Pitch-Synchronous Time-Scaling for High-Frequency Excitation Regeneration

João P. Cabral and Luís C. Oliveira

L^2F Spoken Language Systems Lab.
INESC-ID/IST,
Rua Alves Redol 9, 1000-029 Lisbon, Portugal
{jpcabral,lco}@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt

Abstract

The goal of bandwidth extension of speech (BWE) is to extrapolate the missing low or high frequency components of the wide-band speech (50-8000 Hz) based entirely on information contained in a narrow-band signal (300-3400 Hz). In this paper we propose a new method for high-frequency regeneration of the excitation signal, using the correlation between the shape of the glottal flow waveform and the spectrum of the voice source. The high-band excitation is generated by performing a pitch-synchronous time-scale (PSTS) transformation on the linear prediction narrow-band residual to generate an high-pass signal that retains the periodic characteristics of the original signal but with a larger open quotient. This method is easy to implement and does not introduce discontinuities in the spectrum of the regenerated excitation. It can be used in applications for BWE where no side information is transmitted or for low bit coding of wide-band speech.

1. Introduction

Wide-band speech signals are normally sampled at 16 kHz because almost all the energy of the speech signals is present below 8 kHz. In current analogue telephone networks the bandwidth is limited to about 300-3400 Hz and sampled at a rate of 8 kHz which represents a good compromise between the speech rate and the speech quality. Increasing the bandwidth of the narrow-band signals to 50-8000 Hz results in increased intelligibility and naturalness of speech and gives a feeling of transparent communication. For example, unvoiced speech has more spectral content at higher frequencies and filtering the important energy above 3.4 kHz affects speech intelligibility such as differentiation between fricatives.

The recently emerged end-to-end digital networks, such as the second and third generation wireless systems, ISDN, and voice over packet networks, enable the use of wide-band signals and can provide a speech quality exceeding that of public switched telephone network (PSTN). Algorithms for BWE can be used to transmit wide-band speech at lower bit rates or to regenerate the wide-band speech from the narrow-band speech with no increase in bit rate.

The speech waveform can be obtained by shaping a flat excitation, representing the glottal flow source, with the spectral envelope, that models the vocal tract transfer function. Numerous methods have been proposed for spectral envelope estimation in BWE but for the excitation regeneration less investigations can be found in literature.

In [1] it was demonstrated that the excitation of unvoiced signals is well modeled by white noise but this is a very poor model for the excitation of voiced speech. If voiced signals were truly periodic the spectrum of the flat excitation could be represented as harmonic peaks multiples of the pitch frequency with the same amplitude. Typically, this periodic structure is susceptible to variations over the spectrum, specially in the high-band, which results in a noise component. This irregularity in the spectrum constitutes a drawback to estimate the high-band harmonic structure of the excitation from the corresponding narrow-band component.

Spectral folding is a very popular technique for high-band excitation estimation and was first proposed in [2]. Generally, it consists of up-sampling the narrow-band residual by a factor of two. Another traditional method is to apply a non-linear distortion to the narrow-band excitation of voiced signals to generate harmonics in the missing high-band. A rectifier function is normally used for that purpose [3]. The more recent and relevant excitation regeneration methods are the noise modulation [4] an the sinusoidal transform coding (STC) [5]. In STC the excitation in the high-band is synthesized using the sinusoidal model and the degree of voicing controls a random component in the phase. This is the most complex model.

In this paper a new approach for high frequency excitation regeneration is proposed. A perceptual listening test was conducted to compare this new technique against the spectral folding and the noise modulation, using our own implementation of these methods.

2. High-frequency regeneration of the excitation

For our tests, the narrow-band signals were generated using wide-band speech signals that were low-pass filtered at 3.4 kHz and down-sampled to 8 khz. The coefficients of the synthesis filter were obtained from a 16^{th} order LPC analysis of the wide-band signal and an 12^{th} order LPC analysis of the narrow-band signal extracts a spectrally flat residual. The analysis and synthesis were performed using the overlap-and-add technique with frames of 20 ms length and 10 ms interval. Hamming windows were used in the analysis and Hanning windows were used in the synthesis. The spectral envelope of the unprocessed wide-band signal was also used for the synthesis.

2.1. Spectral folding (SFE)

The spectral folding technique can be easily implemented by adding one zero-valued sample between successive samples of the narrow-band residual $x(n)$. The resulting signal, $y(n)$, has a sample rate of two times the original rate:

$$y(n) = \begin{cases} x(n/2) & \text{n even} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The high-frequency spectrum of this modified signal is a mirrored image of the narrow-band spectrum. This method produces a gap in the spectrum of the wide-band estimated excitation around the Nyquist frequency $f_c = 4$ kHz due to the narrow-band filter response with cut-off frequency at 3.4 kHz. Also, it creates a discontinuity in the periodic structure of the excitation signal because the frequencies of the harmonics above 3.4 kHz are not multiples of the fundamental frequency f_0, unless f_c is a multiple of f_0.

2.2. Noise modulation (NME)

The noise modulation technique [4] generates the high-frequency components by adding white noise modulated by the time envelope of the rectified speech signal in the frequency range of 3-4 kHz. We used the time domain envelope of the 2-3.4 kHz rectified bandpass signal to modulate the white noise source in the band 3.4-8 kHz. The modulated signal was then limited to the frequency range of 3.4 to 8 kHz and scaled to match the energy of the narrow-band residual.

3. New method based on modification of the open quotient

The proposed new method for high-band excitation regeneration is based on time-scaling the open phase of the glottal source waveform to transform the open quotient (OQ) which is a time-domain parameter of the glottal flow signal [6]. Figure 1 (a) shows an example of the glottal flow waveform. Using the notations depicted in this figure the OQ can be expressed as:

$$OQ = (T_e + T_a)/T \quad (2)$$

where $T_e + T_a$ is equal to the duration of the open phase.

Changing the shape of the time-domain waveform of the glottal flow corresponds to change the power spectrum of the voice source. In case of the OQ, decreasing this parameter has the same effect in the duration of the return phase T_a and the duration of the peak flow T_e. The decrease of T_a changes the spectral tilt adding a -6dB/oct above the cut-off frequency F_a of a first-order low-pass filter [6]. The spectral effect of decreasing T_e is mainly to shift energy from low frequency harmonics to medium frequency harmonics but it also has some contribution in increasing the frequency F_a. As result, reducing the open quotient for each glottal cycle expands the frequency scale and generates the missing harmonics in the high-band. The effect in the lower part of the spectrum is to decrease the ratio between the amplitudes of the first and second harmonic [6].

In our first experiments, the duration of the return phase was reduced in order to verify its effect in the missing high-band. This produced harmonics but with a higher spectral decay. By varying the open quotient a more uniform energy distribution in the high-band was achieved. Another advantage of changing the open quotient is that the high-frequency spectrum is less periodic and more similar to the spectral structure of the original excitation signal in the 3.4 to 8 kHz frequency range.

3.1. Description of the method (PSTS)

The method uses a pitch-synchronous analysis where the glottal closure instants are computed from the narrow-band signal for voiced segments while for unvoiced speech the pitch marks are

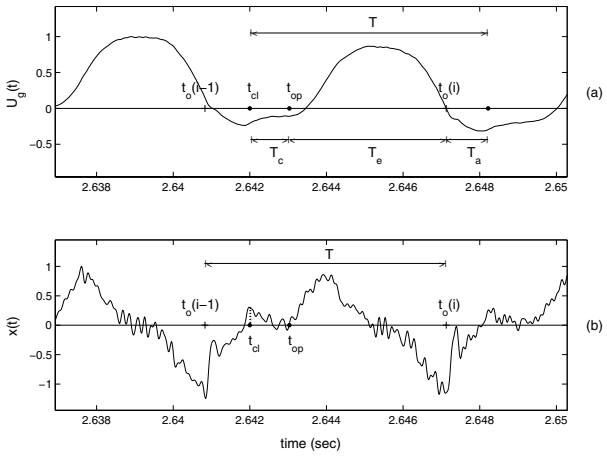

Figure 1: Representation of the time instants in the glottal flow waveform (a) and its time-derivative (b).

equally spaced 10 ms apart. Before computing the LPC model for each pitch period the narrow-band signal is high-pass filtered with a pre-emphasis filter ($\alpha = 0,97$) to attenuate the frequency ripple and noise of the residual. The LPC parameters are computed using a Hamming window centered in each pitch mark and with duration 20 ms. The residual signal is then obtained by inverse filtering the speech signal by a time varying all-zeros filter using the LPC coefficients calculated for each pitch mark. The residual signal is then segmented into short-time signals separated by the pitch marks. Thus, the length of the short-time signal $s_i(t)$ associated with the pitch mark $t_o(i)$ is equal to the estimated pitch period $T(i)$:

$$T(i) = t_o(i) - t_o(i-1) \quad (3)$$

In order to modify the OQ for voiced segments we extract two instants of the short-time signal $x_i(t)$, represented in Figure 1 (b). The glottal closure, t_{cl}, is estimated as the instant of the first peak after the first zero crossing. The glottal opening, t_{op}, is obtained using the threshold based method described in [7]. The duration of the return phase is estimated from t_{cl}:

$$T_a = t_{cl} \quad (4)$$

while the peak flow duration T_e and the close phase duration T_c are calculated by:

$$T_e = T - t_{op} \quad (5)$$

and

$$T_c = t_{op} - t_{cl} \quad (6)$$

Since the duration T_a in equation (4) corresponds to the return phase of the previous glottal cycle we assume that the duration of the return phase is approximately equal for neighbor short-time signals.

According to equation (2) the return phase and the peak flow segments are time-scaled by a factor L to decrease the open quotient ($L < 1$). In our implementation we used the factor $L = 0.6$ to decrease the open quotient in 40%. Before the scaling operation the signal is up-sampled to 16kHz and two times over-sampled to avoid aliasing due to the spectral extension.

After the time-scale transformation it is necessary to adjust the duration of the closed phase T_c to maintain the pitch constant. We choose to insert another signal in the middle point of

the closed phase ($t\prime_m$). To explain how we perform this operation, consider the two parts of the modified short-time signal $z_i(t)$ separated by $t\prime_m$:

$$z_l(t) = z_i(t), \quad 0 \leq t < t\prime_m \tag{7}$$

$$z_r(t) = z_i(t), \quad t\prime_m \leq t \leq T \tag{8}$$

The padding region will be inserted between $z_l(t)$ and $z_r(t)$. To obtain a smooth transition in the the new region we use a first-order linear window with varying length D:

$$w_D(t) = \begin{cases} t/D, & 0 \leq t \leq D \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

Then, two neighbor segments are obtained by windowing the original short-time signal $x_i(t)$ from the center of the closed phase t_m to the left and right side:

$$x_a(t) = x_i(t)\left(1 - w_{T_a}(t - t_m + T_a)\right) \tag{10}$$

$$x_b(t) = x_i(t) w_{T_b}(t - t_m) \tag{11}$$

Figure 2 (a) shows one short-time signal with normalized amplitude and the position of the weighting windows (dashed lines). The durations T_a and T_b, are chosen so that $T_a + T_b$ is the difference between the pitch period and the duration of $z_i(t)$. The durations of the windows are limited to a maximum length. If the sum of the durations is smaller than the pitch period than the windowed signal $x_b(t)$ is extended by concatenating a segment with the adequate length and zero value at the end of this signal. Finally, the two parts of the modified short-time signal and the two windowed signals are time-shifted by the adequate values to preserve the continuity of the energy envelope when the segments are added:

$$\begin{aligned} y_i(t) \quad = \quad & z_l(t) + x_b(t + t_m - t\prime_m) \\ & + x_a(t + t_m - t\prime_m - T_b) + z_r(t - T_a - T_b) \end{aligned} \tag{12}$$

Figure 2 (b) shows the short-time signal after time-scaling the open phase and the length adjustment.

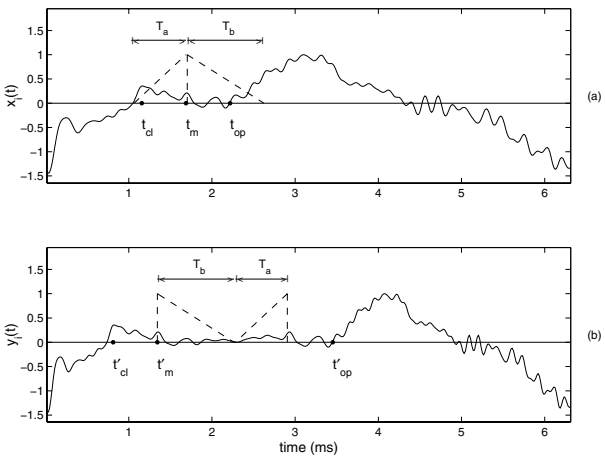

Figure 2: (a) One cycle of the glottal flow derivative waveform and the weighting windows that multiply the signal (b) Short-time signal with decreased open quotient.

Figure 3 represents the spectrum of a 20 ms segment of the narrow-band residual signal and the spectrum of the residual after time-scaling the open phase. It shows the effect of decreasing the open quotient in the amplitude of the first two harmonics (H1-H2 decreases) and in the expansion of the harmonic structure to the high-band.

In the unvoiced regions, the high-band excitation was modeled with white Gaussian noise. The noise variance was scaled by the energy of the narrow-band residual signal in order to match the energy of the original high-band excitation.

Since we are not interested in modifying the narrow-band signal, the modified excitation signal is synthesized with the true wide-band spectral envelope and it is high-pass filtered with cut-off frequency 3.4 kHz. Finally, the high-band component is added to the original narrow-band signal to obtain the wide-band signal. The high-frequency band of the resulting excitation signal is almost perfectly reconstructed and does not contain any discontinuities in the spectrum, as shown in Figure 4.

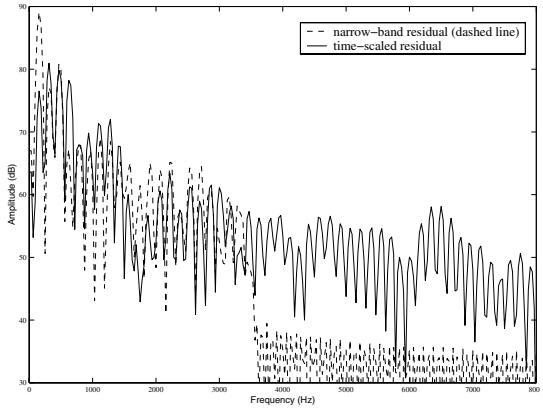

Figure 3: Spectrum the original narrow-band excitation and the spectrum of the modified excitation signal.

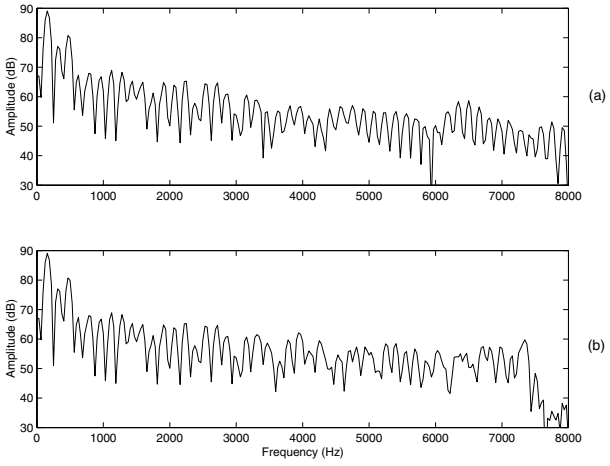

Figure 4: Spectrums of the regenerated wide-band excitation (a) and the original wide-band excitation signal (b).

4. Evaluation of the methods

In order to evaluate the performance of the excitation estimation methods described in the previous sections a MUSHRA (Multi Stimulus test with Hidden Reference and Anchor) listening test [8] was performed. Utterances of about 4 sec from two female and two male speakers were used. The listening panel was composed of 18 listeners. Subjects could directly compare all the test signals in each trial of four and grade them with a high degree of resolution (0 to 100). The set of impaired signals consisted of the original signal with bandwidth 300-8000 Hz (hidden reference), its low-pass filtered version with bandwidth 300-3400 Hz (hidden anchor) and the signals obtained by implementing the three methods under test. We used the high-pass filtered signal at 300 Hz as the reference instead of the unprocessed signal with bandwidth 50-8000 Hz because the tested methods only estimate the high-frequency excitation signal. The resulting mean scores and the corresponding 95% confidence intervals are shown in Table 1. The graphical representation of the results is presented in Figure 5.

Test Signals	Female		Male		Total	
Original	98.6	$\pm$1.6	96.8	$\pm$2.7	97.7	$\pm$1.6
PSTS	93.3	$\pm$3.2	86.9	$\pm$3.1	90.1	$\pm$2.3
SFE	80.2	$\pm$5.8	90.9	$\pm$4.4	85.5	$\pm$3.8
NME	80.8	$\pm$4.4	87.7	$\pm$4.1	84.2	$\pm$3.1
Narrowband	54.0	$\pm$6.0	52.0	$\pm$5.1	53.0	$\pm$3.9

Table 1: Results of the MUSHRA listening for the two female sentences, the two male sentences and all the sentences.

Bandwidth extended speech obtained with the three methods was significantly preferred over the narrow-band speech. Also, listeners scored higher the quality of the proposed method (PSTS) than the spectral folding and noise modulation methods. The enhancement of female speech obtained worse results than male speech for SFE and NME. This is due to the higher fundamental frequency and to a more regular harmonic structure of the high-frequency spectrum of the female voices when compared with male voices. So that, the metallic effect due to strong harmonics in the high-band for SFE and the roughness in NME were more perceptive in the case of female speech. PSTS technique outperformed clearly these methods for the female utterances. For the male utterances, listeners found, in general, the expanded and original signals difficult to distinguish.

We think the shortcoming of an increase in quality with the PSTS method is that it requires a robust pitch marking algorithm that can be a limitation for a real-time system.

In this study we used the true wide-band spectral envelope for the synthesis of the broadband speech, but further tests should be made to evaluate the performance of this method together with a spectral envelope estimation technique.

The audio signals used in our test are available at "http://www.l2f.inesc-id.pt/ jpcabral/psts_hfer".

5. Conclusions

In this paper, we proposed a new approach to generate the high-frequency excitation signal from the narrow-band signal. This technique consists in modifying the open quotient parameter of the glottal flow waveform using a pitch-synchronous time-scaling (PSTS) transformation of the linear prediction residual.

A subjective listening test was performed to evaluate the new method by comparing it with two traditional techniques for excitation regeneration. All methods produced speech with quality very close to the original and with clear preference over the narrow-band speech. The proposed method achieved the best performance in the test. The major drawback of the proposed technique is the dependence on a robust pitch marking algorithm.

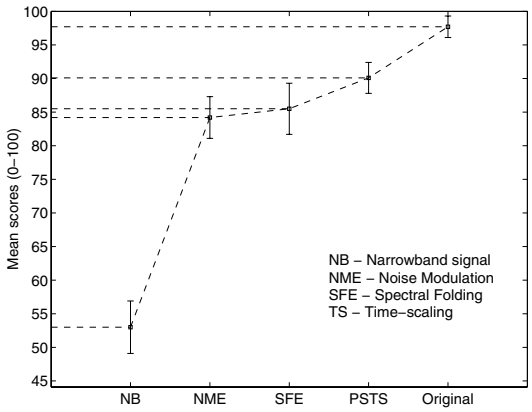

Figure 5: Graph showing the mean scores and the 95% confidence intervals obtained for the different systems from the MUSHRA listening test.

6. Acknowledgements

This work was partially founded by the Portuguese Foundation for Science and Technology (FCT) within the project POSI/SRI/41071/2001.

7. References

[1] Kubin, G., Atal, B. S. and Kleijn, W. B., "Performance of noise excitation for unvoiced speech", IEEE Workshop on Speech Coding for Telecommunications, pp. 35-36, 1993.

[2] Makhoul, J. and Berouti, M., "High-frequency regeneration in speech coding coding systems", Proc. IEEE Int. Conf. Acoust. Speech Sign. Process., pp. 428-431, 1979.

[3] Valin, J. M. and Lefebvre, R., "Bandwidth extension of narrowband speech for low bit-rate wideband coding", IEEE Workshop on Speech Coding, pp. 130-132, Delavan, USA, 2000.

[4] McCree, A., "A 14 kb/s wideband speech coder with a parametric highband model", Proc. IEEE Int.Conf. Acoust. Speech Sign. Process., pp. 1153-1156, 2000.

[5] Epps, J., Wideband extension of narrowband speech for enhancement and coding, Ph.D. thesis, University of New South Wales, Australia, 2000.

[6] Doval, B. and d'Alessandro, C., "Spectral Correlates of Glottal Waveform Models: an Analytic Study", ICASSP 97, pp. 1295-1299, Munich, 1997.

[7] Arroabarren, I. and Carlosena, A., "Glottal source parameterization: a comparative study", Proc. of the ITRW VOQUAL'03, pp. 29-34, Geneva, Switzerland, 2003.

[8] Recommendation ITU-R BS.1534, "Method for the subjective assessment of intermediate quality level of coding systems", Tech. Rep., ITU Radiocommunication, 2001.