WIDEBAND EXTENSION OF TELEPHONE SPEECH USING A HIDDEN MARKOV MODEL

Peter Jax and Peter Vary

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

ABSTRACT

In this paper we propose an algorithm to recover wideband speech from lowpass-bandlimited 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, and minimizes a mean squared error criterion. The new algorithm needs only one single wideband codebook and inherently guarantees the transparency of the system in the base-band. The enhanced speech exhibits a significantly larger bandwidth than the input speech without introducing objectionable artifacts.

1. INTRODUCTION

In current public telephone systems the bandwidth of the transmitted speech is limited to a frequency range of 300 Hz to 3.4 kHz. This fact leads to the characteristic thin and muffled sound of "telephone speech". However, in recent times it can be observed that the demands on the quality of speech comunication systems increase – not only a high intelligibility is desired but also high subjective quality, e.g. in hands-free telephony or for teleconferencing applications. This trend is reflected by ongoing standardizations of wideband speech codecs.

True wideband speech communication requires enhanced speech codecs and increased bitrates and therefore a modification of the transmission link. Hence, for economical reasons, the bandwidth limitation is not likely to change in the near future. Another approach towards a higher bandwidth is to extrapolate the missing low and high frequency components of the speech signal at the receiving side of the transmission link utilizing the bandlimited speech only.

This bandwidth extension of speech signals is only feasible if it is based on a model of the speech production process. Parameters of the wideband source model can be estimated from the bandlimited speech. These parameters can then be used in combination with the source model to estimate and add the missing frequencies.

In this paper only the extension of the bandwidth towards higher frequencies is treated, i.e. the input signal is assumed to contain frequencies below 3.4 kHz only[1]. By adding signal components it shall be extended to frequencies of up to 7 kHz.

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

2. ALGORITHM

According to an auto-regressive (AR) model of the process of speech production the proposed algorithm for bandwidth extension is 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 [1]. A block-diagram of the proposed algorithm is shown in Fig. 1.

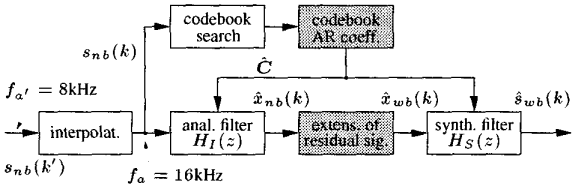

Figure 1: Block-diagram and main signal flow of the proposed algorithm for bandwidth extension.

If the input signal $s_{nb}(k')$ is sampled with a sampling frequency of $f_{a'} = 8$ kHz, the first step of the algorithm consists of a low-pass interpolation filter. The remaining parts of the algorithm process the input signal with a sample rate of $f_a = 16$ kHz. However, the signal $s_{nb}(k)$ still contains only signal components below 3.4 kHz. The further processing is done frame by frame with a frame-size of 20 ms. In the following, the frame index will be denoted by the variable m.

Using the narrowband input signal and a pre-trained codebook an estimate of the *wideband* spectral envelope of the current speech frame is calculated (see section 3). The AR filter coefficients $\hat{C}$ describing this spectral envelope are then taken from a codebook, and used in an FIR filter $H_I(z)$ to estimate the narrowband excitation signal $\hat{x}_{nb}(k)$. Since the base-band of this excitation signal can be assumed to be approximately white for unvoiced frames and to consist of harmonics with equal amplitude for voiced frames respectively, its bandwidth can be extended quite simply (see section 4). The extended excitation signal $\hat{x}_{wb}(k)$ is finally fed into the all-pole synthesis filter $H_S(z)$ thus creating the enhanced output signal $\hat{s}_{wb}(k)$.

Unlike previous algorithms for the bandwidth extension of speech signals (e.g. [1, 2, 3]), the proposed algorithm does not need a combination of several narrow- and wideband codebooks. It utilizes only one single wideband codebook. Hence, the AR coefficients used for the analysis and synthesis filters are identical

7803-6416-3/00/$10.00©2000 IEEE

and the transfer functions of these two filters are mutually inverse

$$H_S(z) = 1/H_I(z). \tag{1}$$

Due to this property the transparency of the system for the base-band signal components can be guaranteed – it is sufficient to make certain, that the base-band of the excitation signal is not modified during the extension of the residual signal.

3. SPECTRAL ENVELOPE

As described in the previous section, the extension of the spectral envelope of the speech signal is based on a wideband codebook. In this codebook several sets of AR coefficients representing the spectral envelopes of typical speech sounds are stored[2]. The training of the codebook is done with a sufficiently large training data set of *wideband* speech and the common LBG algorithm [4], minimizing the Itakuro distance measure (see [1]). 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 only a bandlimited signal is available.

The basis for the codebook search method is a hidden Markov model (HMM) of the speech generation process. Exactly one state S_i of the HMM is assigned to each entry $\boldsymbol{C}_i$ of the codebook. It is further assumed that the state of the source does only change in-between two frames of the input signal. Note, that if wideband speech is available, the – in sense of the model – *true* state sequence can be calculated.

If only the narrowband speech is known, the classification is performed by the following steps: First, a limited number of features is extracted from the narrowband speech. These extracted features are compared with a statistical model of the speech production process. The current speech frame can then either be classified into one of the trained speech sounds or the AR coefficients are estimated directly.

3.1. Features

For each signal frame an N-dimensional vector $\boldsymbol{x}(m)$ of features is extracted from the bandlimited signal. This vector consists of eight cepstral coefficients $c_1 \dots c_8$, the normalized frame energy E_n and the gradient index d_n as defined in [5]

$$d_n = \frac{\sum_{k=2}^{K} \Delta\psi(k)\,|s_{nb}(k) - s_{nb}(k-1)|}{\sqrt{\frac{1}{K}\sum_{k=1}^{K} s_{nb}^2(k)}}. \tag{2}$$

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

Whereas the cepstral coefficients carry information about the shape of the spectral envelope of the narrowband signal, the other two quantities depend on properties of the speech excitation. Additionally, the derivatives over time of all of the above ten primary features are included in the feature vector such that the dimension of this vector $\boldsymbol{x}(m)$ results in $N = 20$.

[2] The i-th set of AR coefficients will be denoted by $\boldsymbol{C}_i$ in the following. The total number of codebook entries is I.

3.2. Statistical Model

For each possible state S_i of the hidden Markov model the features $\boldsymbol{x}$ 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.

3.2.1. *Observation Probabilities* $p(\boldsymbol{x}|S_i)$

Due to the high dimension of the feature vector $\boldsymbol{x}$ these probability density functions (pdf) $p(\boldsymbol{x}|S_i)$ are modeled by *Gaussian Mixture Models* (GMMs): each pdf is approximated as the sum of L gaussian pdfs

$$p(\boldsymbol{x}|S_i) \approx \sum_{l=1}^{L} P_{il}\,\mathcal{N}(\boldsymbol{x}; \mu_{il}, \Sigma_{il}). \tag{3}$$

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

The training of the GMMs, i.e. of the quantities P_{il}, μ_{il} and Σ_{il}, can be performed with the Expectation-Maximization (EM) algorithm (e.g. [6]). The EM algorithm needs a starting point for the iterative refinement, which is determined here by clustering the training data with the LBG algorithm [4].

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.

3.2.2. *Initial State Probabilities* $\pi_i = P(S_i)$

The scalar value π_i describes the probability, that the HMM resides in state S_i without incorporating knowledge of the feature vector $\boldsymbol{x}$ or of preceding or following states.

This probability can be estimated by computing the true state sequence for the training material and evaluating the ratio between the number of occurrences of the state S_i and the total number of speech frames in the training set. The resulting probability values are stored such that the actual bandwidth extension algorithm can later access these a priori state probabilities by table lookups.

3.2.3. *Transition Probabilities* $a_{ij} = P(S_i(m+1)|S_j(m))$

The variable a_{ij} describes the probability of a transition from state S_j to state S_i in the following frame. As the initial state probabilities π_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 S_j to S_i and the total number of occurrences of state S_j.

3.3. Estimation of wideband AR coefficients

The aim of the codebook search algorithm is to calculate an estimate $\hat{\boldsymbol{C}}$ of the wideband AR coefficients which minimizes the distance to the true coefficients $\boldsymbol{C}$.

For the derivation of the estimation rule it is useful to define a helper variable $\alpha_i(m)$ as the joint probability of the partial observation sequence $\boldsymbol{X}(m) = \{\boldsymbol{x}(0), \boldsymbol{x}(1), \dots, \boldsymbol{x}(m)\}$ and the state $S_i(m)$ at frame instant m

$$\alpha_i(m) = P(S_i(m), \boldsymbol{X}(m)). \tag{4}$$

This helper variable can be expressed in a recursive manner in terms of the joint probabilities $\alpha_i(m-1)$ at frame instant $m-1$ and the observation probability $p(\boldsymbol{x}(m)|S_i(m))$ as

$$\alpha_i(m) = \left(\sum_{j=1}^{I} \alpha_j(m-1)\, a_{ij}\right) p(\boldsymbol{x}(m)|S_i(m)). \quad (5)$$

Since the preceding observation vectors are unknown for the first frame, the initial values for $\alpha_i(0)$ have to be calculated from the initial state probabilities π_i

$$\alpha_i(0) = \pi_i\, p(\boldsymbol{x}(0)|S_i). \quad (6)$$

The goal of the proposed MMSE criterion is to minimize the mean squared error between the estimated AR coefficients $\hat{\boldsymbol{C}}$ and the true coefficients $\boldsymbol{C}$ such that the following cost function is minimized

$$\mathcal{R}_{\text{MSE}}(\hat{\boldsymbol{C}}|\boldsymbol{X}) = \iint (\hat{\boldsymbol{C}} - \boldsymbol{C})^T (\hat{\boldsymbol{C}} - \boldsymbol{C})\, p(\boldsymbol{C}|\boldsymbol{X})\, d\boldsymbol{C}. \quad (7)$$

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

$$\hat{\boldsymbol{C}}_{\text{MMSE}} = \iint \boldsymbol{C}\, p(\boldsymbol{C}|\boldsymbol{X})\, d\boldsymbol{C}. \quad (8)$$

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

$$\hat{\boldsymbol{C}}_{\text{MMSE}} = \iint \boldsymbol{C} \left[\sum_{i=1}^{I} p(\boldsymbol{C}|S_i)\, P(S_i|\boldsymbol{X})\right] d\boldsymbol{C} \quad (9)$$

$$= \sum_{i=1}^{I} P(S_i|\boldsymbol{X}) \underbrace{\iint \boldsymbol{C}\, p(\boldsymbol{C}|S_i)\, d\boldsymbol{C}}_{\mathcal{E}\{\boldsymbol{C}|S_i\} = \boldsymbol{C}_i} \quad (10)$$

As shown, the integral at the right hand side of Eq. 10 yields the expected value of $\boldsymbol{C}$ given the occurence of state S_i, i.e. the corresponding codebook vector $\boldsymbol{C}_i$. Applying Bayes rule and substituting the helper variable α_i, we obtain the following estimator

$$\hat{\boldsymbol{C}}_{\text{MMSE}} = \frac{\sum_{i=1}^{I} \boldsymbol{C}_i\, \alpha_i(m)}{\sum_{i=1}^{I} \alpha_i(m)}. \quad (11)$$

The conditional probabilities $p(\boldsymbol{C}|S_i)$ can not be utilized by this estimator due to the indirect modeling of $p(\boldsymbol{C}|\boldsymbol{X})$ via the state probabilities. A superior MMSE estimator could be designed by directly modeling and exploiting $p(\boldsymbol{C}|\boldsymbol{X})$, but this is not a trivial task. Alternatively, the knowledge of $p(\boldsymbol{C}|S_i)$ can be taken into account during the training procedure of the codebook vectors $\boldsymbol{C}_i$.

4. RESIDUAL SIGNAL

Because the narrow-band excitation signal $\hat{x}_{nb}(k)$ can be assumed to be approximately white in the base-band, the wideband excitation signal is calculated as

$$\hat{x}_{wb}(k) = \begin{cases} 2\,\hat{x}_{nb}(k) & ;k = 0, \pm 2, \pm 4 \ldots \\ 0 & ;\text{else.} \end{cases} \quad (12)$$

This operation results in folding of the power spectrum. Thus, in $\hat{x}_{wb}(k)$ there is a spectral gap from 3.4 to 4.6 kHz. Additionally, the harmonic structure in high frequency regions does not match the low frequency components. However, after the synthesis filter $H_S(z)$ these effects are hardly audible.

5. 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 $I = 64$ the enhanced signal was almost indistinguishable from the original wideband speech. Even for extremly small codebooks with down to $I = 3$ entries, acceptable results can be achieved.

The training data was obtained by filtering wideband spech by a lowpass filter with a cut-off frequency of 3.4 kHz. It consisted of about 10 minutes of phonetically balanced clean speech spoken by several male and female speakers.

In many informal and comparative listening tests, the described algorithm 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 codebook search algorithm. However, the more a priori knowledge is utilized in the algorithm, the less frequent are such artifacts.

6. 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 turned out to be an appropriate tool.

7. REFERENCES

[1] H. Carl, "Untersuchung verschiedener Methoden der Sprachcodierung und eine Anwendung zur Bandbreitenvergrößerung von Schmalband-Sprachsignalen". Dissertation. Ruhr-Universität Bochum, 1994

[2] J. Epps, W. H. Holmes, "A New Technique for Wideband Enhancement of Coded Narrowband Speech". IEEE Workshop on Speech Coding, Porvoo, Finland, 1999

[3] N. Enbom, W. B. Kleijn, "Bandwidth Expansion of Speech Based on Vector Quantization of the Mel Frequency Cepstral Coefficients". IEEE Workshop on Speech Coding, Porvoo, Finland, 1999

[4] Y. Linde, A. Buzo, R. M. Gray, "An Algorithm for Vector Quantizer Design". IEEE Trans. on Communications, January 1980

[5] J. Paulus, "Codierung breitbandiger Sprachsignale bei niedriger Datenrate". Dissertation, RWTH Aachen, 1997

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