LOW-BAND EXTENSION OF TELEPHONE-BAND SPEECH

G. Miet[*], *A. Gerrits*[**], *J.C. Valière*[* **]

[*] Philips Consumer Communications, Route d'Angers 72081 Le Mans cedex 9, France
[**] Philips Research Laboratory, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands
[* **] Laboratoire d'Acoustique de l'Université du Maine, CNRS-UMR 9913, Av. O. Messiaen, 72085 Le Mans, France

ABSTRACT

This paper describes a system that generates a low-band signal (100 – 300 Hz) from a telephone-band (300 – 3400 Hz) speech signal to obtain an extended-band speech signal (100 – 3400 Hz). The low-band increases signal naturalness and listening comfort. This system is applied at the receiving end such that compatibility with all current telephone networks is maintained. The described technique splits the telephone-band speech signal into a spectral envelope and a short-term residual. The spectral envelope and the residual are extended separately and recombined to create an extended band signal. This system is evaluated by listening tests and distortion measurement.

1. INTRODUCTION

1.1 Benefits of low-band extension

Current telephone networks use speech with a bandwidth that, in most cases, is limited to 300 – 3400 Hz. Speech with such a bandwidth is referred to as narrow-band or telephone-band speech. This is sufficient for a satisfactory telephone conversation, but a natural speech bandwidth is much wider. By adding the low-band (100 – 300 Hz) to the narrow-band signal, an extended-band signal (100 – 3400 Hz) is obtained. The low-band is important for speech naturalness and speaker identification. This paper proposes a technique to generate this low-band from the narrow-band speech signal. An advantage of such a technique is that it can be applied in terminals like answering machines and telephones without changing current networks.

1.2 Description of the low-band extension system

The low-band signal is generated by a system as is shown in Figure 1. Linear Prediction (LP, [1]) analysis is applied on the narrow-band speech signal to obtain LP parameters. These LP parameters describe the spectral envelope of the narrow-band speech, which corresponds to its short-term correlations. The spectral envelope of the extended-band speech is derived from the narrow-band spectral envelope using an envelope extension method.

The narrow-band LP parameters are used in the analysis filter to remove the short-term correlation in the narrow-band speech. The narrow-band residual signal which is coming out of the analysis filter contains the long-term correlations that are present in the narrow-band signal, resulting in a harmonic structure. If the fundamental frequency is below 300 Hz, the low-band harmonic structure has to be re-instated (harmonics recovery). Also, a noise-like signal is obtained after the removal of the long-term and short-term correlations. To fill the spectral gap between 100 and 300 Hz, synthetic noise is injected to this signal.

By applying the techniques of harmonics recovery and noise injection to the narrow-band residual signal, a synthetic extended-band residual signal is obtained. A synthesis filter, which uses the extended LP parameters, is excited with this extended-band residual signal to produce extended-band synthetic speech. Finally, a band-pass filter removes the useless synthetic signal outside 100-300 Hz. The extended-band speech is simply obtained by adding the band-pass filtered synthetic speech with the narrow-band speech signal as depicted in Figure 1.

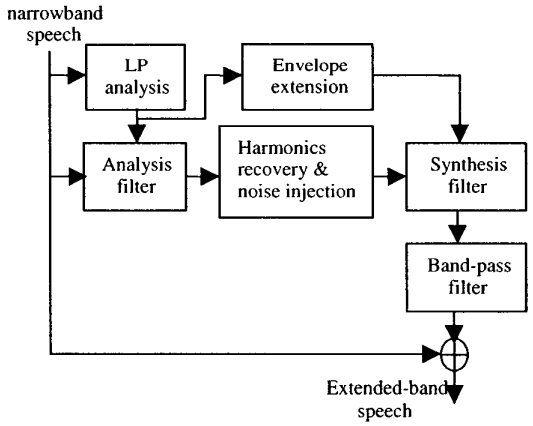

Figure 1: Low-band extension scheme

By separating the extension of the spectral envelope and the residual signal, it is possible to study the effect of each extension.

This paper is organized as follows: Section 2 presents two envelope extension methods and a comparison between them is made. Section 3 describes the residual extension by harmonics recovery and the noise injection. The results of subjective tests with the low-band extension system are given in section 4. Finally, concluding remarks are given in section 5.

2. ENVELOPE EXTENSION

In this section, two different methods to extend the spectral envelope are presented: the Codebook Mapping Method (CMM) and the Extension Matrix Method (EMM).

0-7803-6293-4/00/$10.00 ©2000 IEEE.

2.1 Codebook Mapping Method

The CMM that generates the extended-band LP parameters from the narrow-band LP parameters is derived from [2]. In this method two codebooks are maintained: a narrow-band codebook containing narrow-band LP parameters and a corresponding extended-band codebook containing extended-band LP parameters. The narrow-band LP parameters resulting from LP analysis are compared to the narrow-band LP parameters in the narrow-band codebook and the index of the closest LP parameters is used in the extended-band codebook. The extended-band LP parameters corresponding to this index are taken. This mapping process is shown in Figure 2.

Although it is possible to apply this mapping on the LP parameters themselves, experiments showed that better results are obtained when the LP parameters are converted to a different representation, namely Line Spectral Frequencies (LSF, [4]). The Euclidian norm is used to find the best match in the narrow-band codebook.

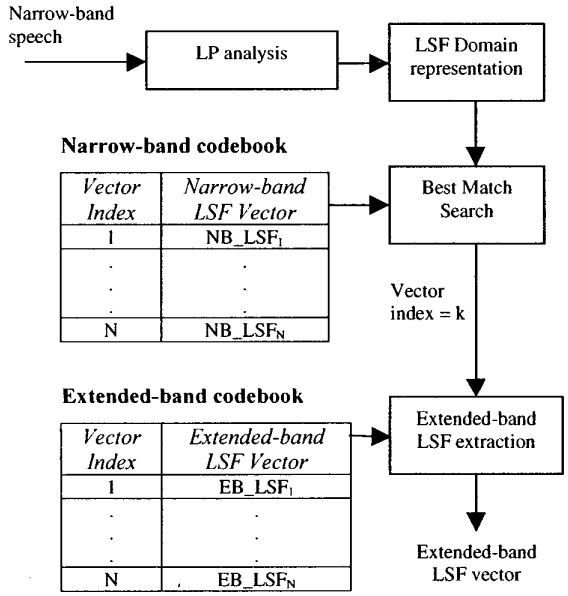

Figure 2: Codebook mapping applied on LSF vectors

The codebooks are obtained by training. A training set of extended-band LSF vectors and corresponding narrow-band LSF vectors is first collected from both extended-band and corresponding narrow-band speech. Then, the extended-band codebook is made using the method given in [3], where the LSF vectors are clustered and the centroïde of each cluster is computed. The corresponding narrow-band codebook is obtained by clustering the narrow-band LSF vectors in the same way as is done for the extended-band codebook. Thus, each LSF vector of the extended-band codebook is linked with a vector of the narrow-band codebook, enabling a mapping function based on the index in the codebooks.

There are a number of limitations to the CMM. First, the number of synthesis filter shapes is limited to the size of the codebooks. Secondly, the generation of the codebooks is not optimal. Here, the assumption is made that if several extended-band LSF vectors belong to a given cluster, their corresponding narrow-band LSF vectors will belong to the same cluster, which is not always the case. Therefore, a mismatch in the mapping from narrow-band to extended-band LP parameters can occur. The number of mismatches increases if the codebook size is increased.

To overcome these limitations, another method called EMM was developed, which is presented in next section.

2.2 Extension Matrix Method

Instead of using codebooks, the EMM uses a linear function to derive the extended-band spectral envelope from the narrowband spectral envelope. Like in the CMM, the LP parameters, which describe the spectral envelope, are converted to a LSF vector.

Let $\tilde{\mathbf{w}}_e = (\tilde{w}_e(1), \tilde{w}_e(2), \ldots, \tilde{w}_e(P))^t$ denote the synthetic extended-band LSF vector and $\mathbf{w}_n = (w_n(1), w_n(2), \ldots, w_n(P))^t$ the narrow-band LSF vector, both being of order P, where $w_n(i)$ represents the i[th] narrow-band LSF and $\tilde{w}_e(i)$ represents the i[th] synthetic extended band LSF. The extension matrix $\mathbf{M}$ is defined by:

$$\tilde{\mathbf{w}}_e^{\,t} = \mathbf{w}_n^{\,t} \cdot \mathbf{M}, \qquad (1)$$

where $\mathbf{M}$ is a PxP matrix.

After conversion of $\tilde{\mathbf{w}}_e^{\,t}$ to LP parameters, these extended-band LP parameters are used in the synthesis filter. Thus, the spectral envelope extension is computed by multiplying the narrow-band LSF vector by the extension matrix giving a synthetic extended-band LSF vector.

Equation (1) requires a pre-calculation of the matrix $\mathbf{M}$. The matrix coefficients $m(i,j)$ are estimated with the same training set as in the CMM. It is composed of N extended-band LSF vectors and their corresponding narrow-band LSF vectors.

The matrix coefficients are chosen to minimize distortion D over the complete training set of size N:

$$D = \sum_{k=1}^{N} \sum_{j=1}^{P} (w_{e,k}(j) - \tilde{w}_{e,k}(j))^2, \qquad (2)$$

with $\tilde{w}_{e,k}(j) = [m(1,j) \quad \ldots \quad m(P,j)] \cdot \mathbf{w}_{n,k}$, where $\mathbf{w}_{e,k}$ and $\mathbf{w}_{n,k}$ are the k^{th} extended-band and narrow-band LSF vectors, respectively.

To solve equation (2), we rewrite equation (1) as $\tilde{\mathbf{W}}_e = \mathbf{W}_n \cdot \mathbf{M}$. $\mathbf{M}$ that minimizes (2) is thus computed by:

$$\mathbf{M} = (\mathbf{W}_n' \mathbf{W}_n)^{-1} \mathbf{W}_n' \mathbf{W}_e, \qquad (3)$$

where each row of $\mathbf{W}_n$, $\mathbf{W}_e$ and $\tilde{\mathbf{W}}_e$ corresponds to a narrow-band LSF vector, its original extended-band LSF vector and the computed extended-band LSF vector, respectively.

The resulting extension matrix, however, is likely to lead to extended-band LP parameters that result in an unstable synthesis filter. This was confirmed by simulations where a considerable number of extended vectors did not fall in the LSF domain.

A sufficient condition in the LSF domain to guarantee a stable filter is:

$$0 < w(1) < w(2) < \cdots < w(P) < \pi \qquad (4)$$

Given that this constraint is strong, a good compromise was found by forcing the matrix coefficients $m(i,j)$ to be positive which gives:

$$\mathbf{M}=\arg\min_{\mathbf{N}}\left\{\left|(\mathbf{W}_e-\mathbf{N}\mathbf{W}_q)^t(\mathbf{W}_e-\mathbf{N}\mathbf{W}_q)\right|, n(i,j)\geq 0, \forall(i,j)\in[1..P]^2\right\}, \quad (5)$$

The algorithm that was used to solve the above formula, called the Non Negative Least Squares (NNLS), is described in [5].

This constraint which only makes sure that the LSFs coefficients are not negative, appeared to be sufficient to fall in the LSF domain in most cases. If, a synthetic extended-band LSF vector does not satisfy (4), its LSF components are modified such that this condition is met.

2.3 Comparison between the two methods

To study the envelope extension, the system as shown in Figure 1 is used with one modification, namely the synthesis filter is excited with the original extended-band residual instead of the synthetic narrow-band residual. Both methods generate effectively the expected effect. However, the CMM produced a slightly rattling background sound in some cases. The speech quality of the extended-band speech, in the system where the EMM is applied, was considered better, mainly due to the absence of annoying artifacts.

This is also confirmed by an objective measurement, where the spectral distortion between the original extended-band LP parameters and the extended-band LP parameters is measured. The Log Spectral Distance (LSD) distance d between the original extended-band LP parameters $\mathbf{a}$ and the synthetic extended-band LP parameters $\tilde{\mathbf{a}}$ is defined as:

$$d = \sqrt{\frac{1}{f_2 - f_1}\int_{f_1}^{f_2}\left[10\log P(f) - 10\log \tilde{P}(f)\right]^2 \cdot df}, \quad (6)$$

with $P(f)=\left|A(e^{\frac{j2\pi f}{f_s}})\right|^{-2}$, $\tilde{P}(f)=\left|\tilde{A}(e^{\frac{j2\pi f}{f_s}})\right|^{-2}$ and $A(z)=\sum_{i=0}^{P} a(i)z^{-i}$

The distortion is only measured in the low-band, so $f_1 = 100Hz$ and $f_2 = 300Hz$ ($f_s = 8kHz$). This distance was averaged over a speech database made by English, French and German speakers. Each language was represented by one male and one female speaker. This database was not used for training the algorithms. The EMM results in a lower mean distance (3.11 dB) than the CMM (4.17 dB) and consequently has a better match in the extended band. Furthermore, the EMM is less memory consuming that the CMM. In case $P = 10$, the EMM requires 10x10 = 100 matrix coefficients to be stored compared to 2x256x10 = 5120 LSF coefficients for the CMM in case the size of the codebooks is 256.

3. RESIDUAL EXTENSION BY HARMONICS RECOVERY AND NOISE INJECTION

In the narrow-band residual, no spectral contribution below 300 Hz is present. As already mentioned, the narrow-band residual signal is extended in two steps: first harmonics recovery is applied followed by noise injection. The aim of harmonics recovery is to fill in the missing harmonics in the low-band. This is only applied for voiced speech where a pitch is present. For unvoiced speech or if the fundamental frequency is higher than 300 Hz, this method is not applied. Therefore, a voicing and pitch detector is incorporated in the system. Examples of both kinds of detectors are described in [1][11]. Let H denote the number of harmonics below 300 Hz and f_0 the fundamental frequency. The harmonic low-band residual signal e is then generated by:

$$e(n) = \sum_{i=1}^{H} A_i \cdot \cos(2\pi i f_0 + \varphi_i), \quad (7)$$

where A_i is the amplitude of the i^{th} harmonic and φ_i is the phase. The amplitudes and the phases have to be determined. Given that the LP analysis filter flattens the spectral envelope, the amplitude is simply set to the averaged amplitude of the harmonics above 300 Hz. The phase is synthesized according to:

$$\varphi_{i,k} = \left[\varphi_{i,k-1} + \pi + 2\pi i L \cdot \left(\frac{3}{4}f_{0,k} - \frac{1}{4}f_{0,k-1}\right)\right]_{2\pi} - \pi, \quad (8)$$

where k represents the k^{th} frame and L represent the frame length. The brackets indicate the modulo-operation. By having such a continuous phase, discontinuities between two speech frames are avoided.

Regardless of the voicing decision, noise is injected in the low-band. This is achieved by adding noise with a bandwidth of 100-300 Hz to the residual signal. The energy of the noise is 10 dB lower than the energy of the narrow-band residual signal.

4. SYSTEM EVALUATION

This section gives the results of subjective tests that are performed with the low-band extension system as is shown in Figure 3. This figure shows the complete system including envelope extension, harmonic recovery and noise injection. For envelope extension the CMM and EMM are used. The order of LP is set to 10. The CMM uses two codebooks of size 256 each.

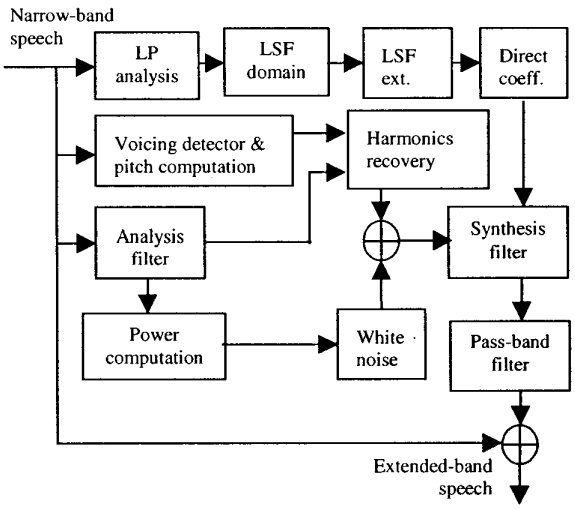

Figure 3: Low-band synthesis system

The subjective evaluation of the low-band synthesis systems was conducted by listening tests derived from the CCR MOS (CMOS) test method described in [6][7][8] with audio samples processed at – 26 dBov as stated in [9]. Six listeners were presented with two different stimulus of the same 4 seconds 100-3400 Hz band audio sample: a reference and a test sample in a random order. The reference sample was either the targeted (original 100-3400 Hz) signal or the corresponding IRS narrow-band signal [8] extended by the CMM. The test sample was obtained by applying the EMM on the IRS narrow-band signal. They had to judge the second stimuli with regard to the first on a seven-quality scale as follows:

0: no preference
1 (-1): slight preference for the second (first)
2 (-2): preference for the second (first)
3 (-3): clear preference for the second (first)

The audio samples used in the experiment were chosen as follows:

- 8 clean speech samples
- above speech samples with car noise (SNR 15 dB)
- above speech samples with street noise (SNR 15 dB)
- 2 music samples

The clean speech was recorded from 4 French speakers (2 males and 2 females). Half of the speakers were part of the training database. The CMOS scores of the EMM versus both the targeted signal and the CMM are given in Table 1 for several tested conditions. The condition "training independent" uses the speakers that are not used for training the codebooks and matrices.

	EMM vs. CMM		**EMM vs. Target**	
	CMOS	*Confidence threshold*	*CMOS*	*Confidence threshold*
All conditions	1.30	0.24	-0.50	-0.27
Training dependent	0.89	0.30	-0.51	-0.41
Training independent	1.67	0.35	-0.42	-0.37
Noisy speech	1.21	0.30	-0.45	-0.34
Clean speech	1.42	0.42	-0.50	-0.48
Music	1.50	1.34	-0.92	-1.62

Table 1: CCR test results of the EM algorithm (Positive values mean that the EMM is better than the CMM/Target)

To compare the EMM with both the CMM and the targeted 100-3400 Hz signal, the 1-tailed T-test was used at the 99% confidence level [10]. Table 1 shows that EMM has a better speech quality than the CMM. However, the Target signal (100-3400 Hz) is slightly preferred over the EMM. Table 1 shows further that the EMM is fairly robust to non-speech signals and much less sensitive to training than the CMM.

The extended-band signal, generated by the EMM, was also compared to the narrow-band signal. The extended-band signal was preferred over the narrow-band signal. However, some listeners judged the low-band effect as too strong, resulting in a considerable inter-listener CMOS variance ($\sigma^2 = 1.56$). A solution can be that the listener adjusts the energy level of the injected noise to control the low-band effect.

5. CONCLUSION

This paper has presented a new system to obtain extended-band speech (100 – 3400 Hz) from narrow-band speech (300 – 3400 Hz) at the receiving end (e.g. a telephone or an answering machine). No additional information about the original extended-band signal is required. Listening tests showed that the quality of the synthetic extended-band speech approaches the quality of the original extended-band speech.

The system creates a synthetic low-band excitation signal to feed a synthesis filter that shapes the low-band spectral envelope. The excitation signal is synthesized by harmonic recovery and noise injection. The synthesis filter can be estimated by two possible envelope extension methods: Codebook Mapping Method and Extension Matrix Method. As a result of subjective tests and distortion evaluations, the Extension Matrix Method is preferred.

Although this technique is used to extend the narrow-band signal in the low-band, similar techniques can be used to extend the narrow-band signal to the high-band (3400 – 7000 Hz) to obtain wide-band speech.

6. REFERENCES

[1] W.B. Kleijn, K.K. Paliwal, "Speech Coding and Synthesis", Elsevier, 1995.

[2] M. Abe, Y. Yoshida, "More natural sounding voice quality over the telephone", NTT Review, Vol.7, No 3, May 1995.

[3] Y. Linde, A. Buzo, R. M. Gray, "An algorithm for Vector Quantizer Design", IEEE Transactions on Communications, Vol. COM-28, No 1, January 1980.

[4] F. Itakura, "Line spectrum representation of linear predictive coefficients", J.A.S.A, vol. 57 Supplement no. 1, p. S35, 1975.

[5] Lawson and R. J. Hanson, "Solving Least Squares Problems", Prentice-Hall, 1974.

[6] ITU-T Rec. G. 191, "Software Tools Library 96", 1996.

[7] ITU-T Rec. P.800, "Method for subjective determination of transmission quality", 08/96

[8] ITU-T Rec. P.830, "Subjective performance assessment of telephone-band and wideband digital codecs", 02/96

[9] ITU-T Rec. P.56, "Objective measurement of active speech level", 03/93

[10] Thomas H. Wonnacott and Ronald J. Wonnacott "Introductory Statistics for Business and Economics", fourth Edition, John Wiley & Sons, New York, 1990

[11] W. Hess "Pitch determination of speech signals", Springer Verlag, 1983.