Enhancement of Bandlimited Speech Signals

Peter Jax and Peter Vary

Institute of Communications Systems and Data Processing,
Aachen University of Technology, Templergraben 55, D-52056 Aachen, Germany
E-mail: jax@ind.rwth-aachen.de

Abstract— **In this contribution we present an algorithm to produce wideband speech from bandlimited "telephone speech". The narrowband input signal is classified into a limited number of speech sounds for which the information about the wideband spectral envelope is taken from a pre-trained codebook. For the codebook search algorithm a statistical approach based on a hidden *Markov* model is used, which takes different features of the bandlimited speech into account. Several estimators are presented which take various amounts of a priori knowledge into consideration. The algorithmic approach inherently guarantees the transparency of the system in the baseband. The enhanced speech exhibits a significantly larger quality and transparency than the input speech without introducing objectionable artifacts.**

I. Introduction

In current public telephone systems the bandwidth of the transmitted speech is typically limited due to constraints of the old analogue telephone system to a frequency range of up to 3.4 kHz. This bandwidth limitation causes the characteristic sound of "telephone speech". In the age of digital hands-free communication the demand for improved speech quality and increased speech intelligibility is rising. This trend is reflected by ongoing standardizations of wideband speech codecs (e.g. [1], [2], [3]). Listening experiments have shown that an improved frequency bandwidth of speech signals contributes significantly to the perceived speech quality [4], [5].

True wideband speech communication requires a modification of the transmission link by enhanced speech codecs and increased bitrates. Hence, for economical reasons the bandwidth limitation is not likely to change on a broad scale in the near future. An alternative approach towards a higher (audio) bandwidth is the bandwidth extension: missing low and high frequency components of the speech signal are recovered at the receiving side of the transmission link utilizing only the bandlimited speech.

This extension of the bandwidth of speech signals is only feasible due to mutual dependencies in the frequency bands of speech signals. One possible concept to explore these redundancies is based on a linear model of the speech production process, as it was proposed in [6], [7]: First, the parameters of the source model are estimated from the available bandlimited speech. In a second step, these parameters can then be used in combination with the source model to estimate and add the missing frequency components. Note that there exists an information-theoretic bound on the quality of the extended speech due to the limited redundancy with respect to the missing frequency components [8].

In this paper the extension of the bandwidth towards higher frequencies is treated, i.e., the input signal is assumed to contain frequencies lower than 3.4 kHz[1], and it shall be extended artificially up to 7 kHz.

II. Algorithm

A simple linear source-filter model of the speech production process, which is commonly used in speech processing algorithms, consists of an *autoregressive* (AR) filter (corresponding to the vocal tract) which is excited by a spectrally flat excitation signal. The parameters of the model, i.e., the coefficients of the AR filter as well as the characteristics of the excitation signal are time-variant.

According to this model the algorithm for bandwidth extension can be divided into two tasks, which are to a certain extent mutually independent: the extension of the spectral envelope of the speech signal and of its excitation signal [6]. A block-diagram of a resulting algorithm [9] is shown in Fig. 1. A distinctive feature of the illustrated algorithm is that all modifications of the narrowband signal are performed in one signal path — a special treatment of the baseband signal is not necessary.

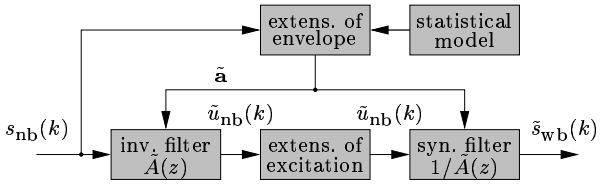

Fig. 1. Block-diagram and main signal flow of the proposed algorithm for bandwidth extension.

In Fig. 1 it is assumed that the bandlimited input signal is already sampled at a sampling frequency that is sufficient to represent the extended wideband speech signal (e.g. $f_g = 16$ kHz). The processing is performed frame by frame with a frame-size of 20

[1]This frequency band will be defined as the *baseband* in the following.

ms. The frame index will be denoted by the variable m in the following.

The first step in the bandwidth extension algorithm consists of the estimation of the spectral envelope of the original wideband speech signal in the two upper blocks. For this purpose several approaches have been proposed in literature, e.g., codebook based methods [6], [10], algorithms based on a linear mapping [11], or based on statistical estimation [7], [12], [9]. The result is a set of coefficients $\tilde{\mathbf{a}}$ of the all-pole vocal tract filter of the source model. By utilizing these filter coefficients in an FIR analysis filter $\tilde{A}(z)$ operating on the narrowband input signal $s_{\mathrm{nb}}(k)$, an estimate $\tilde{u}_{\mathrm{nb}}(k)$ of the bandlimited excitation signal can be derived. Note that the frequency response of the analysis filter is the inverse of the frequency response of the vocal tract (synthesis) filter. The algorithm for the estimation of the AR coefficients $\tilde{\mathbf{a}}$ is described in detail in section III.

The second important block of the bandwidth extension algorithm uses the estimate $\tilde{u}_{\mathrm{nb}}(k)$ of the bandlimited excitation signal and calculates an extended version $\tilde{u}_{\mathrm{nb}}(k)$ of the excitation. In this step the advantageous characteristics of the excitation signal in the model of the speech production process — especially its spectral flatness — can be utilized for a very efficient realization (see section IV). Finally, the estimated wideband excitation signal is fed into the all-pole synthesis filter $1/\tilde{A}(z)$, thus creating the enhanced output speech signal $\tilde{s}_{\mathrm{wb}}(k)$.

Unlike previous algorithms for the bandwidth extension of speech signals the proposed algorithm uses the same AR coefficients $\tilde{\mathbf{a}}$ in the analysis filter $\tilde{A}(z)$ and the synthesis filter $1/\tilde{A}(z)$. Hence, the transfer functions of these two filters are exactly mutually inverse. Due to this property of the algorithm the transparency of the system for the baseband signal components can be guaranteed — the baseband of the excitation signal is not modified during the extension of the excitation signal.

III. Spectral Envelope

The basis of the algorithm for the estimation of the AR coefficients a representing the spectral envelope of the speech signal is the introduction of a *hidden Markov model* (HMM) for the process of speech generation [9]. The states of the HMM are defined by the levels of a *vector quantizer* (VQ) of the coefficient sets of the *wideband* auto-regressive vocal tract filter: each centroid of the vector quantizer represents the spectral envelope of a typical speech sound. In the VQ codebook the AR coefficients are contained, e.g., as *line spectral frequencies* (LSF) due to the advantageous properties of this representation concerning quantization or averaging (e.g. [13]). One state $\mathcal{S}_i$ of the HMM is assigned to each entry $\hat{\mathbf{a}}_i$ of the vector quantizer codebook such that there are as many states in the HMM as there are entries in the codebook. It is further assumed that the state of the source does only change in-between two frames of the input signal. The number of codebook entries is denoted by N_S.

If wideband speech is available, e.g., in the training phase of the algorithm, the *true* state sequence can be calculated by minimizing the quantization error

$$\mathcal{S}_{\mathrm{true}} = \mathcal{S}_{i_{\mathrm{opt}}}, \quad \text{with } i_{\mathrm{opt}} = \arg\min_{i=1}^{N_S} ||\mathbf{a} - \hat{\mathbf{a}}_i||^2. \quad (1)$$

The training of the VQ codebook is performed with a sufficiently large training data set of wideband speech and utilizing the well-known LBG algorithm [14]. As a result of the LBG algorithm the codebook entries are the expected values of the AR coefficient vectors given the state of the source

$$\hat{\mathbf{a}}_i = E\{\mathbf{a}|\mathcal{S}_i\}. \quad (2)$$

Although the training of the codebook with wideband speech material guarantees that proper representatives of the different speech sounds are contained in the codebook, it also raises the challenge, that for the classification of the input signal into these speech sounds only the bandlimited signal is available. Therefore, for this case a more sophisticated estimator has to be employed which exploits the maximum relevant information contained in the narrowband speech. As illustrated in Fig. 2 the proposed estimator can be divided into the following steps:

1. To reduce the dimensionality of the estimation problem a limited number of features is extracted from each frame of the narrowband speech signal.
2. These extracted features are compared with a pre-trained statistical model of the process of speech production. Additional a priori knowledge about the state sequence can be utilized to calculate a posteriori probabilities.
3. The current speech frame is then either classified into one of the trained speech sounds (i.e. HMM states) or the AR coefficients are estimated directly.

The statistical parameters of the HMM, i.e., observation, state, and transition probabilities provide the a priori knowledge that is later-on exploited by the MAP and MMSE estimation rules. In the following subsections each of the steps is described in detail.

A. Feature Extraction

For each signal frame a b-dimensional vector $\mathbf{x}(m)$ of features is extracted from the bandlimited input signal. The elements of this vector should be selected such that the resulting feature vector contains the maximum information about the state of the HMM.

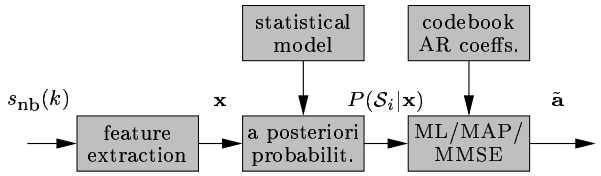

Fig. 2. Steps in the estimation of the spectral envelope of the wideband speech signal.

In this work the vector consists of the first eight cepstral coefficients $c_1, c_2, \ldots c_8$, the normalized frame energy E_n, and a gradient index d_n of the speech signal as defined in [15] as a measure for a voiced/unvoiced classification

$$d_n = \frac{\sum_{\kappa=2}^{N_\kappa} \Psi(\kappa)\,|s_{\mathrm{nb}}(\kappa) - s_{\mathrm{nb}}(\kappa-1)|}{\sqrt{\frac{1}{N_\kappa}\sum_{\kappa=1}^{N_\kappa} s_{\mathrm{nb}}^2(\kappa)}}. \qquad (3)$$

In this equation N_κ is the number of samples per frame, and the variable $\psi(\kappa)$ denotes the sign of the gradient $s_{\mathrm{nb}}(\kappa) - s_{\mathrm{nb}}(\kappa-1)$, i.e., $\psi(\kappa) \in \{-1, 1\}$, and $\Psi(\kappa) = 1/2\,|\psi(\kappa) - \psi(\kappa-1)|$.

Whereas the cepstral coefficients carry information describing the shape of the spectral envelope of the narrowband signal, the other two quantities mainly depend on properties of the excitation of the speech.

Additionally, the derivatives over time of all of the above ten primary features are included in the feature vector such that the dimension of the vector $\mathbf{x}(m)$ results in $b = 20$. The set of all observed feature vectors up to the m-th frame is defined by the observation sequence

$$\mathbf{X}(m) = \{\mathbf{x}(1), \mathbf{x}(2), \ldots \mathbf{x}(m)\}. \qquad (4)$$

B. Statistical Model

For each possible state $\mathcal{S}_i$ of the hidden *Markov* model the features $\mathbf{x}$, which are generated by the speech production process, exhibit different statistical properties. To describe these properties a statistical model consisting of the following three parts is used.

B.1 Initial State Probabilities $P(\mathcal{S}_i)$

The scalar values $\alpha_i = P(\mathcal{S}_i)$ describe the probabilities, that the HMM resides in a state $\mathcal{S}_i$ without incorporating knowledge of the feature vector $\mathbf{x}$ or of preceding or following states.

These probabilities can easily be estimated by computing the true state sequence for the wideband training material and evaluating the ratio between the number of occurrences of state $\mathcal{S}_i$ and the total number of speech frames in the training set. The resulting probability values are stored in a table such that the actual bandwidth extension algorithm can later access the a priori state probabilities by simple table lookups.

B.2 Transition Probabilities $P(\mathcal{S}_i(m+1)|\mathcal{S}_j(m))$

The variable $\beta_{ij} = P(\mathcal{S}_i(m+1)|\mathcal{S}_j(m))$ describes the probability of a transition from state $\mathcal{S}_j$ to state $\mathcal{S}_i$ from one frame to the following one. As the initial state probabilities α_i, the transition probabilities can be stored in a table, which is now two-dimensional. In the training procedure the individual entries of this table are estimated (with knowledge of the true state sequence) as the ratio between the number of occurrences of the particular transition from $\mathcal{S}_j$ to $\mathcal{S}_i$ and the total number of occurrences of state $\mathcal{S}_j$.

B.3 Observation Probabilities $p(\mathbf{x}|\mathcal{S}_i)$

Due to the high dimension of the feature vector $\mathbf{x}$ the *probability density functions* (PDF) $p(\mathbf{x}|\mathcal{S}_i)$ are modeled by *Gaussian mixture models* (GMMs): each PDF is approximated as the sum of L *Gaussian* PDFs (e.g. [16])

$$p(\mathbf{x}|\mathcal{S}_i) \approx \sum_{l=1}^{L} \rho_{il}\,\mathcal{N}(\mathbf{x}; \mu_{il}, \Sigma_{il}). \qquad (5)$$

In this equation $\mathcal{N}(\mathbf{x}; \mu_{il}, \Sigma_{il})$ denotes the l-th N-dimensional *Gaussian* distribution of the GMM with mean vector μ_{il} and covariance matrix Σ_{il}. Each *Gaussian* distribution is weighted by a scalar factor ρ_{il}, with $\sum_{l=1}^{L} \rho_{il} = 1$.

The training of the GMMs, i.e., of the quantities ρ_{il}, μ_{il} and Σ_{il}, can be performed with the iterative *expectation-maximization* (EM) algorithm (see e.g. [17], [16]), which is guaranteed to converge to a local maximum of the mean log-likelihood function

$$\mathcal{L}(\Theta, \tilde{\Theta}) = E\{\log p(\mathbf{x}, \mathcal{S}_i; \Theta)\,|\,\mathbf{x}, \tilde{\Theta}\}, \qquad (6)$$

in which $\tilde{\Theta}$ denotes the estimated parameters ρ_{il}, μ_{il} and Σ_{il}, and Θ describes the true source parameters. The EM algorithm needs a starting point for the iterative refinement, which is determined here by clustering the training data with the LBG algorithm [14].

For each state of the hidden *Markov* model one distinct GMM has to be trained, using the subset of the complete training material for which the *true* state is equal to the currently trained one.

C. Classification vs. Estimation

The goal of the codebook search algorithm is to calculate an estimate $\tilde{\mathbf{a}}$ of the wideband AR coefficients which minimizes the distance to the true coefficients $\mathbf{a}$. In this process the statistical model defined in the previous subsection can be utilized in different ways.

C.1 ML Classification

The simplest method is to select that entry $\tilde{\mathbf{a}}_{i_{\mathrm{ML}}}$ of the codebook for which the observation probability density $p(\mathbf{x}(m)|\mathcal{S}_i(m))$ is maximized (*maximum*

likelihood, ML)

$$\tilde{\mathbf{a}}_{\mathrm{ML}} = E\{\mathbf{a}|\mathcal{S}_{i_{\mathrm{ML}}}\} = \tilde{\mathbf{a}}_{i_{\mathrm{ML}}(m)}, \tag{7}$$

with

$$i_{\mathrm{ML}}(m) = \arg\max_{i=1}^{N_S} P(\mathbf{x}(m)|\mathcal{S}_i(m)). \tag{8}$$

Note, that in this classification rule neither the initial state probabilities α_i nor the transition probabilities β_{ij} are exploited.

C.2 MAP Classification

The *maximum a posteriori* (MAP) classification rule selects that codebook entry $\tilde{\mathbf{a}}_{i_{\mathrm{MAP}}}$, for which the a posteriori probability $P(\mathcal{S}_i(m)|\mathbf{X}(m))$ of the occurrence of state $\mathcal{S}_i$ given the observed sequence $\mathbf{X}(m)$ is maximized

$$\tilde{\mathbf{a}}_{\mathrm{MAP}} = E\{\mathbf{a}|\mathcal{S}_{i_{\mathrm{MAP}}}(m)\} = \tilde{\mathbf{a}}_{i_{\mathrm{MAP}}(m)}, \tag{9}$$

with

$$i_{\mathrm{MAP}}(m) = \arg\max_{i=1}^{N_S} P(\mathcal{S}_i(m)|\mathbf{X}(m)). \tag{10}$$

For the derivation of this rule it is useful to define a variable $\phi_i(m)$ as the joint probability of the partial observation sequence $\mathbf{X}(m)$ and the state $\mathcal{S}_i(m)$ at frame instant m

$$\phi_i(m) = P(\mathcal{S}_i(m), \mathbf{X}(m)). \tag{11}$$

This joint probability density can be expressed in a recursive manner in terms of the joint probabilities $\phi_j(m-1)$ at frame instant $m-1$ and the observation probability $p(\mathbf{x}(m)|\mathcal{S}_i(m))$ as

$$\phi_i(m) = \left(\sum_{j=1}^{N_S} \beta_{ij}\, \phi_j(m-1)\right)\, p(\mathbf{x}(m)|\mathcal{S}_i(m)). \tag{12}$$

The first term of this equation can be interpreted as the a priori probability of the occurrence of state $\mathcal{S}_i$ at frame instant m given the past observation sequence $\mathbf{X}(m-1)$

$$\sum_{j=1}^{N_S} \beta_{ij}\, \phi_j(m-1) = P(\mathcal{S}_i(m)|\mathbf{X}(m-1)). \tag{13}$$

Since the preceding observation vectors are unknown for the first frame, the initial values for $\phi_i(1)$ have to be calculated from the initial state probabilities α_i

$$\phi_i(1) = \alpha_i\, p(\mathbf{x}(1)|\mathcal{S}_i(1)). \tag{14}$$

Applying *Bayes* rule to Eq. (10) and using the helper variable from Eq. (11) a rule for the MAP classification can be found in which only known variables occur

$$\begin{aligned} i_{\mathrm{MAP}}(m) &= \arg\max_{i=1}^{N_S} \frac{P(\mathcal{S}_i(m), \mathbf{X}(m))}{p(\mathbf{X}(m))} \\ &= \arg\max_{i=1}^{N_S} \phi_i(m) \end{aligned} \tag{15}$$

C.3 MMSE Estimation

This method differs from the previous described classification methods in the property that its results are not bound to the entries of the codebook. The goal of the *minimum mean-square error* (MMSE) criterion is to minimize the mean squared error between the estimated AR coefficients $\tilde{\mathbf{a}}$ and the true coefficients $\mathbf{a}$, such that the following cost function is minimized

$$\mathcal{R}_{\mathrm{MSE}}(\tilde{\mathbf{a}}|\mathbf{X}) = \iint (\mathbf{a}-\tilde{\mathbf{a}})^T(\mathbf{a}-\tilde{\mathbf{a}})\, p(\mathbf{a}|\mathbf{X})\, d\mathbf{a}. \tag{16}$$

A solution for this optimization problem can be found by taking the root of the derivative of the cost function

$$\tilde{\mathbf{a}}_{\mathrm{MMSE}} = \iint \mathbf{a}\, p(\mathbf{a}|\mathbf{X})\, d\mathbf{a}. \tag{17}$$

Since we don't have an explicit model of the conditional probability $p(\mathbf{a}|\mathbf{X})$, this quantity has to be expressed indirectly in terms of the state probabilities

$$\begin{aligned} \tilde{\mathbf{a}}_{\mathrm{MMSE}} &= \iint \mathbf{a} \left[\sum_{i=1}^{N_S} p(\mathbf{a}|\mathcal{S}_i)\, P(\mathcal{S}_i|\mathbf{X})\right] d\mathbf{a} \\ &= \sum_{i=1}^{N_S} P(\mathcal{S}_i|\mathbf{X}) \underbrace{\iint \mathbf{a}\, p(\mathbf{a}|\mathcal{S}_i)\, d\mathbf{a}}_{E\{\mathbf{a}|\mathcal{S}_i\} = \hat{\mathbf{a}}_i} \end{aligned} \tag{18}$$

As shown, the integral at the right hand side of Eq. 18 yields the expected value of $\mathbf{a}$ given the occurence of state $\mathcal{S}_i$, i.e., the corresponding codebook vector $\hat{\mathbf{a}}_i$. Applying Bayes rule and substituting the helper variable ϕ_i, we obtain the following estimator

$$\tilde{\mathbf{a}}_{\mathrm{MMSE}} = \frac{\sum_{i=1}^{I} \hat{\mathbf{a}}_i\, \phi_i(m)}{\sum_{i=1}^{N_S} \phi_i(m)}. \tag{19}$$

It is interesting to note, that the conditional probabilities $p(\mathbf{a}|\mathcal{S}_i)$ can not be utilized by this estimator due to the indirect modeling of $p(\mathbf{a}|\mathbf{X})$ via the state probabilities. A superior MMSE estimator can probably be designed by directly modeling and exploiting $p(\mathbf{a}|\mathbf{X})$, however, this is not a trivial task. Alternatively, the knowledge of $p(\mathbf{a}|\mathcal{S}_i)$ can be taken into account during the training procedure of the codebook vectors $\hat{\mathbf{a}}_i$.

There is a strong relationship between the MMSE estimator and the ML/MAP classificators. If the observation probability is sufficiently high for one particular entry of the codebook, the a posteriori probability of this entry becomes dominant and the solution of Eq. (19) approximates the ML or MAP

classificators from Eq. (8) and (15). Only in cases in which the model of the observation probabilities is not sufficient to classify the input frame, the result is averaged from the most probable codebook entries. Accordingly, the proposed MMSE estimator can be regarded as a *soft* classification rule.

IV. Excitation Signal

According to the simplifying linear model of speech production the excitation signal is spectrally flat: in voiced sounds it contains sinusoids at integral multiples of the fundamental (pitch) frequency of the speech segment; all harmonics have the same amplitude. During unvoiced sounds the excitation resembles a spectrally flat noise signal.

Due to these properties of the excitation signal its extension can be performed by a modulation of the estimate $\tilde{u}_{\mathrm{nb}}(k)$ of the baseband excitation signal with a sinusoid with the modulation frequency ω_m [6], [18], [19]

$$\tilde{u}_{\mathrm{hb}}(k) = \tilde{u}_{\mathrm{nb}}(k)\, 2\cos(\omega_m k)) \tag{20}$$

This modulation in time-domain causes the desired spectral shift of the signal spectrum in frequency-domain

$$\tilde{U}_{hb}(e^{j\omega}) = \tilde{U}_{nb}(e^{j(\omega-\omega_m)}) + \tilde{U}_{nb}(e^{j(\omega+\omega_m)}) \tag{21}$$

Note, that by the modulation with a real valued sinusoid *two* shifted versions of the baseband spectrum are created. Hence, to prevent an overlapping of the signal spectrum of the modulated signal with the baseband input signal, one of the two shifted spectra must be suppressed by a high-pass filter prior to combination with the narrowband excitation signal (see Fig. 3). An alternative would be to calculate the analytic signal of the baseband excitation prior to modulation, however, this approach does not yield any advantage concerning subjective quality, computational complexity, or algorithmic delay [19].

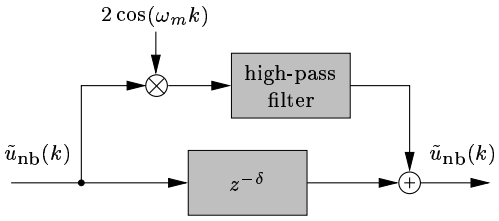

Fig. 3. Extension of the excitation signal by modulation. The algorithmic delay δ of the high-pass filter has to be compensated for in the path of the baseband signal.

By selecting the modulation frequency ω_m one out of several modulation schemes can be chosen:

- A modulation with the *Nyquist* frequency, i.e., $\omega_m = \pi$, corresponds to the method of *spectral mirroring* as proposed in [20]. In this special case the two shifted copies of the baseband spectrum coincide, such that the high-pass filter from Fig. 3 is not needed. Thus, this method can be implemented *very* efficiently.
  However, due to the cutoff frequency of the narrowband input signal, there is a spectral gap in $\tilde{u}_{\mathrm{nb}}(k)$ between 3.4 and 4.6 kHz. Furthermore, the harmonic structure in the extended frequency band does not match the low frequency components.
- To prevent the spectral gap the modulation frequency can be chosen such that the shifted spectrum starts in continuation of the baseband spectrum, e.g.,

$$\omega_m = \omega_o, \quad \text{with } \omega_o = 2\pi \frac{3.4\text{ kHz}}{f_g}. \tag{22}$$

  The variable f_g denotes the sampling rate of the signal. With this method there is still a misalignment of the harmonics in the extended frequency band.
- A further possibility to control ω_m takes the pitch frequency ω_p of the current speech frame into account: the modulation frequency is adapted in such a way that it is always an integral multiple of the estimated pitch frequency [18], e.g.,

$$\omega_m = \left\lceil \frac{\omega_o}{\omega_p} \right\rfloor \omega_p. \tag{23}$$

  By this method it can be guaranteed that the harmonics in the extended frequency band do always match the harmonic structure of the baseband. Due to the rounding operation in (23) there is in general a small spectral gap with a width in the order of the pitch frequency.
  It must be emphasized, that the pitch-adaptive modulation method reacts quite sensitive to small errors of the estimate of the pitch frequency, because these are significantly enlarged by the factor $\lceil \omega_o/\omega_p \rfloor$. Therefore, a very good pitch estimator is needed.

We have performed many informal listening tests which have shown that — on the pre-condition that the bandwidth extension of the spectral envelope works well — the human ear is amazingly insensitive to distortions of the excitation signal at high frequencies above 3.4 kHz. For example, spectral gaps of moderate width as produced by band-stop filters are almost inaudible. Further, distortions of the harmonic structure of speech at high frequencies do not significantly degrade the subjective quality of the enhanced speech signal.

Due to these properties of the human auditory system, all of the described methods for the extension of the excitation signal perform well, when a good estimate of the wideband spectral envelope is available. A good compromise between subjective

quality of the output signal and computational complexity is given by the modulation with the fixed modulation frequency of $\omega_m = \omega_o$.

V. Evaluation

For the evaluation of the proposed algorithm, codebooks of different sizes were trained. When listening to the *best* possible output of the algorithm, i.e. with knowledge of the *true* state sequence, it was found that for codebook sizes beyond $N_S = 64$ the enhanced signal is almost indistinguishable from the original wideband speech. Even for extremely small codebooks with down to $N_S = 3$ entries acceptable results can be achieved.

In many informal and comparative listening tests, all of the described algorithms yielded good results — a significant extension of the bandwidth is audible. Occasionally, there are audible artifacts, mainly during unvoiced fricatives like /s/ or /f/, which result from wrong classifications by the envelope estimation algorithm. However, the more accurate the a priori knowledge is in the algorithm, the less frequent is the occurence of such artifacts. As expected, the best results are achieved by the MMSE estimator.

VI. Complexity

The computational complexity of the proposed algorithm strongly depends on the parameterization: a good quality of the enhanced speech signal can be achieved with less than 5 WMOPS — the main part of the computational power is needed for the calculation of the features $\mathbf{x}$ of the narrowband speech signal. If the algorithm is placed behind a speech decoder, the complexity can further be reduced by adopting variables from the decoder.

VII. Conclusion

The proposed method allows a bandwidth extension of lowpass-bandlimited speech to a frequency range of up to 7 kHz. The results proove, that there is enough information in the low frequency regions to securely estimate the missing high frequency components — however, for this estimation more features of the narrowband speech than only its spectral envelope should be utilized. For this purpose the proposed statistical framework turns out to be an appropriate tool.

References

[1] CCITT, "7 kHz Audio Coding Within 64 kBit/s," Recommendation G.722, vol. Fascile III.4 of Blue Book, 1988.

[2] 3GPP, "Speech Codec Speech Processing Functions; AMR Wideband Speech Codec; General Description," TS 26.171, version 1.01, Mar. 2001.

[3] C. Erdmann, P. Vary, K. Fischer, W. Xu, M. Marke, T. Fingscheidt, I. Varga, M. Kaindl, C. Quinquis, B. Kovesi, and D. Massaloux, "A Candidate Proposal for a 3GPP Adaptive Multi-Rate Wideband Speech Codec," in *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, May 2001.

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

[5] S. Voran, "Listener Ratings of Speech Passbands," in *Proc. of IEEE Workshop on Speech Coding*, Pocono Manor, 1997, pp. 81–82.

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

[7] Y.M. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical Recovery of Wideband Speech from Narrowband Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 544–548, Oct. 1994.

[8] M. Nilsson, S.V. Andersen, and W.B. Kleijn, "On the Mutual Information Between Frequency Bands in Speech," in *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, June 2000, vol. 3, pp. 1327–1330.

[9] P. Jax and P. Vary, "Wideband Extension of Telephone Speech Using a Hidden Markov Model," in *Proc. of IEEE Workshop on Speech Coding*, Delavan, Wisconsin, Sept. 2000, pp. 133–135.

[10] J. Epps and W.H. Holmes, "A New Technique for Wideband Enhancement of Coded Narrowband Speech," in *Proc. of IEEE Workshop on Speech Coding*, Porvoo, Finland, Sept. 1999.

[11] C. Avendano, H. Hermansky, and E.A. Wan, "Beyond Nyquist: Towards the Recovery of Broad-Bandwidth Speech from Narrow-Bandwidth Speech," in *Proc. of European Conf. on Speech Communication and Technology (EUROSPEECH)*, Madrid, Sept. 1995.

[12] K.-Y. Park and H.S. Kim, "Narrowband to Wideband Conversion of Speech using GMM-based Transformation," in *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, June 2000, vol. 3, pp. 1847–1850.

[13] P. Vary, U. Heute, and W. Hess, *Digitale Sprachsignalverarbeitung*, Teubner-Verlag, Stuttgart, 1998, (in German).

[14] Y. Linde, A. Buzo, and R.M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[15] J. Paulus, *Codierung breitbandiger Sprachsignale bei niedriger Datenrate*, Ph.D. thesis, RWTH Aachen, 1997, (in German).

[16] S.V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, Wiley, Teubner, 1996.

[17] T.K. Moon, "The Expectation-Maximization Algorithm," *IEEE Signal Processing Magazine*, pp. 47–60, Nov. 1996.

[18] J.A. Fuemmeler and R.C. Hardie, "Techniques for the Regeneration of Wideband Speech from Narrowband Speech," in *IEEE Workshop on Nonlinear Signal and Image Processing*, Baltimore, June 2001.

[19] U. Kornagel, "Spectral Widening of the Excitation Signal for Telephone-Band Speech Enhancement," in *Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Darmstadt, Sept. 2001.

[20] J. Makhoul and M. Berouti, "High-Frequency Regeneration in Speech Coding Systems," in *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1979.