SPECTRAL WIDENING OF TELEPHONE SPEECH USING AN EXTENDED CLASSIFICATION APPROACH

Ulrich Kornagel
Signal Theory, Institute for Communication Technology
Darmstadt University of Technology
Merckstrasse 25
D-64283 Darmstadt, Germany
e-mail: ulrich.kornagel@nt.tu-darmstadt.de

ABSTRACT

The quality degradation of transmitted speech in current telephone systems due to the band limitation can be reduced by application of an enhancement system which generates "pseudo" wide-band speech from its telephone-band version. In this contribution, an extended classification approach is presented as a component of such an enhancement system. The incoming telephone-band signal is divided into the spectral envelope and the residual signal by means of an LPC-error-filter. The proposed method focuses on the enhancement of the spectral envelope. At this, an appropriate set of spectral wide-band envelopes is selected from a pre-trained codebook in two steps. In the first step a subset of proper wide-band envelopes is selected. This is then further reduced in the second step by means of a cost-function which takes account of the relationship between the wide-band envelopes of the subset and the resulting wide-band envelope of the classification procedure one time step earlier. Listening tests have shown that the extended classification is a proper approach to enhance the quality of the synthesized signal components.

1 INTRODUCTION

In current telephone systems, the transmitted speech is usually band-limited to the range from 0.3 kHz to 3.4 kHz resulting in the characteristic quality degradation of telephone speech. The aim of the spectral widening idea is to define an enlarged frequency band, usually the whole band up to 8 kHz, and to generate the frequency components missing in this sense.

For this purpose, the incoming telephone-band signal is separated into consecutive overlapping blocks. For each block, the signal is divided into the *spectral envelope* and the *residual signal* using an LPC-error-filter. Each part has to be enhanced separately (see section 2 and 3, respectively). Finally, the desired speech signal with an enlarged bandwidth is generated by a shaping filter defined by the enlarged spectral envelope and driven by the enhanced residual signal. That is, the enhanced residual signal is used as an excitation signal (Figure 1).

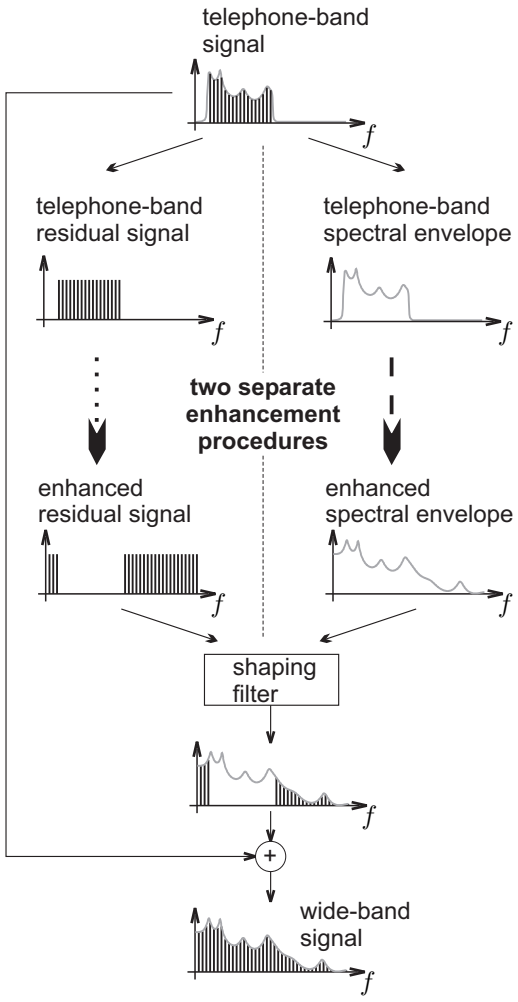

Figure 1: Principle of the spectral enhancement procedure, illustrated in the frequency domain.

The enhancement procedure of the spectral envelope contains several components that are also used in speech recognition applications [5]. Especially the feature extraction step is very similar. One important difference to the spectral widening application is that the latter

has no restrictions to the speech contents, i.e. it is not allowed to confine the algorithm to a finite number of words, sentences, etc.

However, it is possible to exploit the knowledge that the variability of speech characteristics over time is restricted. This knowledge is utilized in an extended classification procedure proposed below. With this, it is possible to *restrict the dynamics of the synthesized speech in the classification process.*

2 ENHANCEMENT OF THE SPECTRAL ENVELOPE

The basis of determining the spectral wide-band envelope is a wide-band codebook containing these envelopes as parameters of autoregressive models, i.e. as vectors of LPC-coefficients. By means of a classification procedure, the most suitable spectral wide-band envelope is determined.

2.1 Feature Extraction

The first step of the classification process is to obtain as much information from the telephone-band signal as possible. The resulting set of features should at least contain information concerning the spectral envelope of the telephone-band signal. Usual representations for this are the vector of cepstral coefficients or the LPC-coefficients. Both representations can uniquely be transformed into one another.

Further features can be the normalized short-time energy [3] and the degree of voicing [2].

2.2 Extended Classification

The extended classification procedure can be divided into three sub-procedures, namely the classification 1, the classification 2 and the linear combination (Figure 2).

Let $\mathbb{U} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be the set of N LPC-coefficient vectors defining the N spectral wide-band envelopes of the wide-band codebook and $\mathbb{P} = \{p_1, p_2, \ldots, p_N\}$ be an associated set of classification measures. The latter indicates the cost of classification 1 of each vector in $\mathbb{U}$ for the current block.

2.2.1 Classification 1

The classification 1 reduces the set $\mathbb{U}$ to a subset $\tilde{\mathbb{U}} = \{\mathbf{x}_{\xi(1)}, \mathbf{x}_{\xi(2)}, \ldots, \mathbf{x}_{\xi(M)}\}$ containing M vectors, i.e. $\tilde{\mathbb{U}} \subset \mathbb{U}$ and $1 < M < N$. This is done by selecting the M vectors $\mathbf{x}_{\xi(j)}$ with minimal costs $p_{\xi(j)}$ for $1 \leq j \leq M$. The function $i = \xi(j)$ maps the vector indices j of the set $\tilde{\mathbb{U}}$ to the vector indices i of the set $\mathbb{U}$ with $1 \leq \xi(j) \leq N$ for $1 \leq j \leq M$ (Figure 3).

One possible basic realization of this classification step is to generate a second codebook with spectral telephone-band envelopes (e.g.[1]), the so called

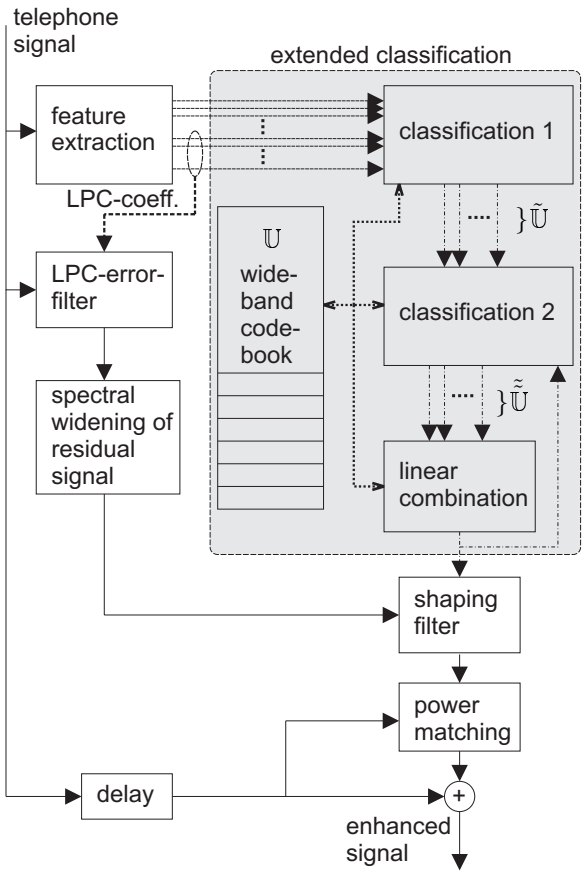

Figure 2: Overall view of the entire system.

telephone-band codebook. Both codebooks are connected, i.e. each vector of the telephone-band codebook belongs to exactly one vector of the wide-band codebook with the same spectral characteristic within the telephone-band. The cost of classification of one vector is then defined by a spectral distortion measure between the spectral envelope of the current block (of the incoming telephone-band signal) and the respective spectral envelope from the telephone-band codebook. Because of the connection of both codebooks, the M most suitable spectral wide-band envelopes are known when the M telephone-band envelopes with lowest costs are chosen from the telephone-band codebook.

2.2.2 Classification 2

The classification 2 reduces the subset $\tilde{\mathbb{U}}$ to the subset $\tilde{\tilde{\mathbb{U}}}$ containing L vectors, i.e. $\tilde{\tilde{\mathbb{U}}} \subset \tilde{\mathbb{U}}$ and $1 < L < M$. This second classification procedure is based on another cost-function which takes the dynamics of the later synthesized speech signal into consideration. At this, each vector of the set $\tilde{\mathbb{U}}$ is defined to be the cheaper the more its spectral envelope resembles the spectral wide-band envelope determining the shaping filter of the antecedent block. This reduces unrealistic variances of the wide-band envelopes for consecutive blocks due to uncertain-

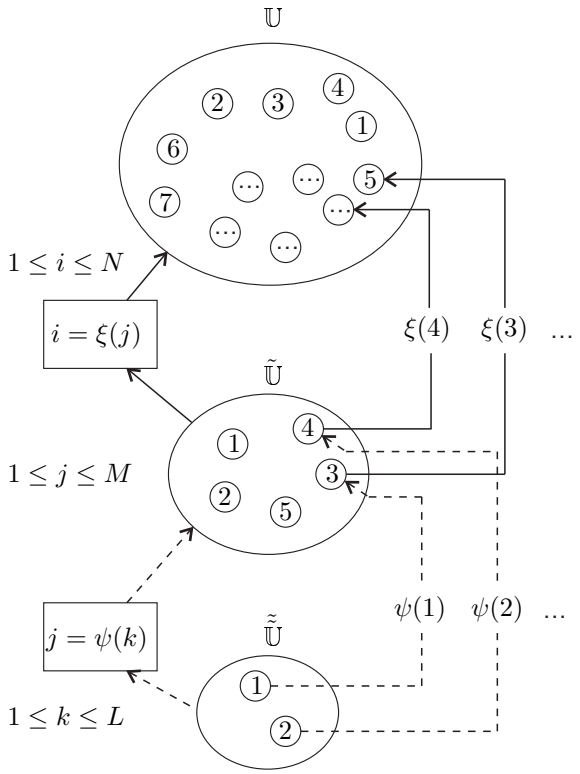

Figure 3: Mapping of indices. The figure shows the three sets which occur in the extended classification. For each set the vectors are symbolised by circles containing the respective vector indices. The two mapping functions $i = \xi(j)$ and $j = \psi(k)$ connect the vector indices of the three sets.

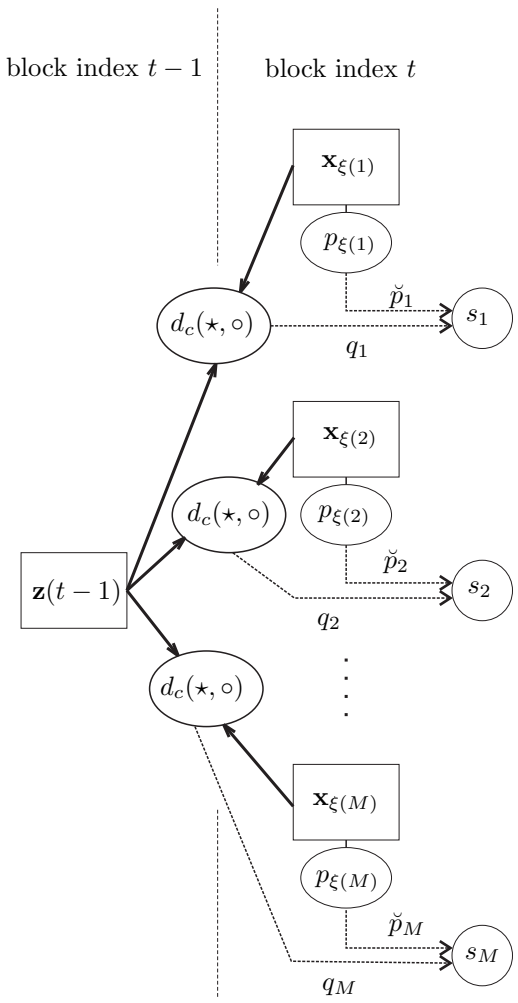

Figure 4: Overall view of the cost-function in the classification 2. The abbreviated arguments "$\star$" and "$\circ$" of the cepstral distance $d_c(\star, \circ)$ have to be substituted by $\mathbf{z}(t-1)$ and $\mathbf{x}_{\xi(j)}$ (for $1 \le j \le M$), respectively.

ties in the first classification step.

Let t be the index of the current block and $\mathbf{z}(t-1)$ be the vector of LPC-coefficients describing the shaping filter (i.e. the spectral wide-band envelope) of the last block, i.e. at block index $t-1$. Furthermore, let $\breve{p}_j$ be the normalized costs of the first classification step

$$\breve{p}_j = \frac{p_{\xi(j)}}{\sum_{\nu=1}^{M} p_{\xi(\nu)}} \quad ; \quad 1 \le j \le M$$

and q_j be the normalized cepstral distance d_c [5] between the vector $\mathbf{z}(t-1)$ and the vector $\mathbf{x}_{\xi(j)}$

$$q_j = \frac{d_c(\mathbf{x}_{\xi(j)}, \mathbf{z}(t-1))}{\sum_{\nu=1}^{M} d_c(\mathbf{x}_{\xi(\nu)}, \mathbf{z}(t-1))} \quad ; \quad 1 \le j \le M$$

again interpreted as costs.

The entire cost-function is then defined for each vector in the set $\tilde{\mathbb{U}}$ as

$$s_j = \breve{p}_j + q_j \quad ; \quad 1 \le j \le M.$$

All costs are collected in the set $\mathbb{S} = \{s_1, s_2, \ldots, s_M\}$. An overall view is given in Figure 4. The classification 2 selects the L vectors with minimal costs $s_{\psi(k)}$ to define the subset $\tilde{\tilde{\mathbb{U}}}$. At this, the connection of the indices k of the set $\tilde{\tilde{\mathbb{U}}}$ with the indices j of the sets $\tilde{\mathbb{U}}$ and $\mathbb{S}$ is defined by the function $j = \psi(k)$ with $1 \le \psi(k) \le M$ for $1 \le k \le L$ (Figure 3).

2.2.3 Linear Combination

The vector $\mathbf{z}(t)$ of LPC-coefficients, which defines the shaping filter for the current block, is computed as a weighted sum of the L vectors in the subset $\tilde{\tilde{\mathbb{U}}}$. For this purpose, it is necessary to convert the vectors in $\tilde{\tilde{\mathbb{U}}}$ into a representation which allows to ensure the stability of the resulting shaping filter. One possible representation is the vector $\mathbf{\Gamma}$ of reflection coefficients. It can be easily calculated from the LPC-coefficients by means of the step-down algorithm [6]. Thus, L vectors $\mathbf{\Gamma}_k$ for $1 \le k \le L$ are computed from the L vectors of $\tilde{\tilde{\mathbb{U}}}$.

The weighting factor for each vector $\mathbf{\Gamma}_k$ is calculated as the inverse cost $1/s_{\psi(k)}$ normalized to the sum $\sum_{\nu=1}^{L} 1/s_{\psi(\nu)}$. Therefore, the weighted sum $\mathbf{\Gamma}_{\sum}$ of the vectors $\mathbf{\Gamma}_k$ is obtained as

$$\mathbf{\Gamma}_{\sum} = \frac{1}{\sum_{\nu=1}^{L} 1/s_{\psi(\nu)}} \sum_{\mu=1}^{L} 1/s_{\psi(\mu)} \mathbf{\Gamma}_{\mu}.$$

Stability of the shaping filter is guaranteed for all components of $\mathbf{\Gamma}_{\sum}$ to have an absolute value bounded by unity. This is always true because each single vector in $\mathbb{U}$ describes a stable filter and the linear combination does not affect this property since the sum of all weighting factors equals one.

Finally, the vector $\mathbf{\Gamma}_{\sum}$ is converted back to the vector $\mathbf{z}(t)$ of LPC-coefficients by means of the step-up procedure [6].

3 ENHANCEMENT OF THE RESIDUAL SIGNAL

The excitation signal for the shaping filter ($\mathbf{z}(t)$) can be computed from the telephone-band residual signal which is calculated by means of an LPC-error-filter. The necessary LPC-coefficients are already known from the feature extraction step. The enhancement procedure can be divided into the enhancement towards frequencies below 0.3 kHz and towards frequencies above 3.4 kHz.

The lowpass components can be generated by application of a quadratic function followed by a lowpass filter with a cutoff frequency $f_c = 0.3$ kHz. The highpass components can be obtained by a modulation approach [4]. The resulting excitation signal has spectral components within the 8 kHz band except for the telephone-band since the shaping filter only has to generate spectral components complementary to the telephone band.

4 SYNTHESIS

The spectral components missing in the sense of an 8 kHz bandwidth are generated by exciting the shaping filter with the enhanced residual signal (see section 3). At this, the shaping filter is defined as an autoregressive model which is determined by the LPC-coefficients $\mathbf{z}(t)$ from the extended classification for block index t (see section 2).

In the synthesis step, the original telephone-band signal and the synthesized signal are added. At this, both signals have to be synchronized by means of a delay in the telephone-band signal path. Furthermore, the power of the synthesized signal has to be matched to the power of the telephone-band signal.

5 DISCUSSION

The splitting of the classification procedure and the application of the wide-band envelope feedback in classification 2 improves the quality of the synthesized speech in comparison with a one-step-classification only based on telephone-band features.

Furthermore, the results are superior to the ones produced by directly applying the linear combination step to the L cheapest vectors of classification 1, i.e. by discarding the classification 2 procedure.

The benefit of the proposed approach is that inhomogeneities and chippering in the synthesized signal components can be reduced significantly.

To ensure a reliable classification process the values of M and L should not be too large in comparison with the value of N. As an example, the values $N = 512$, $M = 5$ and $L = 2$ showed reasonable results.

References

[1] Carl, H. and Heute, U.: *Bandwidth enhancement of narrow-band speech signals* , Signal Processing VII, Theories and applications, pp. 165-168, 1994.

[2] Epps, J. and Holmes, W.H.: *A new technique for wideband enhancement of coded narrowband speech*, IEEE Workshop on Speech Coding, Proceedings, pp. 174-176, 1999.

[3] Jax, P. and Vary, P.: *Wideband extension of telephone speech using a hidden markov model*, IEEE Workshop on Speech Coding, Proceedings, pp. 133-135, 2000.

[4] Kornagel, U. : *Spectral widening of the excitation signal for telephone-band speech enhancement*, International Workshop on Acoustic Echo and Noise Control, Darmstadt, Germany, Conference Proceedings, pp. 215 - 218, 2001.

[5] Rabiner, L. and Juang, B.-H. : *Fundamentals of speech recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[6] Hayes, M.H.: *Statistical digital signal processing and modeling*, John Wiley & Sons, New York, 1996.