Neural Networks Versus Codebooks in an Application for Bandwidth Extension of Speech Signals

Bernd Iser, Gerhard Schmidt

Temic Speech Dialog Systems
Soeflinger Str. 100, 89077 Ulm, Germany
`bernd.iser@temic-sds.com gerhard.schmidt@temic-sds.com`

Abstract

This paper presents two versions of an algorithm for bandwidth extension of speech signals. We focus on the generation of the spectral envelope and compare the performance of two different approaches – neural networks versus codebooks – in terms of objective and subjective distortion measures.

1. Introduction

Speech signals that are transmitted over current public telephone networks have only a very limited bandwidth, e.g. 300 Hz up to 3400 Hz for analog lines. When comparing those speech signals to other audio sources such as radio or CD the quality difference is obvious and bothersome. Thus, great efforts have been made to increase the quality of telephone speech signals in recent years. Wideband codecs are able to increase the bandwidth up to 7 kHz or even higher at only moderate complexity. Nevertheless, applying these codecs would mean to exchange current networks. Another (cheaper) possibility is to extend the bandwidth after transmission over the unchanged network. The basic idea of these enhancements is to estimate the speech signal components above 3400 Hz and below 300 Hz and to complement the signal in the new frequency bands with this estimate.

The generation of this estimate can be divided into two separate tasks assuming that the well-known source-filter model of speech generation [1] is applied. First, a so-called excitation signal is required. This excitation signal corresponds to the signal that can be observed directly behind the vocal chords, which means that this signal contains information about voicing and pitch but not about formant structures or the spectral shaping in general. Consequently, this excitation signal has to be weighted with the spectral envelope of the speech signal. Thus, one key element in bandwidth extension of speech signals is the estimation of the spectral envelope. We investigate two methods for this estimation. Mapping the narrow-band envelope to a broadband envelope by either using a codebook or a neural network. This paper contains a comparative study of the performances of codebook and neural network approaches in terms of objective and subjective distortion measures.

2. Bandwidth Extension

As mentioned before the generation of speech can be divided into the generation of an excitation signal and the spectral shaping of this excitation signal. Fig. 1 shows the basic scheme of the bandwidth extensions implemented in this study. Note that the multiplication units within Fig. 1 symbolize multiplications in the frequency domain for spectral shaping and power adjustment. After generating the excitation signal $x_e(n)$ and
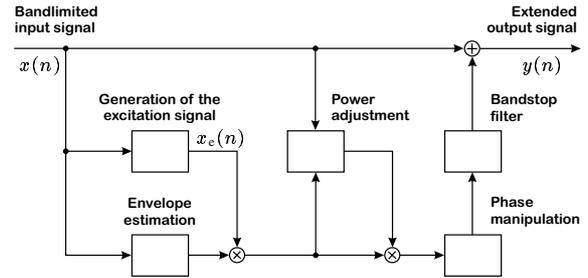

Figure 1: *Schematic diagram of the bandwidth extension.*

weighting $x_e(n)$ with the spectral envelope, power adjustment of the synthesized signal to the input signal $x(n)$ is necessary. Before adding the complementary signals the phase of the extended frequency bands can be manipulated. For computing the bandwidth extension block processing in the frequency domain is applied. The input signal is divided into overlapping blocks of length $N_B = 256$ (sampling rate = 11025 Hz). The blocks are overlapping by 75 percent resulting in a frameshift of 64 samples. In the following sections we will discuss the particular processing units of the bandwidth extension depicted in Fig. 1.

2.1. Generation of the Excitation Signal

For the generation of the excitation signal $x_e(n)$ several methods have been proposed. Most of them can be divided into two classes: nonlinearities and frequency shifting approaches. In the latter case a part within the telephone band (e.g. 500 ... 800 Hz) is copied into the extension frequency bands (e.g. 0 ... 300 Hz). If a fixed scheme is applied the pitch structure of voiced sequences might be destroyed. Thus, the frequency range is adjusted individually in each processing frame according to an estimate of the pitch frequency. Furthermore, the spectral envelope of the speech signal has to be removed by means of predictor error filtering. The performance of these approaches depends crucially on the quality of the pitch estimation. Especially in the case of moderate or even high background noise level these methods tend to produce artifacts.

In the second class nonlinearities, such as quadratic or cubic characteristics, are applied directly to the bandlimited input signal. In this case the extended excitation signal is generated in the time domain and no pitch estimation is required. The drawback of these methods is that the power of the resulting signal does not correspond with the power of the original signal in a simple way. For this reason, sophisticated power adjustment techniques are required. In this comparison a cubic characteristic is used to generate the excitation signal. Fig. 2 shows the result of this method for a voiced sound. The application of a cubic characteristic in the time domain corresponds to the dou-
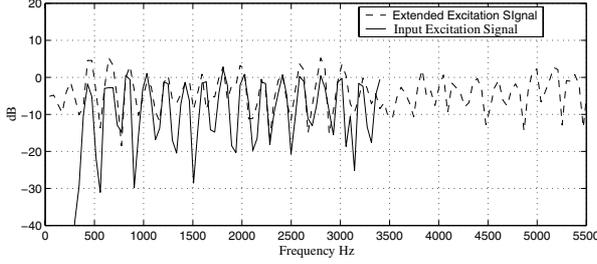
Figure 2: *Example for the generation of the excitation signal.*

ble convolution of the signal within the frequency domain:

$$\tilde{x}_e(n) = x^3(n) \; \circ\!\!-\!\!\bullet \; X(e^{j\Omega}) * X(e^{j\Omega}) * X(e^{j\Omega}). \qquad (1)$$

In the case of voiced sounds the spectrum of the excitation signal is similar to a line spectrum showing the harmonics of the pitch frequency within the telephone band. Convolving this signal with itself in the frequency domain complements the missing harmonics and the pitch frequency itself. Note that the application of a cubic characteristic also generates spectral components around $\Omega = 0$ that have to be removed. A quadratic characteristic would serve as well but the cubic characteristic results in sharper harmonics. In the case of unvoiced sounds the excitation signal is similar to white noise and the application of a cubic characteristic extends these signals too without changing the basic noise properties except that the signal is no longer a white signal. After the cubic characteristic has been applied to the input signal the resulting envelope is the result of the double convolution in the frequency domain of the input envelope. Thus, before weighting the extended excitation signal with the extended spectral envelope in the frequency domain a spectral flattening has to be done. For this reason a predictor error filter of order $p = 12$ is applied:

$$x_e(n) = \tilde{x}_e(n) - \sum_{i=1}^{p} \tilde{a}_i(n)\, \tilde{x}_e(n-i). \qquad (2)$$

The filter coefficients $\tilde{a}_i(n)$ are adapted individually for each frame to the correlation properties of the signal $\tilde{x}_e(n)$. Hence, the coefficients $\tilde{a}_i(n)$ have both, a time index n and a coefficients index i. Fig. 3 shows the schematic diagram of the generation of the excitation signal.
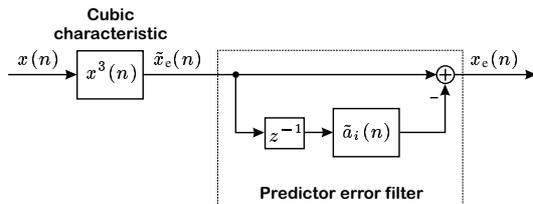

Figure 3: *Scheme of the excitation-signal generation.*

2.2. Generation of the Spectral Envelope

After the generation of the excitation signal $x_e(n)$ the (wideband) spectral envelope needs to be estimated. Again, two different classes of estimation techniques have been suggested. Originating in the field of speech coding codebook approaches have been proposed. In codebook based methods the spectral envelope in terms of low order all-pole models of the current (bandlimited) frame is first mapped on one of e.g. $N_{CB} = 1024$ codebook entries according to a predefined distance measure [2]. The codebook is usually trained with a large corpus

of pairs of bandlimited and wideband speech sequences. Each entry of the codebook for bandlimited envelopes has its wideband counterpart. This counterpart is utilized as an estimate for the wideband spectral envelope of the current frame. The second class of estimation techniques is based on the application of neural networks. The training of the neural network is performed – like the codebook generation – with pairs of parameter sets corresponding to bandlimited and wideband speech sequences. Typical parameter sets are predictor coefficients, cepstral coefficients, or line spectral frequencies. In order to compare both approaches one scheme of each class has been implemented. The schematic diagram of the envelope generation is depicted in Fig. 4. Before extracting the coefficients of an all-
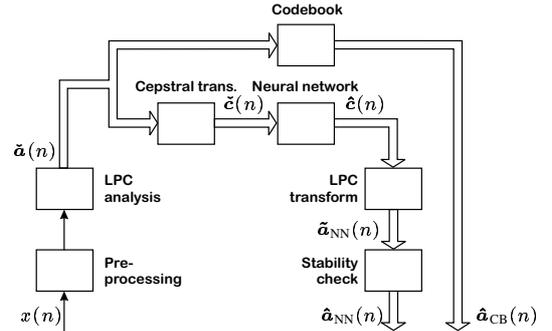

Figure 4: *Schematic diagram of the envelope generation.*

pole LPC model some preprocessing is performed. Within the preprocessing unit the sampling frequency is reduced in order to remove frequency parts above 3400 Hz. Due to this subsampling the LPC analysis can focus on the representation of the telephone band and is not optimized for reproducing the edge around 3400 Hz. For the same reason the frequencies below 300 Hz are set to a mean value calculated out of a few samples above 300 Hz. As a matter of course, this type of preprocessing has also been used for the generation of the codebook and during the training of the neural network, respectively. Both approaches generate the coefficients $\hat{a}_i(n)$ of a predictor error filter of order $P = 20$. The spectral envelope is estimated for each frame according to

$$\frac{1}{\hat{A}\left(e^{j\Omega}, n\right)} = \frac{1}{1 - \sum_{i=1}^{P} \hat{a}_i(n)\, e^{-j\Omega i}}. \qquad (3)$$

In Fig. 5 four time-frequency analyses are depicted. On the left side spectrograms of the original and the bandlimited signal are depicted. In the right part of the figure spectrograms of the extended signals are presented.

2.2.1. Neural Network Approach

The neural network used in this paper is a multilayer perceptron (MLP) in feed forward operation with three layers. The number of neurons in each layer is listed in Tab. 1. For the

Table 1: *Distribution of neurons per layer.*

Layer	Number of neurons
input	36
hidden	36
output	30

extraction of the narrowband envelope an LPC analysis of order $p = 12$ has been used (after preprocessing as described in the last section). The neural network has been trained with the *standard-backpropagation algorithm*. The (quadratic) difference between desired and produced predictor coefficients is
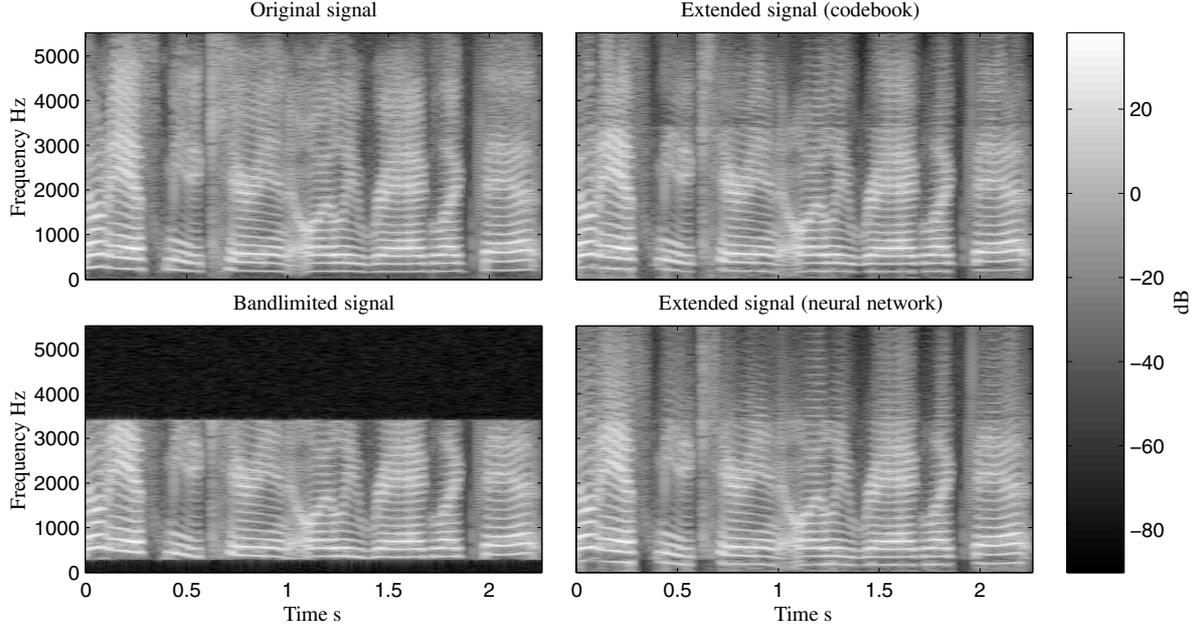
Figure 5: *Spectrograms of the original and the bandlimited speech sequence as well as analyses of two extended sequences.*

not a well-suited cost function for network training. Thus, the p predictor coefficients have been transformed recursively into $\frac{3}{2}p$ cepstral coefficients according to

$$\check{c}_i(n) = \begin{cases} \check{a}_i(n) + \frac{1}{i} \sum_{k=1}^{i-1} k\, \check{c}_k(n)\, \check{a}_{i-k}(n), & \text{for } i = 1 \ldots p, \\ \frac{1}{i} \cdot \sum_{k=1}^{i-1} k \cdot \check{c}_k(n)\, \check{a}_{i-k}(n), & \text{else.} \end{cases} \tag{4}$$

Using cepstral coefficients instead of LPC coefficients results in the well-known cepstral distance measure:

$$d_{\text{CEPS}}^2\left(\ldots, n\right) = \sum_{i=1}^{\frac{3}{2}P} \left(c_i(n) - \hat{c}_i(n)\right)^2. \tag{5}$$

Paying respect to the fact that speech can be modeled as a quasi stationary process we used the cepstral coefficients of the current block as well as of the predecessor to implemented some kind of memory. For this reason we get 36 input neurons out of $2 \cdot 12$ LPC coefficients. For the broadband coefficients we used an LPC order of $P = 20$ resulting in 30 neurons for the output layer. The cepstral coefficients $\hat{c}_i(n)$ are the output values of the network and the coefficients $c_i(n)$ are cepstral coefficients computed out of the wideband signals during the training. After the inverse transformation (cepstral coefficients into LPC coefficients) the resulting all-pole model is examined if there are poles outside the unit circle. If so, these poles are reflected inside the unit circle for stability issues. For the training of the neural network as well as for the codebook an extract of the TIMIT speech data base consisting of 100 female and 100 male speakers with 10 utterances per speaker has been used.

2.2.2. Codebook Approach

The codebook used in this paper has been trained with the *LBG algorithm* [3]. As a distance measure the likelihood ratio distance measure (LR) has been applied to the spectral envelopes of the broadband and the bandlimited signals [1]:

$$d_{\text{LR}}\left(\ldots, n\right) = \int_{-\pi}^{\pi} \frac{\left|\hat{A}(e^{j\Omega}, n)\right|^2}{\left|A(e^{j\Omega}, n)\right|^2} \frac{d\Omega}{2\pi} - 1. \tag{6}$$

If all-pole models according to Eqn. 3 are utilized to describe the spectral envelopes Eqn. 6 can be written in matrix vector notation:

$$d_{\text{LR}}\left(\ldots, n\right) = \frac{\hat{\boldsymbol{a}}^T(n)\, \boldsymbol{R}_{xx}(n)\, \hat{\boldsymbol{a}}(n)}{\boldsymbol{a}^T(n)\, \boldsymbol{R}_{xx}(n)\, \boldsymbol{a}(n)} - 1. \tag{7}$$

This distance measure is consistent with the LPC analysis. The vectors $\hat{\boldsymbol{a}}(n)$ and $\boldsymbol{a}(n)$ contain the predictor coefficients of the estimated envelope and the original broadband envelope, respectively. The matrix $\boldsymbol{R}_{xx}(n)$ is the autocorrelation matrix of the current frame. The resulting codebook consists of 1024 narrowband envelopes represented by LPC vectors of length $p = 12$ and of 1024 broadband envelopes represented by LPC vectors of length $P = 20$.

2.3. Power Adjustment

Before adding the synthesized signal in the complementary frequency bands to the input signal the power of the synthesized signal has to be adjusted. Hence, the energy of the synthesized signal within the telephone band is compared to the energy of the input signal assuming that the estimated spectral envelope is almost identical to the spectral envelope of the input signal in this band. Before the power adjustment is applied the power ratio is smoothed by a first order IIR filter in consideration of the fact that the generated spectral envelope might jitter from one block to another.

2.4. Phase Manipulation

The phase of the synthesized signal in the extension band is determined by the application of the cubic characteristic. The spectral flattening that is done after the application of the cubic characteristic as well as the weighting with the spectral envelope is only applied to the magnitude. To avoid bothersome artifacts it is necessary to assign a reasonable phase to the synthesized signal. This is done by extracting the phase of the input signal in a predefined part of the telephone band and assigning the same phase to the synthesized signal in the extension band.

3. Quality Assessment

For the evaluation of the quality of the different bandwidth extensions we have investigated one subjective and several objective distance measures.

3.1. Objective Distance Measures

Some well known objective distance measures are L_p norm based spectral distortion measures [4]:

$$d_p^p = \int_{-\pi}^{\pi} \left| \log \frac{1}{\left| A(e^{j\Omega}, n) \right|^2} - \log \frac{1}{\left| \hat{A}(e^{j\Omega}, n) \right|^2} \right|^p \frac{d\Omega}{2\pi}, \quad (8)$$

where the most common choices for p are $1, 2, \infty$ resulting in the *city block distance*, *Euclidean distance* (SD) and the *Minkowski distance*. Another well known distance measure is the log area ratio distance measure (LAR):

$$d_{\text{LAR}}(..., n) = \sum_{i=1}^{P} \left[\ln \frac{\left(1 + k_i(n)\right)\left(1 - \hat{k}_i(n)\right)}{\left(1 - k_i(n)\right)\left(1 + \hat{k}_i(n)\right)} \right]^2, \quad (9)$$

where $\hat{k}_i(n)$ and $k_i(n)$ denote the reflection coefficients of the extended and the broadband signal, respectively. Both can easily be derived from LPC coefficients [1].

3.2. Subjective Distance Measures

In order to evaluate the subjective quality of the extended signals a mean-opinion-score (MOS) test in terms of a comparison rating has been executed. About 30 people of different age and gender have participated in the test. The subjects were asked to compare the quality of two signals (pairs of bandlimited and extended signals) by choosing one of the statements listed in Tab. 2. The extended signals have not been used for network

Table 2: *Conditions of the MOS test.*

Score	Statement
-3	A is much worse than B
-2	A is worse than B
-1	A is slightly worse than B
0	A and B are about the same
1	A is slightly better than B
2	A is better than B
3	A is much better than B

training or codebook generation, respectively. Finally, the subjects were asked whether they prefer the signal which was extended by the neural network or the one which was extended with the codebook.

4. Discussion

Before we present the results of the evaluation some general comments are presented. An interesting fact is that both the codebook as well as the neural network are not representing the higher frequencies appropriately where the behavior of the neural network is even worse. At least the power of the higher frequencies produced by the neural network and the codebook is mostly less than the power of the original signal so that bothersome artifacts do not attract attention to a certain degree.

For evaluating the quality of the envelope generation, the four distance measures presented in Eqs. 6-9 have been applied to a validation set of speech sequences consisting of 400 sentences. The averaged results are presented in Tab. 3. Note that the distance is measured between the extended signal and the broadband signal. Even if the neural network produces better envelopes in terms of distortion measures, the MOS test shows

Table 3: *Measured distortion.*

	SD	CEPS	LAR	LR
Neural Netw.	0.527	0.851	4.62	2.17
Codebook	0.609	0.990	4.58	4.49

a different result. The subjects rated the signals of the network approach with an average mark of 0.53 (between equal and slightly better than the bandlimited signals) and the signals resulting from the codebook scheme with 1.51 (between slightly better and better than the bandlimited signals). When choosing which approach produces better results around 80 percent voted for the codebook based scheme. Fig. 6 shows the results of the MOS test.
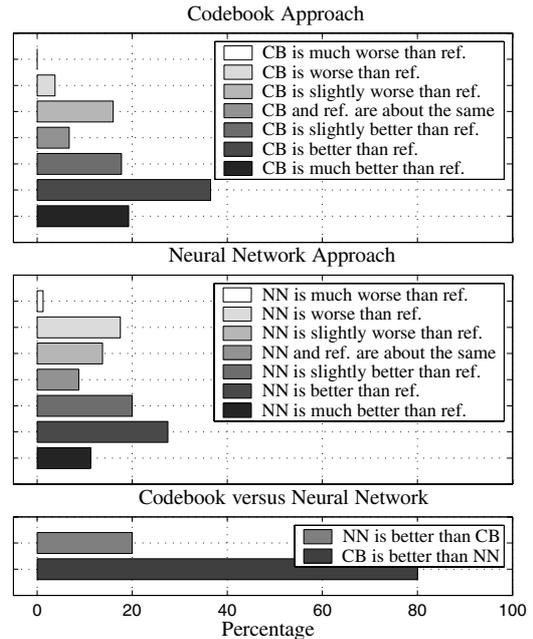

Figure 6: *Results of the MOS test.*

5. Conclusions

In this contribution two schemes for bandwidth extension of speech signals have been compared. A variety of distance measures have been investigated in order to validate the approach. Unfortunately only one of the distance measures was able to match the subjective quality obtained from a MOS test. Even if the codebook approach seems to be the better choice, neural networks are also able to enhance the speech quality, but at much lower computational complexity.

6. References

[1] Deller Jr., J. R., Hansen, J. H. L. and Proakis, J. G. "Discrete-Time Processing of Speech Signals", IEEE Press, 2000.

[2] Kornagel, U. "Spectral Widening of Telephone Speech Using an Extended Classification Approach, Proc. EUSIPCO 2002, vol. 2, pp. 339 - 342, 2002.

[3] Linde, Y., Buzo, A. and Gray, R. M. "An Algorithm for Vector Quantizer Design", IEEE Trans. Comm., vol. COM-28, no. 1, pp. 84 - 95, Jan. 1980.

[4] Gray, R. M. e. a. "Distortion Measures for Speech Processing", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28, no. 4, pp. 367 - 376, Aug. 1980.