Bandwidth extension of narrowband speech

Miquel Expósito Pérez[(1)], Josep M. Salavedra[(1)]
miquel.exposito.91@gmail.com , josep.salavedra@upc.edu
[(1)]Dept. of Signal Theory and Communications. TALP Research Centre. UPC University
Campus Nord UPC, Edifici D5, c/ Jordi Girona 1-3, 08034 Barcelona, Spain

***Abstract*- Recently, 4G mobile phone systems have been designed to process wideband speech signals whose sampling frequency is 16 kHz. However, most part of mobile and classical phone network, and current 3G mobile phones, still process narrowband speech signals whose sampling frequency is 8 kHz. During next future, all these systems must be living together. Therefore, sometimes a wideband speech signal (with a bandwidth up to 7,2 kHz) should be estimated from an available narrowband one (whose frequency band is 300-3400 Hz). In this work, different techniques of audio bandwidth extension have been implemented and evaluated. First, a simple non-model-based algorithm (interpolation algorithm) has been implemented. Second, a model-based algorithm (linear mapping) have been designed and evaluated in comparison to previous one. Several CMOS (Comparison Mean Opinion Score) [6] listening tests show that performance of Linear Mapping algorithm clearly overcomes the other one. Results of these tests are very close to those corresponding to original wideband speech signal (see Fig.5 and Fig.6).**

I. INTRODUCTION

Since the beginning of history, human speech has been the most natural communication system used by humans. That's the reason why phone system has been so successful. Signals emitted in that age were analogical and limited in frequency band, due to the physical restrictions of the acoustic components and the bandwidth capacity. With the beginning of the digital transmissions, using the modulation by coded impulses (PCM), a sampling frequency of 8 kHz and an audio bandwidth from 300 to 3400 Hz were adopted. That was useful to have a good adaptation and compatibility with the analogic network.

Currently, speech quality does not satisfy a large part of the users. Specially by considering that other hearing systems, like the compact disc or the radio have better quality. The aim is to achieve a level of quality where the voice does not sound dull and the naturalness of the speech is not missed due to the lack of high frequency components

Speech quality degradation in analogic phone systems is due to the use of limitation filters, inside the amplifiers, used for maintaining a specific signal level. These filters have a bandwidth from 300Hz to 3,4kHz and they have the goal of reducing the crosstalk between different channels. However, digital network systems are able to transmit a better quality of speech signal (frequency components lower than 300 Hz and those higher than 3,4 kHz).

During last years, many research works have been done regarding the bandwidth extension of phone speech signals. Some codecs have been developed to transmit phone speech with a bandwidth up to 7,2 kHz. But it has an important problem: it needs to use a large part of the nowadays phone network that works under narrowband conditions. A change of the whole phone network, used up to now, is not feasible due to its economic cost. So, the option of extending the bandwidth of the signals after the transmission is the best transitory choice. The idea of this improvement consists of estimating frequency components that would be below 300 Hz and above 3400 Hz, add them to the original narrowband speech signal, and obtain a broadband signal (see Fig 1).

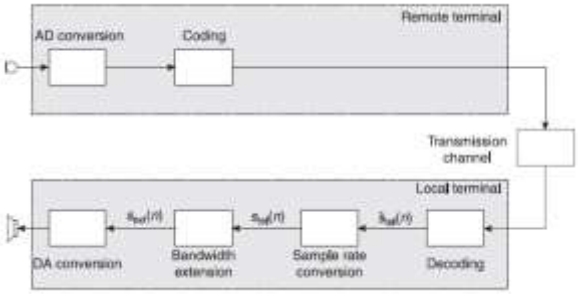

Fig. 1. Diagram of bandwidth extension in the reception host

II. NON-MODEL-BASED ALGORITHM

This interpolation technique makes use of the spectral components that appear while upsampling a signal, with a filter showing only a slow decay above half of the desired sampling frequency. That filter has been designed as a low-pass filter with cut-off frequency at 90% of the sampling frequency.

This method profits from the noise-like nature of the excitation signal concerning unvoiced utterances. Unvoiced frames have a broadband spectrum with most of its energy in the higher frequency regions and, therefore, these portions are exactly mirrored. Temporal behaviour of unvoiced frames below the cut-off frequency strongly correlates with those above the cut-off frequency, which is preserved by applying this mirroring.

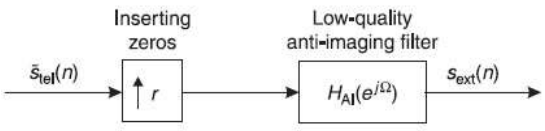

Fig. 2. Diagram of bandwidth extension by interpolation

A drawback of this technique is that lower frequency part of the spectrum is not extended at all. This method works quite well with original signals limited at 8 kHz and bandwidth extended until 12 kHz (for example), but results are poor when it is applied to phone speech signals.

III. SOURCE-FILTER MODEL

Algorithms, presented in section IV of this work, make use of the Source Filter Model. That model is based on the anatomical analysis of the human speech apparatus and how it is generated (see Fig. 3). The source-filter model has been largely accepted in many speech signal applications, specially in speech coding and speech synthesis.

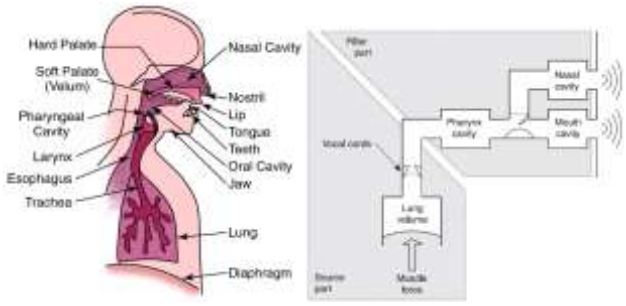

Fig. 3. Real and schematic phonetic vocal tract

The human voice is generated from the lungs as a flow of air which runs through the larynx to reach the vocal cords. The total length of the vocal tract, from the larynx to the lips, is about 17 cm for an adult male, and 13.5 cm for a female. By changing length and cross section profiles of the vocal tract, mostly by moving the lips, jaw, tongue and velum, humans are able to produce different speech sounds [3].

The aim of the source-filter model is recreating two scenarios. To create the excitation signal, this model uses two different generators depending on whether we want voiced or unvoiced utterances:

- For voiced utterances i.e. periodic utterances, we generate a pulse train signal, whose pulses are separated by a period, which is the inverse of the pitch frequency of the speech signal that we want to synthetize.

- For unvoiced utterances we generate white noise signal.

The filter H(z), that allows the reconstruction of the oral cavity influence, is described as a low order filter (only poles)

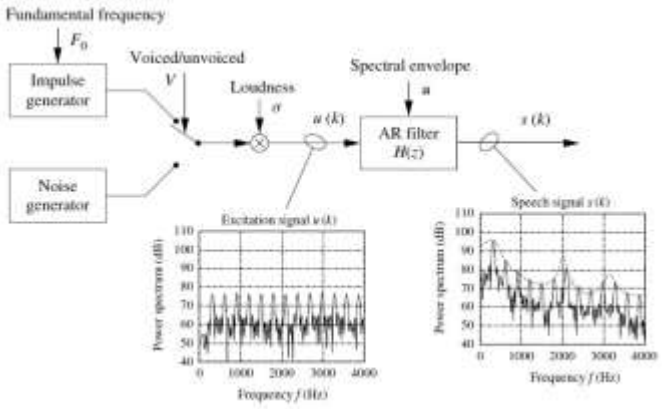

Fig. 4. Source-Filter model scheme

$$H(e^{j\pi}, \mathrm{n}) = \frac{\sigma(\mathrm{n})}{\mathrm{A}(e^{j\pi},\mathrm{n})} = \frac{\sigma(\mathrm{n})}{1-\sum_{i=1}^{Npre} ai(n)e^{-j\Omega i}} \qquad (1)$$

Filter order, called N_{pre}, is subjectively chosen and is usually between 8 and 20. Excitation signal for this filter is spectrally plain (see figure 4), and transfer function of this model represents directly the spectral envelope of the speech signal. Coefficients, $a_i(n)$, in expression (1), have been calculated by solving the Yule-Walker equations [3], by using Levinson-Durbin recursive model:

$$\underline{\underline{R_{ss}}}.\underline{\tilde{a}} = \underline{r_{ss}} \qquad (2)$$

This model also obtains the filter gain as follows:

$$\sigma(n) = \sqrt{r_{ss,0}(n) - \sum_{i=1}^{N_{pre}} \tilde{a}_i(n)\, r_{ss,i}(n)} = \sqrt{r_{ss,0}(n) - \tilde{\mathbf{a}}^{\mathrm{T}}(n)\mathbf{r}_{ss}(n)} \qquad (3)$$

Due to the non-stationary speech behaviour, these parameters have to be estimated every frame, where speech signal can be considered as a stationary signal.

The spectral envelope has a very compact representation. That's why the prediction coefficients $a_i(n)$ have a lot of importance in speech coding and bandwidth extension. These coefficients have been transformed to Cepstrum coefficients. Cepstrum coefficients allow us to take into account the characteristics of the human auditory perception.

IV. MODEL-BASED ALGORITHM

In contrast to previous algorithm, this technique makes use of the explained source-filter model, and thereby of a previous knowledge of the speech signal. It has been separated in two main sub-tasks:

- A) Generation of the broadband excitation signal.
- B) Estimation of the broadband spectral envelope.

The input signal has been sampled to 8 kHz and filtered, so it has got only frequency information up to 3,4 kHz. After that, excitation signal is extracted by using a simple prediction filter:

$$e_{nb}(n) = \mathrm{s_{tel}(n)} + \sum_{i=1}^{Npre,nb} \alpha_{nb,i}(n)\, s_{tel}(n-i) \qquad (4)$$

To estimate broadband excitation signal, we make use of this narrowband excitation signal. On the other hand, spectral envelope of the narrowband speech signal is used to estimate spectral envelope of the broadband speech signal. Finally, we combine these two parts of the model with an inverse prediction error filter:

$$s'(n) = \hat{\mathrm{e}}_{\mathrm{bb}}(\mathrm{n}) - \sum_{i=1}^{Npre,bb} s'(n-i)\alpha_{bb,i}(n) \qquad (5)$$

A. Generation of the broadband excitation signal

To generate broadband excitation signal, two techniques have been considered: frequency modulation and nonlinear processing.

Modulation technique consists of multiplying input signal with a modulation function, i.e. a cosine. Main purpose of this technique is based on signals and systems theory basis: multiplication of all the samples of narrowband excitation signal with a cosine, with frequency equal to 4 kHz, corresponds to the convolution of the signal with two Dirac impulses in the frequency domain at this frequency.

A special technique of modulation has been used in order not to alter low frequencies of the input signal. This technique is called Spectral Folding, which consists in modulating input signal to 4 kHz, so we will obtain the same information we had, but now between 4 and 8 kHz. After that, this new speech signal is filtered with a high-pass filter with a cut-off frequency of 4 kHz and it's added to the original narrowband excitation signal.

B. Estimation of the broadband spectral envelope

The estimation of the spectral envelope is done following next three steps and they are applied each frame of the narrowband signal and the broadband signal.

- Feature extraction: From each narrowband speech signal frame, several features x which carry information on the state of the source model are extracted, that is, indirectly, on the estimate spectral envelope of the missing frequency band. By the feature extraction algorithm, the dimension and complexity of the estimation problem are significantly reduced.

- A priori knowledge: A previously knowledge of the broadband signal is needed to compare vectors x and y'. That knowledge is previously done, in a training stage.

- Classification or estimation: It is the last step and corresponds to estimating y; that means the spectral envelope of the broadband speech signal. To execute this task, there are several techniques like codebook mapping, the linear mapping or piecewise-linear mapping, Bayesian estimation… [4] In this work, linear mapping technique has been considered.

Linear mapping technique consists of a multiplication of feature vector of the narrowband speech signal (it contains the narrowband spectral envelop information) with a matrix that allows to estimate features of the spectral envelope of the broadband speech signal. This algorithm has both low computational complexity and low memory requirements.

Vector x(n) contains narrowband spectral envelope information, and vector y(n) contains wideband spectral envelope information. So, estimation of broadband spectral envelope can be done by applying the following linear operation:

$$y(n) = \boldsymbol{W}(\boldsymbol{x}(n) - m_x) + m_y \quad (6)$$

In this work, these features are Cepstrum coefficients, but it is possible to use other ones like prediction coefficients, line spectral frequencies, etc. Multiplication of the band limited feature vector x(n) with the matrix W can be interpreted as a set of N_y FIR filter operations. Each row of W corresponds to an impulse response which is convolved with the signal vector x(n), resulting in one element of the wideband feature vector y(n). As common in linear estimation theory the mean values of the feature vectors m_x and m_y are estimated within a pre-processing stage. So, W_{opt} matrix is defined as follows:

$$\boldsymbol{X} = [x(0) - m_x, x(1) - m_x, \dots, x(N-1) - m_x],$$
$$\boldsymbol{Y} = [y(0) - m_y, y(1) - m_y, \dots, y(N-1) - m_y],$$
$$\boldsymbol{W}opt = \boldsymbol{Y}\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}})^{-1} \quad (7)$$

In order to improve this algorithm, several matrixes W have been used depending on the differences between the feature vectors. In this work, two matrixes have been considered, one for voiced frames and another one for unvoiced frames (W_v and W_u). That implies that a voiced/unvoiced classification system is previously needed. After that, for all Cepstrum coefficients (voiced and unvoiced frames for the narrowband and broadband signals), the arithmetic average has been calculated and applied.

Voiced-unvoiced classification has been done by using zero-crossing rate (ZCR) technique in conjunction of the frame energy [5]. This technique is a very good indicator of the frequency, at which the energy is concentrated, in the signal spectrum. Voiced utterances present a zero-crossing rate value lower than unvoiced ones (they are similar to white noise). The energy of the audio frame is another parameter for voiced or unvoiced frame classification. Voiced utterances present a higher energy due to their periodicity. On the other hand, unvoiced utterances present low energy.

V. RESULTS

As it was expected, performance of the model-based algorithm is much better in comparison to non-model-based algorithm. Narrowband speech signal has been sampled at f_s=8kHz and a frame length of 128 samples has been considered. Different values of AR model order have been evaluated, from 8 to 20. Estimated wideband speech signals, obtained from available narrowband speech signals, have been compared to original wideband speech signal (sampled at f_s=16kHz), in terms of several spectral distances and listening tests. An AR order of 16 has been obtained as a good trade-off between quality and computational complexity

Performance of model-based algorithm is shown in fig.7, where a time speech signal and its spectrogram are depicted. Of course, there is no distortion in the lower part of speech spectrum. Although, estimated speech spectrum (in the upper part) of some unvoiced frames don't have a lot of energy, in contrast to voiced frames (and several unvoiced frames), significant differences can be appreciated during listening.

By using spectral folding method, some systematic errors are produced. Since pitch frequency of speech signal is not available, reproduced discrete structure in the extended frequency band is inconsistent during voiced sounds, and so discrete frequency components are not correctly placed at integer multiples of the pitch frequency, resulting in a metallic sound or 'ringing' of the enhanced speech. Further, the position of the extended frequency band is invariably determined by the sampling rate and the band limits of the input signal. That means that in telephony there will be a gap between 3.4 and 4.6 kHz. In addition, the upper band limit of the folded signal is determined by the lower band limit of the input speech.

In order to compare estimated wideband speech signals and original ones, several subjective CMOS listening tests have been done. A group of non-trained listeners has been created to do this task, as it is defined in [6]. Every time, listeners must compare two speech signals and give a subjective value from -3 (first speech signal sounds much

better than second one) to +3 (first speech signal sounds much worse than second one). In fig.5, original narrowband speech (as first signal) has been compared to both estimated non-model-based speech (left) and estimated model-based speech (right). Most part of listeners clearly prefers estimated model-based speech in relation to original narrowband one that is also preferred in comparison to estimated wideband non-model-based speech, that's the worse one of all.

In theory, it seems that most part of listeners should clearly prefer original wideband speech in relation to original narrowband one. However, that's not completely true, as it is shown in fig.6, where value +3 is not de most selected one (left side). This CMOS listening test (see right side) was also done in German language (in TU Wien) and results are even slightly worse. Although listeners groups corresponding to fig.5 and fig.6 were different, results shown in fig.6 make even better those results depicted in fig.5. It must be noted that several listeners (see fig.6) prefer original narrowband speech (historic phone speech) in relation to wideband one.

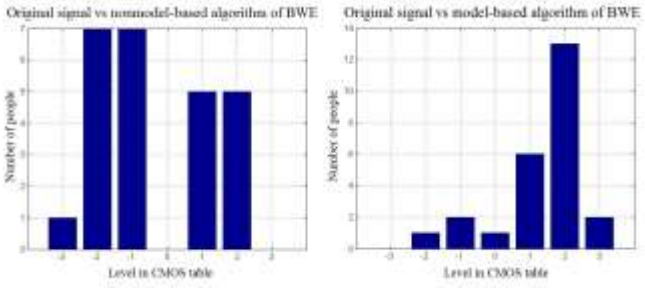

Fig. 5. CMOS tests: (left side) original narrowband speech versus estimated non-model-based one; (right side) original narrowband speech vs estimated model-based speech.

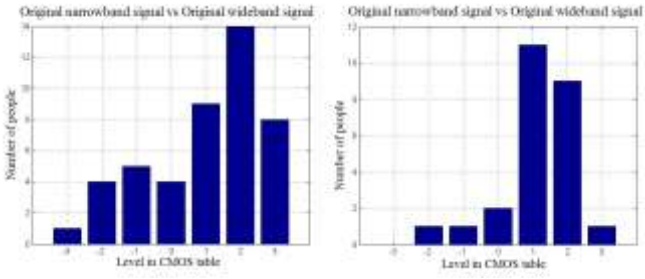

Fig. 6. CMOS test: comparison of original narrowband speech vs original wideband speech in Spanish (left) and German (right).

VI. CONCLUSIONS

In this work, two different bandwidth extension techniques of narrowband speech signals have been presented and evaluated. First algorithm developed here does not make use of the source-filter model, which models phonetic vocal tract, so the results obtained were not so good. Second developed algorithm makes use of the source-filter model, so complexity of this technique is higher and its implementation has been split into two sub-tasks: the extension of the excitation signal (applying the techniques of modulation and linear processing) and the extension of the spectral envelope (solved by using the linear mapping technique).

First algorithm doesn't obtain a very good quality. Its output speech signal sounds quite artificial and metallic, but we obtain a little improve. We can conclude that the other algorithm of bandwidth extension of narrowband speech, model-based algorithm, leads to significantly better results and a quality nearer to the original broadband signal.

ACKNOWLEDGMENTS

To Gerhard Doblinger for his support during the project development.

To all the members that participated in the different CMOS listening tests done to obtain a subjective evaluation of the different techniques presented here.

REFERENCES

[1] E. Larsen, R. M. Aarts. *Audio Bandwidth Extension. Application of Psychoacoustics, Signal Processing and Loudspeaker Design*, 1st ed. John Wiley & Sons, England, 2004.

[2] B.Iser, W.Minker, G.Schmidt: *Bandwidth extension of speech signals*. Lecture notes in electrical engineering Volume 13.Springer, Germany, 2008.

[3] E.Hänsler, G.Schmidt: *Speech and audio processing in adverse enviroments*. 1st ed. Springer. Germany, 2008.

[4] L. Laaksonen. *Artificial bandwidth extension of narrowband speech – enhanced speech quality and intelligibility in Mobile devices*. Aalto University publication series, Finland, 2013.

[5] R.G. Bachu, S. Kopparthi, B. Adapa, B.D. Barkana. "*Separation of Voiced and Unvoiced using Zero crossing rate and Energy of Speech Signal*". Department of electrical engineering. University of Bridgeport, EUA. 2010.

[6] ITU, *Methods for Subjective Determination of Transmission Quality*, ITU-T, 1996.

This project was sponsored by:
Speech and Audio Recognition for Ambient Intelligence,
TEC2010-21040-C02-01

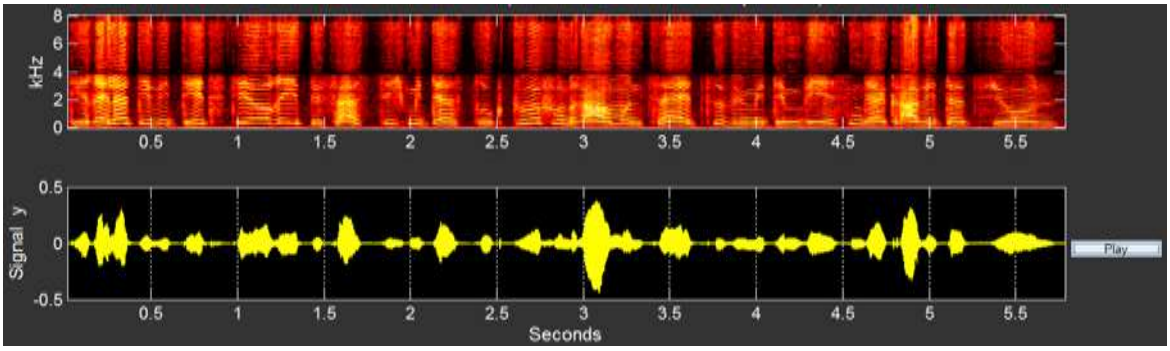

Fig. 7. Estimated wideband speech signal corresponding to model-based algorithm (AR order=16).