SPEECH ENHANCEMENT VIA FREQUENCY BANDWIDTH EXTENSION USING LINE SPECTRAL FREQUENCIES

S. Chennoukh, A. Gerrits, G. Miet and R. Sluijter*

Philips Research Labs., Eindhoven, Netherlands
(*) Philips Consumer Communications, Le Mans, France
email: samir.chennoukh@philips.com

ABSTRACT

This paper contributes to narrowband speech enhancement by means of frequency bandwidth extension. A new algorithm is proposed for generating synthetic frequency components in the highband (i.e., 4-8kHz) given the lowband ones (i.e., 0-4kHz) for wideband speech synthesis. It is based on linear prediction (LPC) analysis-synthesis. It consists of a spectral envelope extension using efficiently line spectral frequencies (LSF) and a bandwidth extension of the LPC analysis residual using a spectral folding. The lowband LSF of the synthesis signal are obtained from the input speech signal and the highband LSF are estimated from the lowband ones using statistical models. This estimation is achieved by means of four models that are distinguished by means of the first two reflection coefficients obtained from the input signal linear prediction analysis.

1. INTRODUCTION

In most speech transmission systems, the bandwidth is limited to a range of 0.3-3.4kHz. This speech bandwidth represents a good compromise between speech quality and transmission bandwidth for voiced sounds in general and often a poor one for unvoiced sounds. This setting leads to muffling characteristics in the telephone speech. The human preference for the wideband speech has been proven in ITU evaluation [1]. Indeed, wideband speech, whose bandwidth is defined in the range 0.05-7kHz, spans all distinctive speech frequency components. Therefore, the wideband speech sounds clear and can provide a more natural conversation over the telephone lines. Yasukawa [2], as many other investigators [3], proposed solutions to generate the missing frequency bands from the received narrowband speech without any extra-information being transmitted. Then, to reconstruct the wideband speech, these bands are added to the received narrowband speech (Figure 1).

Generating synthetic components in the highband and the lowband given the telephone band without any extra information assumes that the frequency components in these bands are correlated to the telephone bandwidth components. When using low- and high- pass filters to add these missing frequency bands to the narrowband signal, the energy of these bands need to be scaled correctly, otherwise the quality of the reconstructed wideband signal is degraded with significantly perceivable distortions. In order to avoid such unconvenient processing, one would like to achieve the synthesis such as to keep the amplitude spectral properties in the range 300-3400Hz identical to the input narrowband signal. Therefore, one needs an efficient spectral envelope extension technique. In this paper, a new algorithm is proposed to achieve such a synthesis using a linear prediction analysis-synthesis. In this scheme, the wideband signal is reconstructed, unlike in the previous studies [2], without using a high-pass filter (HPF) for the highband extraction and a band-pass filter (BPF) for the narrowband extraction (Figure 1). It is a time domain processing and its description is given in section 2. The line spectral frequencies (LSF) are proposed for the spectral envelope extension. The extension is achieved according to the spectral envelope shape characterised by the first two reflection coefficients. These coefficients showed interesting variation as a function of the nature of the speech signal. In this study, the wideband speech synthesis from narrowband speech is limited to generate only a highband (i.e., 4-8kHz) to an original 4kHz full-bandwidth speech.

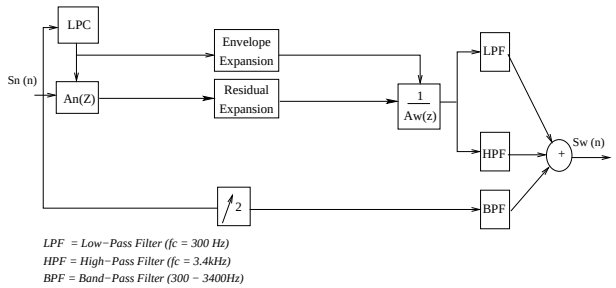

Figure 1: *Wideband synthesis based on linear prediction analysis-synthesis*

2. BLOCK DIAGRAM FOR WIDEBAND SYNTHESIS

Figure 2 shows the proposed block diagram to achieve the synthesis. The input speech signal, which is sampled at 8kHz, is first up-sampled by two (i.e., insert zeros between every successive samples). The obtained signal is now sampled at 16kHz. It has the same spectrum in the lowband, i.e. 0-4kHz, as the input signal and a folded version of it in the highband, i.e. 4-8kHz. This signal is then low-pass filtered (LPF) to remove the folded version in order to recover the same spectral properties as the input signal but sampled at 16kHz.

In the lower branch, the signal is first down-sampled by 2. Then, the down-sampled signal is modelled using an

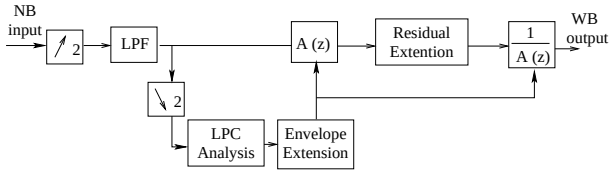

Figure 2: *Bandwidth extension to 8kHz from a 4kHz speech bandwidth*

auto-regressive LPC model. The model transfer function represents the spectrum of the input speech. An extension of the spectral envelope is achieved using the LPC coefficients. Then, the output signal of the low-pass filter is analysed using the extended LPC analysis filter. The analysis residual, that is expected to have a flat spectrum, is successively down-sampled and up-sampled by 2 (i.e., put to zero every other sample) which realises a spectral folding [4]. The obtained sparse signal is used as an excitation to the wideband synthesis filter (i.e., extended LPC model.)

2.1. Spectral envelope extension using line spectral frequencies

The line spectral frequencies (LSF) were first published by Itakura [5]. These frequencies are obtained from the roots of two transfer functions, where one is the difference and the other is the sum between the LPC analysis filter transfer function and its conjugate [6].

In this section, LSF are used to achieve the spectral envelope extension. They are estimated from the input signal. The obtained LSF are located in the range 0-π in a 4kHz bandwidth of a speech signal sampled at 8kHz. If we model the corresponding wideband speech (i.e., 8kHz bandwidth) using an LPC model with twice the order of the input signal LPC model, the input signal LSF should represent the wideband LSF in the lowband range 0-$\pi/2$. Thus, the lowband LSF of the wideband speech are given as the input signal LSF divided by 2.

A simulation of the wideband speech synthesis is achieved where the lowband LSF are obtained from input speech signal as mentioned above and the highband LSF are taken from the corresponding wideband speech. For the linear prediction analysis residual, a spectral folding is applied. The simulation showed a very good wideband synthesis quality. Thus, a spectral envelope extension is developed to estimate the highband LSF from the lowband LSF.

2.1.1. Lowband-highband mapping using a single matrix

The highband LSF of the wideband speech synthesis are obtained using a linear mapping function [7] given as:

$$f_h = f_l.A$$

where f_h is the vector of the highband LSF, the f_l is the vector of the lowband LSF obtained from input speech and A is the transform matrix of dimension pxp where p is the linear prediction model order of the input speech. This matrix is obtained by training using a long list of examples of vector pairs, where on one side are the lowband LSF and on other side are the corresponding highband LSF. This list is generated from a speech signal database. The matrix A is obtained as follows:

$$A = (F_l^T F_l)^{-1} F_l^T F_h$$

where F_h is a matrix of the LSF training data in which the rows represent the highband LSF of each example and F_l is a matrix of the LSF training data in which the rows represent the lowband LSF of each example.

The use of one matrix for the mapping between the lowband LSF and the highband LSF has given significant distortions. For this reason, a study has been performed to use a higher number of matrices to achieve the mapping. This experience has shown that the relationship between the lowband and the highband differs as a function of the spectral shape.

2.1.2. Lowband-highband mapping using 4 matrices

To use multiple matrices for more accurate prediction of the highband spectral envelope, one needs to find relevant parameters that can be used to cluster the spectral shapes. One way of doing it is to look at the reflection coefficients [8]. It is known that the first reflection coefficient provides the overall tilt of the amplitude spectrum. Then, the problem is to establish the same kind of interpretation on the shape of the amplitude spectrum when used jointly with the second reflection coefficient.

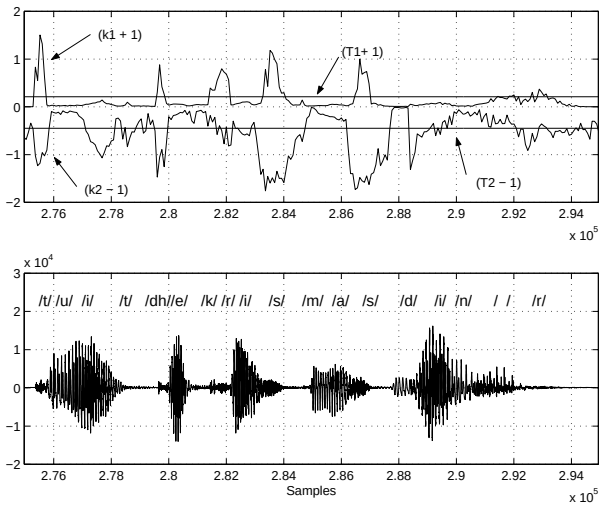

Figure 3: *Illustration of the variation of the first two reflection coefficients obtained from a 4kHz bandwidth input speech signal with the corresponding wideband speech signal spoken by an english male speaker saying: "to eat the christmas dinner." $k1$ and $k2$ are the reflection coefficients, and $T1 = -0.7$ and $T2 = 0.55$ are their respective thresholds.*

Figure 3 shows a wideband speech signal of a male speaker simultaneously with the corresponding curve of variation of $(k_1 + 1)$ (the curve in the top of the upper-figure) and $(k_2 - 1)$ (the curve in the bottom of the upper-figure) where k_1 and k_2 are respectively the first and the second reflection coefficient obtained from input speech linear prediction analysis. The first observation is that there is some

Table 1: *Reflection coefficient thresholds (T_1, T_2) for 4 matrices*

Ref. Coeff. Threshold	T_1 = -0.7	T_2 = 0.55
M_1	$k_1 \leq$	$k_2 \geq$
M_2	$k_1 <$	$k_2 <$
M_3	$k_1 >$	$k_2 >$
M_4	$k_1 \geq$	$k_2 \leq$

correlation between the first two reflection coefficients and the nature of the speech signal, namely, whether it is a noise like signal such as unvoiced speech sounds or a periodic like and energetic signal such as voiced speech sounds. This correlation can be emphasised by putting a threshold on both reflection coefficients that can distinguish the two categories of sounds. Thresholds are fixed experimentally such as to span the voiced sounds within the same category. The first reflection coefficient threshold is set to T_1 = -0.7 and the second one is set to $T_2 = 0.55$. The threshold T_1 is represented by $(T_1 + 1)$ in the upper-figure 3 and the threshold T_2 is represented by $(T_2 - 1)$ in the same figure.

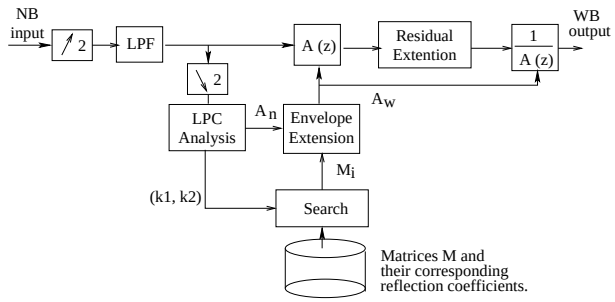

Figure 4: *Bandwidth extension block diagram based on line spectral frequencies using 4 matrices*

In this way, the speech spectral shapes are classified into 4 clusters and thus, 4 matrices are created to achieve the lowband-to-highband mapping (Figure 4). Figure 4 shows the modified block diagram for wideband synthesis to account for the multiple matrices. Table 1 provides the matrix allocation to the different combinations of the first two reflection coefficients as a function of their values relative to their respective thresholds. The matrix M_1 represents the voiced speech sounds where the lowband is energetically more significant than the highband (Figure 5). In the opposite, the matrix M_4 represents the speech sounds where the high-frequency band is more energetic than the low-frequency band as is the case for /s/ speech phones. The remaining two matrices are transition matrices that take different shapes. The matrix M_2 can be a silence or a transition from any sound (whose M_1 is the representative) to another (whose M_4 is the representative) or vice-versa. The matrix M_3 could be an unvoiced sound such as /ch/ or represent a transition speech sound as the matrix M_2.

Given the definition of the matrix representations, four matrices are created. A database of speech signal frames is clustered into 4 clusters where each cluster is represented by one of the 4 matrices. Then, an estimation of the matrix coefficients for each cluster that maps the lowband LSF to

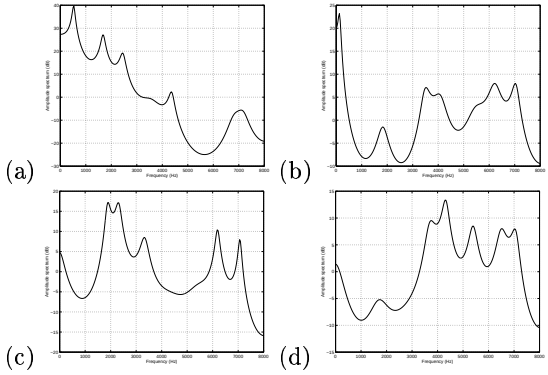

Figure 5: *Amplitude spectral envelope shape represented by the 4 matrices from Table 1. (a) a typical amplitude spectral envelope shape for matrix M_1. (b) an example of an amplitude spectral envelope shape for matrix M_2. (c) an example of an amplitude spectral envelope shape for matrix M_3. (d) a typical amplitude spectral envelope shape for matrix M_4*

the highband LSF is achieved as given in section 2.1.1.

Given the matrices, the spectral envelope extension from input spectral envelope is achieved as follows. For each speech entry, linear prediction and reflection coefficients are estimated. Then, the LSF are computed and divided by 2 to represent the lowband of the wideband speech synthesis (Figure 6). Given the first two reflection coefficients of each entry, a matrix is selected from table 1. The selected matrix is used to estimate the highband LSF from the lowband LSF. Once the highband LSF are estimated, they are appended to the lowband ones. From the obtained array of LSF, we determine the linear prediction coefficients of the extended spectral envelope.

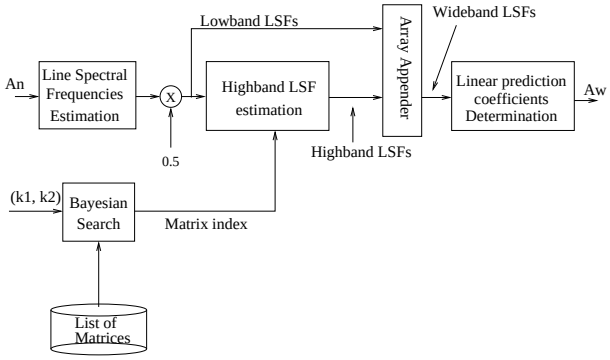

Figure 6: *Spectral envelope extension based on statistical models of the relationship between the lowband and the highband line spectral frequencies*

2.2. Discussions

A good synthesis quality is obtained using the proposed algorithm. To reach this quality, a post-processing discussed in this section is applied to the estimated highband LSF to eliminate some undesirable sounds due to the statistical models (i.e. matrices). The estimated highband LSF create sometimes a whistling on the wideband synthesis signal. The source of these noises has been identified as being the high order LSF component getting closer to or taking a

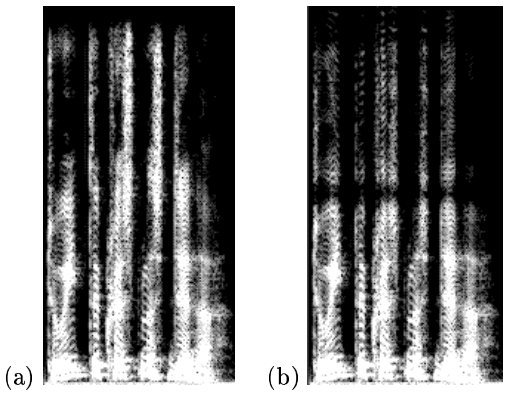

Figure 7: *Bandwidth extension based on four statistical models mapping lowband LSF to highband LSF. English male speaker saying: "to eat the christmas dinner" (a) Wideband original signal, (b) Wideband synthesis*

higher value than 3.0 *radians*. In order to encounter this problem, the higher order LSF component has been limited to [2.5762 2.82] radians. In this LSF post-processing, the entire set of the highband LSF are re-adjusted to fit the high order LSF in the limited range. The dynamics of the LSF as a function of time may also be an issue. They are sometimes quite significant to generate slight clicks perceivable in the wideband synthesis signal. So in order to reduce these dynamics of the LSF, a smoothing is applied to each LSF as a function of time. The smoothing function is a sum of weighted LSF involving the current one and the previous ones in time (in our implementation to the third frame back in time). It provides a smooth synthesis signal where the clicks are removed. However, the quality of the synthesis in reaction to this smoothing looses some brightness. So the weighting coefficients need to be tuned in order to obtain the desired quality. A synthesis by interpolation [9] has been tried but it gave a lower brightness than the LSF smoothing based synthesis.

Figure 7 shows spectrograms of a wideband original and a wideband synthesis for a male English speaker. The synthesis is performed frame-by-frame of $5ms$ (i.e., 40 samples) length of the input signal. Each frame is analysed using an $8th$ order LPC and thus, a $16th$ order LPC is used for the wideband synthesis. Then, four matrices of 8x8 dimension have been created by training and used to achieve the spectral envelope extension.

The use of the proposed algorithm for telephone speech (i.e., 300-3400Hz) enhancement does not require any change to the described algorithm. However, the signal bandwidth of 4kHz, given its sampling frequency, and the band-pass filtering for narrowband speech lead the wideband synthesis signal to a lack of information in the frequency range 3.4-4.6kHz. This is a result of the spectral folding (relative to the 4kHz frequency) of the LPC analysis residual which is limited in frequency to the telephone speech bandwidth (i.e., 3.4kHz). One solution would be to down-sample the telephone speech from 8kHz to 7kHz sampling frequency. Then, the extension can be performed from 3.5kHz bandwidth to 7kHz bandwidth [10].

Finally, the proposed algorithm trained on speech has also been used for music bandwidth extension and it gave an acceptable quality.

3. CONCLUSIONS

In this paper, a new algorithm for bandwidth extension is proposed for telephone speech enhancement. The extension is achieved using efficiently the line spectral frequencies (LSF). The extension is processed according to the first two reflection coefficients. The variation of these coefficients showed interesting behaviour as a function of the nature of the input speech signal (i.e., noise like or periodic.) This observation has been exploited to create four clusters of distinctive spectral shapes. These clusters are represented by four transform matrices that map the lowband LSF to the highband LSF. In addition, the proposed algorithm has low complexity and requires reduced amount of memory for a storage of the matrices.

4. REFERENCES

[1] "Paired Comparison test of wide-band and narrow-band telephony," ITU, COM 12-9-E, March 1993.

[2] Yasukawa H., "Wideband Speech Recovery from bandlimited Speech Using LPC Analysis/Synthesis," *NTT DSP Symposium*, pp. 409-412, 1997.

[3] Miet G., Gerrits A. and Valiere J.C., "Low-Band Extension of Telephone-Band Speech," *Proceedings of Int. Conf. on Acoustics, Speech and Signal Proc.*, 2000.

[4] Makhoul J., Berouti M., "High-Frequency Regeneration in Speech Coding Systems," *Proceedings of Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 428-430, 1979.

[5] Itakura F., "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals," *J. of Acoust. Soc. of Am.*, 57, supplement No 1, S35, 1975.

[6] Kang G.S. and Fransen L.J., "Application of Line Spectrum Pairs to Low Bit-Rate Speech Encoders," *Proceedings of Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 7.3.1-7.3.4, 1985.

[7] Epps J. and Holmes W.H., "A New Technique for Wideband Enhancement of Coded Narrowband Speech," *Proc. Speech Coding Workshop*, pp. 174-176, 1999.

[8] Markel J.D. and Gray A.H. Jr., "*Linear Prediction of Speech*," Springer-Verlag, Berlin Heidelberg New-York, 1976.

[9] Nakatoh Y., Tsushima M., and Norimatsu T., "Generation of Broadband Speech From Narrowband Speech Using Piecewise Linear Mapping," *Proceedings of EUROSPEECH*, Vol. 3, pp. 1643-1646, 1997

[10] Yasukawa H., "Signal Restoration of Broad Band Speech Using Nonlinear Processing," *Proceedings of EUSIPCO*, pp. 987-990, 1996.