of the harmonic structure of the signal. This information might partially be taken from the narrowband speech codec which is used in the conventional communication system.

3. BWE WITH SIDE INFORMATION

The proposed system for wideband enhancement of narrowband speech signals by combined coding and BWE is illustrated in Figure 2. At the transmitting terminal, the high band spectral envelope of the wideband input signal is analyzed and the side information is determined. The resulting *message* m is *encoded* either separately or jointly with the narrowband speech signal. At the receiver, the decoded side information is used to support the estimation of the wideband envelope within the bandwidth extension algorithm.

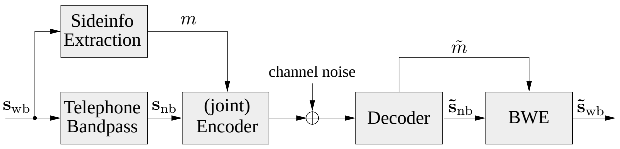
Sideinfo Extraction
m
$\tilde{m}$
channel noise
$\mathbf{s}_{\mathrm{wb}}$
Telephone Bandpass
$\mathbf{s}_{\mathrm{nb}}$
(joint) Encoder
Decoder
$\tilde{\mathbf{s}}_{\mathrm{nb}}$
BWE
$\tilde{\mathbf{s}}_{\mathrm{wb}}$

Figure 2: Overview of the transmission system.

The choice for the message m is the topic of the following section. Its usage within the decoder side BWE algorithm will be explicated in Section 3.2 and a practical application example will be presented in Section 4.

3.1. Choice of Side Information

The feature extraction from Figure 1 lacks sufficient information on the upper band's spectral envelope. In order to support the wideband envelope estimation we additionally extract a spectral representation of frequencies from $3.4\,\mathrm{kHz}$ to $7\,\mathrm{kHz}$ from the *wideband* signal (which is only available at the sending side). This subband envelope is computed by *selective linear prediction*, i.e., computation of the wideband power spectrum followed by an IDFT of its upper band components and a subsequent Levinson-Durbin recursion of order 8. The resulting subband LPC coefficients are converted into the cepstral domain and are finally quantized by a vector quantizer with a codebook of size $M \doteq 2^N$. The training of this vector quantizer shows that a $6\,\mathrm{bit}$ codebook ($M = 64$) is sufficient to achieve a mean log spectral distortion (see Equation (8)) of less than $3\,\mathrm{dB}$ for the *upper band*. This has been adopted here as our target "speech quality" since the examination of "near transparent" wideband coding (AMR-WB [4] at $23.05\,\mathrm{kbit/s}$) yields similar values [1]. Now, the side information message m is defined as the codebook index of the quantized cepstral vector for each speech frame:

$$m \in \mathbb{M} \doteq \{1, \ldots, M\}. \tag{1}$$

For a frame length of, e.g., $\tau = 20\,\mathrm{ms}$ this results in a *side information data rate* of $r_{\mathrm{s}} = 300\,\mathrm{bit/s}$. For the actual transmission, the message m is mapped to a real value (or vector) $\mathbf{x}$.

3.2. Revised Estimation Rules

There are several possibilities to make use of the side information as defined by (1) within the BWE algorithm. The most intuitive approach is to omit the envelope estimation in Figure 1 and to replace it with a block which assembles the narrowband envelope (available from the speech signal $\mathbf{s}_{\mathrm{nb}}$) and the upper band envelope (available as the side information m) and then outputs the desired coefficients $\mathbf{a}_{\mathrm{wb,BWE}}$. However, more sophisticated mechanisms take the statistical model of the BWE algorithm into account. Hence, appropriate estimation rules can be formulated, which will be addressed now.

If $\mathbf{x}_{\mathrm{recv}} \in \mathbb{R}^d$ denotes the received side information vector (which has been disturbed by some effective noise $\mathbf{n}_{\mathrm{eff}}$) and a *likelihood function* $p(\mathbf{x}_{\mathrm{recv}} | m)$ is available at the receiver, the cepstral vector $\tilde{\mathbf{c}}$ representing the upper band's envelope can, in analogy to [1] and [2], be estimated with either a *maximum likelihood* (ML), a *maximum a posteriori* (MAP) or a *minimum mean square error* (MMSE) rule. The latter one reads:

$$\tilde{\mathbf{c}}_{\mathrm{MMSE}} \doteq \sum_{m \in \mathbb{M}} \hat{\mathbf{c}}_m \cdot P(m | \mathbf{x}_{\mathrm{recv}}). \tag{2}$$

So MMSE estimation essentially is a weighted sum of the M vector codebook centroids $\hat{\mathbf{c}}_m$ with a posteriori probabilities $P(m | \mathbf{x}_{\mathrm{recv}})$ as weighting factors. These probabilities can be computed using a priori probabilities $P(m)$ and the likelihood function $p(\mathbf{x}_{\mathrm{recv}} | m)$ by applying the mixed form of Bayes' rule:

$$P(m | \mathbf{x}_{\mathrm{recv}}) = \frac{P(m) \cdot p(\mathbf{x}_{\mathrm{recv}} | m)}{\sum\limits_{m' \in \mathbb{M}} P(m') \cdot p(\mathbf{x}_{\mathrm{recv}} | m')}, \tag{3}$$

where the marginal density $p(\mathbf{x}_{\mathrm{recv}})$ is expressed as the sum in the denominator. This approach is closely related to *error concealment* by *softbit speech decoding* (see [3], Equation (18))[2].

Our new *combined estimation approach* extends the calculation of the a posteriori probabilities (3) and reintroduces dependencies on the narrowband features $\mathbf{x}_{\mathrm{f}}$ (compare to Figure 1). In (3) we substitute the likelihood function $p(\mathbf{x}_{\mathrm{recv}} | m)$ by the conditional joint probability density function (PDF) of the received side information $\mathbf{x}_{\mathrm{recv}}$ *and* the features $\mathbf{x}_{\mathrm{f}}$. These new a posteriori probabilities can be used in several estimation rules. For the case of MMSE estimation their insertion into (2) gives:

$$\tilde{\mathbf{c}}_{\text{J-MMSE}} \doteq \sum_{m \in \mathbb{M}} \hat{\mathbf{c}}_m \cdot \frac{P(m) \cdot p(\mathbf{x}_{\mathrm{recv}}, \mathbf{x}_{\mathrm{f}} | m)}{\sum\limits_{m' \in \mathbb{M}} P(m') \cdot p(\mathbf{x}_{\mathrm{recv}}, \mathbf{x}_{\mathrm{f}} | m')}, \tag{4}$$

[2] For simplicity, our formulation of the MMSE criterion in (2) and (3) only uses a priori knowledge of order 0 (AK0), i.e., the state probabilities $P(m)$. As shown in [1], [2] and [3], the estimation scheme can also be formulated for HMMs of higher order.

which we label "*joint MMSE* estimation rule". The conditional joint PDF can be assumed to be separable:

$$p(\mathbf{x}_{\text{recv}}, \mathbf{x}_{\text{f}} \big| m) = p(\mathbf{x}_{\text{recv}} \big| m) \cdot p(\mathbf{x}_{\text{f}} \big| m), \qquad (5)$$

where the first factor is our likelihood function and the latter one is an integral part of the envelope estimation block and thus available within the BWE algorithm.

Now these two factors independently contribute to the estimation of the spectral envelope in the missing upper band. In case the side information transmission is unreliable we will find $p(\mathbf{x}_{\text{recv}} \big| m) \approx p_1 \equiv \text{const}$ for $m \in \mathbb{M}$ and the factor $p(\mathbf{x}_{\text{f}} \big| m)$ will primarily determine the estimation result since then:

$$\begin{aligned} & \sum_{m \in \mathbb{M}} \hat{\mathbf{c}}_m \cdot \frac{P(m) \cdot p(\mathbf{x}_{\text{recv}} \big| m) \cdot p(\mathbf{x}_{\text{f}} \big| m)}{\sum\limits_{m' \in \mathbb{M}} P(m') \cdot p(\mathbf{x}_{\text{recv}} \big| m') \cdot p(\mathbf{x}_{\text{f}} \big| m')} \\ \approx \; & \sum_{m \in \mathbb{M}} \hat{\mathbf{c}}_m \cdot \frac{p_1 \cdot P(m) \cdot p(\mathbf{x}_{\text{f}} \big| m)}{p_1 \cdot \sum\limits_{m' \in \mathbb{M}} P(m') \cdot p(\mathbf{x}_{\text{f}} \big| m')} \\ = \; & \sum_{m \in \mathbb{M}} \hat{\mathbf{c}}_m \cdot \frac{P(m) \cdot p(\mathbf{x}_{\text{f}} \big| m)}{\sum\limits_{m' \in \mathbb{M}} P(m') \cdot p(\mathbf{x}_{\text{f}} \big| m')}. \qquad (6) \end{aligned}$$

So totally unreliable side information transmission transforms (4) into the usual MMSE estimation rule for BWE from [1] and [2]. Vice versa, if the features $\mathbf{x}_{\text{f}}$ do not allow for a dependable estimation (i.e., the second factor in (5) is constant), the received side information can dominate the estimation and joint MMSE estimation (4) simplifies to (2).

The described mechanism can be interpreted as an improved form of error concealment which utilizes more than one source of information for its parameter estimation. This can be seen as a kind of diversity "reception" (see also [3]) where one part of the information which is available to the decoder is actually transmitted (coding) and the other part is provided by the statistical model of the BWE algorithm.

4. EXAMPLE APPLICATION

We will now briefly present a practical application for BWE with side information. This application makes use of our J-MMSE estimation rule (4).

4.1. Backwards Compatible Wideband Telephony

In the future it will be desirable to convert existing telecommunication networks from narrowband to wideband speech coding and transmission. Since this means a major change to the whole system, a certain transition period has to be allowed for. Within this period, where both wide- and narrowband terminals are active, BWE is an attractive option which retains backwards compatibility w.r.t. both the sending terminal and the communication network.

In case the backwards compatibility w.r.t. the network is vital and changes to all terminals are applied more easily, BWE with side information transmission might become an option. There is no need to alter the network since there exists the possibility of *embedding* the side information as, e.g., defined by (1) into the narrowband speech signal. The appropriate technology is known as "digital watermarking".

4.2. The "Watermark Channel"

In [5] we describe BWE with side information transmission via a digital "watermark channel". The principle is illustrated by Figure 3. Watermarking in this context is rather to be interpreted as a kind of "steganography". The side information bitstream is "hidden" in the narrowband speech signal. A related approach of using information hiding to enable wideband speech has been taken by [6]. Our watermarking scheme (Λ-QIM) is similar to Chen's proposal [7] for "Quantization Index Modulation" (QIM).

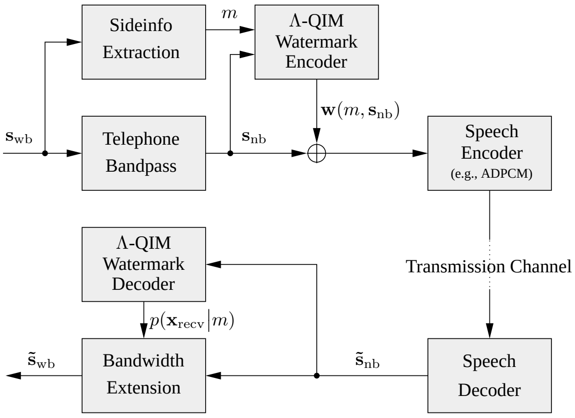

Figure 3: BWE with side information transmission via a digital "watermark channel".

The Λ-QIM watermark decoder is able to output a likelihood function $p(\mathbf{x}_{\text{recv}} \big| m)$ which can be inserted into several estimation rules, e.g., (4). If the equivalent channel noise is assumed to be i.i.d. Gaussian, the likelihood function may take the following form:

$$p(\mathbf{x}_{\text{recv}} \big| m) \approx \frac{1}{\sqrt{(2\pi\tilde{\sigma}^2)^n}} \exp\left(-\frac{\Delta_m}{2\tilde{\sigma}^2}\right), \qquad (7)$$

where $\tilde{\sigma}$ is an estimate for the standard deviation of the effective noise $\mathbf{n}_{\text{eff}}$ and n is the dimension of the signal vectors. The values Δ_m $(m \in \mathbb{M})$ can be directly measured within the watermark decoder (see [5]).

4.3. Performance Evaluation

The performance of bandwidth extension with and without side information and, for reference, of wideband speech coding have been measured utilizing the objective subband spectral distortion measure (i.e., the distance between the subband spectral envelopes) of the original and the bandwidth extended (or wideband coded) speech:

$$d^2_{\mathrm{LSD,\,hb}} \doteq \frac{1}{2\pi} \int_{-\pi}^{\pi} (20 \lg \frac{\sigma_{\mathrm{rel}} \cdot |A_{\mathrm{hb,BWE}}(e^{j\Omega})|}{\sigma_{\mathrm{rel,BWE}} \cdot |A_{\mathrm{hb}}(e^{j\Omega})|})^2 \, \mathrm{d}\Omega, \qquad (8)$$

with $\Omega \doteq \pi \cdot \mathrm{sgn}(f) \cdot (|f| - f_1)/(f_2 - f_1)$, $f_1 \leq |f| \leq f_2$. The considered subband (hb $\hat{=}$ high band) is bounded by $f_1 = 3.4\,\mathrm{kHz}$ and $f_2 = 7\,\mathrm{kHz}$. A_{hb} represents the frequency response of the linear prediction filter for this band and σ_{rel} is the high band gain normalized by the narrowband gain. In practice, $d_{\mathrm{LSD,\,hb}}$ is approximated by using the first 9 coefficients $c_0, \ldots, c_8$ of the real cepstrum in the high subband:

$$d^2_{\mathrm{LSD,hb}} \approx \left(\frac{10}{\ln 10}\right)^2 \left((c_0 - c_{0,\mathrm{BWE}})^2 + 2\sum_{i=1}^{8}(c_i - c_{i,\mathrm{BWE}})^2\right). \quad (9)$$

Our stand-alone BWE algorithm achieves mean log spectral distortions down to about $6\,\mathrm{dB}$. On the other hand, the examination of the AMR-WB codec at $23.05\,\mathrm{kbit/s}$ — an example of “near transparent” wideband speech coding — gives a mean distortion of about $3\,\mathrm{dB}$.

4.4. Simulation Results

Simulations of our new transmission scheme with a side information data rate of $300\,{}^{\mathrm{bit}}/_{\mathrm{s}} = 6\,{}^{\mathrm{bit}}/_{\mathrm{frame}}$ show a clear advantage over stand-alone BWE in terms of subband spectral distortion. In Figure 4 some results are shown for transmission of the watermarked signal over several narrowband speech codecs.

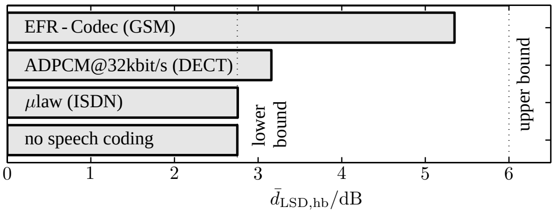

Figure 4: Simulation results for BWE with side information transmission via “digital watermarks”.

Totally reliable watermark transmission yields values for $\bar{d}_{\mathrm{LSD,hb}}$ which are *lower bounded* by the quantization distortion of the $6\,\mathrm{bit}$ vector quantizer described in Section 3.1 ($2.75\,\mathrm{dB}$). The J-MMSE estimation rule (4) ensures that the distortion measure is *upper bounded* by the value which can already be achieved with stand-alone BWE ($\approx 6\,\mathrm{dB}$).

Informal listening tests confirm our simulation results and show a consistent preference for BWE with watermark transmitted side information over stand-alone BWE. However, the AMR-WB codec is still superior due to a yet suboptimal excitation extension.

5. CONCLUSION

We have introduced a system for artificial bandwidth extension of speech signals which benefits from the transmission of additional *side information* such that better performance is achieved than the theoretical bound for pure feature based BWE suggests. The new *joint MMSE* estimation rule provides an *error concealment* mechanism which ensures that unreliable side information transmission does not reduce the resulting speech quality to values lower than already achieved by stand-alone BWE. In typical cases, however, the transmitted side information biases the envelope estimation towards the correct decision.

Additionally, an example transmission system has been presented where the side information is *embedded* into the narrowband speech signal using techniques of *digital watermarking*. At the decoder, the retrieved watermark message can be converted into a quantized representation for the spectral envelope of the missing upper subband. Depending on the codec used for speech transmission, our new transmission scheme yields a significant gain in terms of spectral distortion. This measure correlates reasonably well with the perceived speech quality.

6. REFERENCES

[1] Peter Jax, *Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds*, Ph.D. thesis, RWTH Aachen University, Aachen, Germany, 2002.

[2] Peter Jax, “Bandwidth extension for speech,” in *Audio Bandwidth Extension*, Erik Larsen and Ronald M. Aarts, Eds., chapter 6, pp. 171–236. Wiley and Sons, Nov. 2004.

[3] Tim Fingscheidt and Peter Vary, “Softbit speech decoding: A new approach to error concealment,” *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 3, pp. 240–251, Mar. 2001.

[4] 3GPP TS 26.290, “Extended AMR wideband codec; transcoding functions,” Sept. 2004.

[5] Bernd Geiser, Peter Jax, and Peter Vary, “Artificial bandwidth extension of speech supported by watermark transmitted side information,” in *Proc. of European Conf. on Speech Communication and Technology (INTERSPEECH / EUROSPEECH)*, Lisbon, Portugal, Sept. 2005.

[6] Heping Ding, “Wideband audio over narrowband low-resolution media,” in *Proc. of ICASSP*, Montreal, Canada, May 2004, vol. 1, pp. 489–492.

[7] Brian Chen and Gregory W. Wornell, “Quantization index modulation: A class of provably good methods for digital watermarking and information embedding,” *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.