BANDWIDTH EXTENSION OF NARROW-BAND SPEECH

Ronald M. Aarts, Jo Smeets and Piet C.W. Sommen

Eindhoven University of Technology
Dep. Electrical Engineering, Building Eh 3.34
P.O.Box 513, 5600 MB Eindhoven
The Netherlands
p.c.w.sommen@tue.nl
Ronald.M.Aarts@philips.com

ABSTRACT

In this paper we describe a first study of a system which can extend the bandwidth of band-limited speech at the low-frequency side. For example speech in telephone communication (narrow-band speech) is band-limited from 300 to 3400 Hz.

The kernel of the system is the harmonics generator, which produces sub-harmonics and higher harmonics of the input signal. The sub-harmonics form a substitute for the low frequencies that are not present in the narrow-band signal. Adding harmonics only works for voiced speech. Extending unvoiced speech does not result in forming a substitute of the original signal, but it has been determined that it did not cause an annoying distortion.

Depending on the harmonic structure of the signal the setting of the harmonics generator must sometimes be modified. When these settings are done manually for a female speech sample the result sounds quite reasonable. Further improvement is obtained by adapting the loudness of the synthesised low frequency part to that of the original low frequency part. Finally an algorithm that can alter the element's settings automatically is discussed.

1. GENERAL INTRODUCTION

Speech can be (re-)produced via several systems. Using those systems it is not always evident, that the whole (audible) spectrum of speech is presented, like in a CD-player for example. Wide-band speech is band-limited from 50 Hz to 7000 Hz and contains almost all information present in a normal speech signal. Narrow-band speech is typically band-limited from 300 Hz to 3400 Hz. Narrow-band speech thus does not contain all information present in normal speech, but is still intelligible.

Telephone communication is done via a narrow-band transmission channel. The advantage is that the bandwidth is small and hence a low sample rate (8 kHz) may be used. Disadvantages are a poor sound quality and an unnatural speech reproduction. Nowadays better performance is expected. Therefore narrow-band speech could be exchanged with wide-band speech. This would not only mean a doubling in sample rate and bandwidth, but also very expensive and time-consuming modifications to the present network.

In [1] a system is described, which applies bandwidth extension to a narrow-band telephone signal at the high end (4–8 kHz). For recovering the envelope, LPC (Linear Prediction Coding) and code-mapping techniques are used. In [2] a spline neural network is used to recover the envelope. The low and high band extension are different, so two neural networks are used. In this paper we present an alternative technique with which it is also possible to extract wide-band speech from its narrow-band equivalent.

The paper is organised as follows. In Section 2 the harmonics generator is discussed. Next the bandwidth extension system is described in Section 3. Then in Section 4 the results are presented and analysed. Finally the conclusions and recommendations are given in Section 5.

2. HARMONICS GENERATOR

The kernel of the system is the harmonics generator. The signal is extended with sub-harmonics as if it were harmonic. Although this is only the case when dealing with the voiced parts of speech, unvoiced speech is extended similarly to voiced speech, which saves the use of a voiced-unvoiced switch.

To compute (sub-) harmonics in a simple way, a non-linear element is used. By non-linear we mean non-linear in frequency but linear in amplitude. The non-linear element used is the so called 'full wave integrator' and is a modification of the device presented in [3]. The system consists of an integrator with a reset function. The integrator integrates the absolute value $|x[n]|$ of the input signal. When only higher harmonics are needed, the reset takes place at every zero-crossing of the input signal (with a positive slope). This results in a saw-tooth containing higher harmonic components.

If sub-harmonics are needed then the element has to be modified. We have to make sure that the output signal contains a periodic component which is a sub-harmonic of the input. This is done by resetting the integrator less often than at every zero-crossing (with a positive slope). When, for example a reset is given after every two zero-crossings with a positive slope, the output signal has a period that is twice that of the input. Its lowest frequency component is therefore half that of the input signal. In Figure 1 examples are presented in the time and frequency domains. A mathematical description of the full wave integrator is given in [3] and [5].

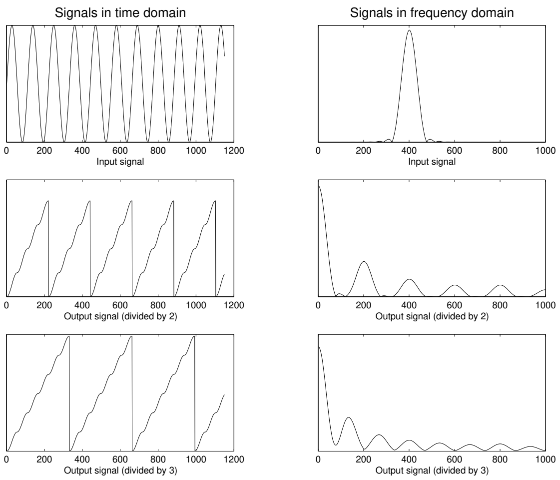

Figure 1: Time and frequency representation of signals using the integrator.

3. DESCRIPTION OF BANDWIDTH EXTENSION SYSTEM

The system[1], is defined as described in this section. A block diagram is given in Figure 2. From the narrow-band input two band-limited parts are taken, one of 300–600 Hz (BPF1a) and one of 300–450 Hz (BPF1b). Two harmonics generators produce two separate streams of harmonics of the selected bands (see also Figure 1). Frequency components in the first band (BPF1a) will be extracted as if the 2nd harmonic is the lowest one present in the narrow-band signal. Frequency components in the second band (BPF1b) will be extracted as if the 3rd harmonic is the lowest one present in the narrow-band signal.

[1] All signals and systems use $f_s = 44.1$ kHz. The speech samples used do not show a pitch below 100 Hz, which means that the lowest component within the narrow-band signal is at most the third harmonic. Signals are taken from [4].

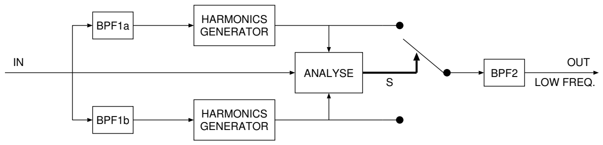

Figure 2: Block diagram of the system. The Harmonics generator in the upper branch generates sub-harmonics one octave below the dominant frequency, while in the lower branch frequencies are divided by three.

The problem now is how to decide which of the two harmonic streams is the most appropriate one. Of course this decision is not valid for the whole excerpt, so this 'best fit decision' should be made every so many samples. Two examples are given in Figures 3 and 4. A wide-band (see Fig. 3) signal (1) is filtered to a narrow-band signal (2). Of this signal the first frequency component (3) is taken to generate two streams of harmonics, one as if this component was the second harmonic of the wide-band signal (4a) and one as if it were the third (4b). Comparison of the plot (2) with plots (5a) and (5b) define which stream is most correlated and will most probably contain the right sub-harmonics (6a) and (6b). In this case (6b) fits best.

For ease we will treat unvoiced speech the same as voiced speech. The most obvious method to be used in an automatic switch that performs comparable with the one used in the manual test which will be described in Section 4.1. The respective spectra of the harmonics streams from the harmonics generators are compared with that of the narrow-band signal.

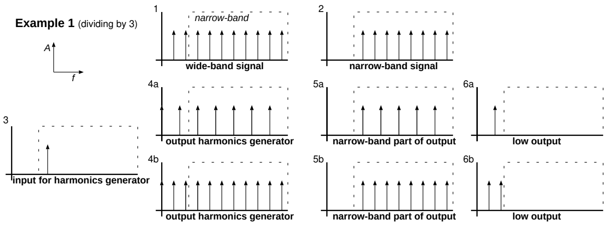

Figure 3: Spectra of the various signals occurring in Example 1.

From both harmonics streams (output from harmonics generators) and the narrow-band signal a frequency band of 300 to 1200 Hz is taken. Then the segmented 1024 point FFTs (spectra) of these signals are computed. Next the amplitudes of the harmonics streams are scaled by dividing them by the square root of their respective auto-correlation of lag zero. From the filtered narrow-band signal the fre-

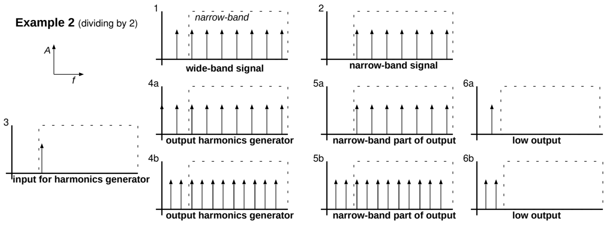

Figure 4: Similar example as in Figure 3. Only in this case (6a) fits best.

quency positions of the peaks above a threshold (5% of the maximum value) are computed. Subsequently a new signal is created with very narrow frequency bins (Kronecker delta functions) at those frequency positions. Then a vector S_i, which is used for decision, is computed as follows:

$$S_i = \sum_{k=1}^{1024} in_k \cdot h_{k,i}, \tag{1}$$

where $h_{k,i}$ are the FFTs of the harmonics streams. The harmonics stream ($h_{k,i}$) which produces the largest output (S_i) is selected. If the frequency of the second peak in the narrow-band signal is close to two times that of the first peak (first harmonic falls within the narrow-band signal), then no sub-harmonics will be generated.

4. RESULTS

4.1. Manual test

A test signal containing a female speech fragment was examined in the frequency domain (instead of using the *analyse*-block). At three points in the sample the settings of the harmonics generator were modified. No changes were made to the harmonics generator during unvoiced speech parts.

As a provisional solution to the loudness level differences between the original and the synthesised low part, a gain function α was derived from the original low part. For each fragment of 1024 samples the amplitude of the synthesised signal was made approximately equal to that of the original low part by multiplying the synthetic signal by α. This function α is low-pass filtered to prevent the result from being modulated. The result after adaptation of the loudness level sounded similar to adding the original low part.

4.2. Analysis of the results

Time analysis shows that the amplitude of the synthesised signal does not match that of the original low part. Because the harmonics generator is linear amplitude, the amplitude of the output is proportional to that of the input. As is shown in Figure 5, most of the energy of a speech signal is contained in the low frequencies. Hence the envelope of the narrow-band signal is almost defined by that of the lower frequencies within this narrow-band signal. In Figure 6 parts of equally scaled waveforms are displayed. The signal, with a duration of 1.13 s (with the same utterance for all panels) is taken from [4]. In these wave forms it can be seen that the synthetic low part does not look like the original.

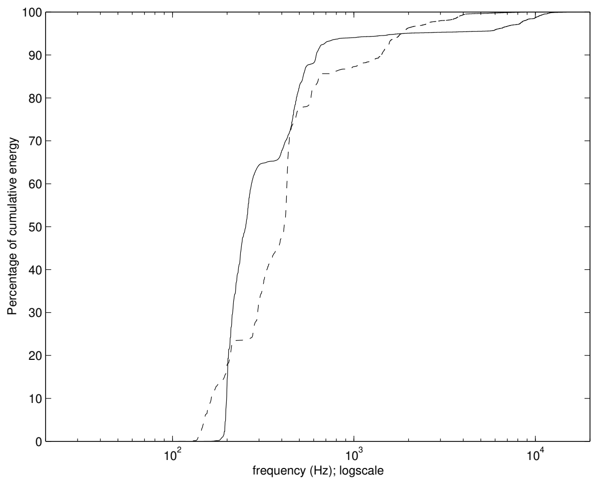

Figure 5: Cumulative energy in a male speech sample (dashed) and a female speech sample (solid). Signals are taken from [4].

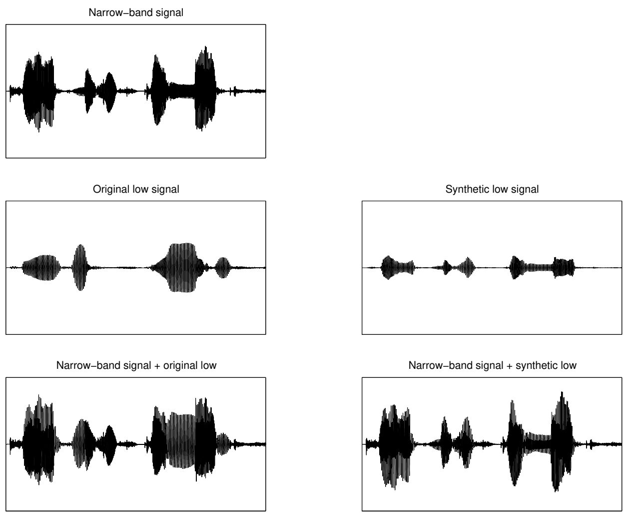

Figure 6: Various wave forms occurring in the system.

The decision signal used to perform the harmonics stream selection, is completely different from the signal constructed as described in Section 4.1. The decision signal shows oscillation. This may be caused by switching while there is unvoiced speech present at the input, although, when dealing with voiced speech sometimes the signal oscillates too.

In Figure 7 this decision problem is displayed. Here the blocks denote the possible frequency for the n-th harmonic, with n placed inside the block. At the left panel the original signals are displayed. The dashed lines denote the bandwidth of the input filters. In the right column the synthesised signals are displayed. Arrows indicate where the different components present within the input bandwidth cause subharmonics. As long as only one of the harmonics is present

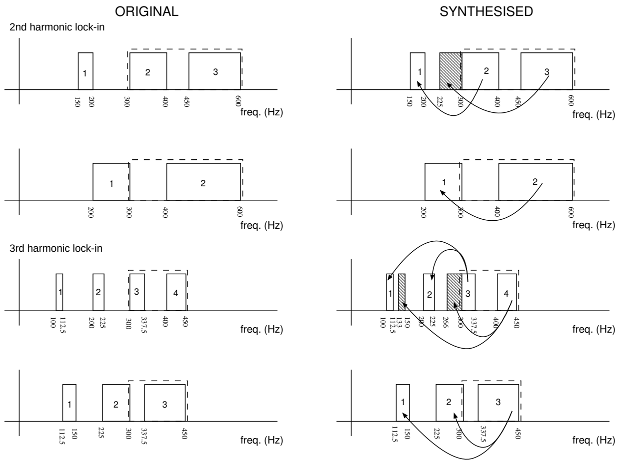

Figure 7: Schematic representation of input (left) and output (right) spectra. The solid blocks denote the harmonics, the dashed blocks denote the input filter's band-widths. As long as just one harmonic component is present in the filter's bandwidth, there is no wrong extension.

within the input bandwidth, the output is satisfying. But when there is (a part of) another harmonic present in this bandwidth this causes undesirable low frequency components which may form annoying artifacts.

4.3. Peak detection

Occasionally the frequency peaks detected as harmonic components within the narrow-band signal are harmonic 'disturbances'. They may be background sounds or noisy components of voiced consonants. As long as such disturbing components are below the threshold it is not detected, but this is not always true. Increasing the threshold is not always a suitable solution, because the higher harmonics which are needed for best fit decision could also be cancelled.

5. CONCLUSIONS AND RECOMMENDATIONS

The loudness levels of the synthesised and the original low part do not always match. The cause of this is that the synthesised low part is amplitude linear to the lowest frequencies within the narrow-band signal and not to the original low part.

When selecting the best fitting sub-harmonics manually, by examining the windowed (segmented) FFT, and adjusting the loudness level with use of the original low part, the result sounded very good. Although the synthesised part is not an exact reconstruction of the original, the extension yielded a much more natural sounding speech signal than that of the narrow-band input signal.

The method described generates (sub-) harmonics of the input signal. Hence it is impossible to extend unvoiced speech perfectly, because in this case one can not talk about pitch or harmonics. However extending unvoiced speech with sub-harmonics does not result in an audible annoying effect. This is unlike unharmonics in voiced speech, which will be noticed at once.

The tested algorithm used the idea that the set of (sub-) harmonics generated are correlated with the narrow-band signal, which of course only holds for voiced speech. If the set of higher harmonics generated by the non-linear element are also present in the narrow-band signal, then it will be very likely, that the sub-harmonics of that set match to the narrow-band signal. However sometimes the wrong harmonics signal is selected or the switch starts to oscillate. This for example happens with background sounds, voiced consonants, or unvoiced speech.

6. REFERENCES

[1] Masanobu Abe and Yuki Yoshida, *More Natural Sounding Voice Quality over the Telephone*, NTT, Review, vol. 7, May 1995, pp. 104-109.

[2] Aurelio Uncini, F. Gobbi and F. Piazza, *Signal Restoration of Broad Band Speech Using Nonlinear Processing*, ICASSP, 1999, pp. 997-1000.

[3] Erik Larsen and Ronald M. Aarts, *Perceiving low pitch through small loudspeakers*, Preprint of the 108th AES Convention Paris, France, 2000.

[4] European Broadcasting Union, *SQAM Compact Disc*, Polygram Hanover, Germany, 1988, tracks 49 and 54.

[5] J. Smeets, *Bandwidth extension of narrow-band speech*, M. Sc. Thesis, Eindhoven University of Technology, August, 1999.