INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

Real time pitch shifting with formant structure preservation using the phase vocoder

Michał Lenarczyk

IPI PAN – Institute of Computer Science of the Polish Academy of Sciences, Warsaw, Poland

m.lenarczyk@phd.ipipan.waw.pl

Abstract

Pitch shifting in speech is presented based on the use of the phase vocoder in combination with spectral whitening and envelope reconstruction, applied respectively before and after the transformation. A band preservation technique is introduced to contain quality degradation when downscaling the pitch. The transposition ratio is fixed in advance by selecting analysis and synthesis window sizes. Real time performance is demonstrated for window sizes having adequate factorization required by fast Fourier transformation.

Index Terms: voice conversion, nonparametric transformation, phase vocoder

1. Introduction

Pitch shifting in speech is a useful technique with applications in other speech technologies, for example voice conversion. It can be realized in several different ways. Simple interpolation by a defined ratio changes the time and frequency scale of the signal and thus affects both the pitch and the formant structure. A more advanced tool is the phase vocoder [1, 2], which allows to scale the frequency range while preserving the time scale. Scaling of the signal in frequency affects both the pitch and the formant positions. For practical use, the pitch must be transformed without affecting the spectral structure, i.e., the spectral envelope must be preserved.

PSOLA technique [3] could be used to accomplish this task. Another way is to use a parametric speech vocoder such as Harmonic/Noise Model (HNM) [4], to synthesise new speech signal with adequately modified pitch profile. These methods rely on precise estimation of the fundamental frequency or period and also on the segmentation of the signal by its voiced/unvoiced state. As is well known, detection of voiced state and the estimation of pitch in voiced segments are difficult tasks when real time regime is enforced, in which only a short time context is available for analysis. In consequence, these methods may perform well in an off-line setting, where the signal is pre-recorded and can be analyzed in an arbitrarily large time context, but not in real time.

This work concerns the use of the phase vocoder for real time pitch transposition with spectral envelope preservation. The phase vocoder is a nonparametric method of pitch transformation, since explicit estimation of either the voicing state of the fundamental frequency or period is not needed. It can thus operate on short signal windows. Indeed, real time implementations of the phase vocoder have been reported in the past [2], but not in a configuration with spectral envelope preservation, as demonstrated in this contribution.

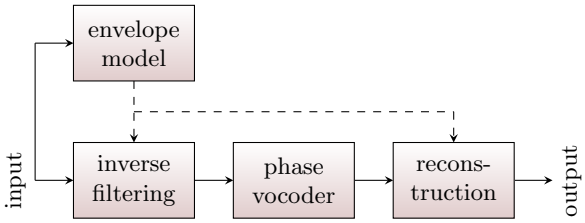

Figure 1: *Phase vocoder based pitch transposition with envelope preservation.*

2. Structure of the system

Pitch transposition with envelope preservation using the phase vocoder has been proposed in [5], where the spectral structure is retained by pre-warping the spectral envelope of the signal before feeding it into the phase vocoder. The shape of the transformation function is chosen so as to cancel the effect of the phase vocoder. A different approach is considered in [6], where the spectral envelope is first estimated and removed from the signal by inverse filtering, and restored as the final reconstruction stage in which phase vocoder transformed residual is used as input. This approach, illustrated in Fig. 1, is more flexible and is followed in this work.

2.1. Envelope modeling

Two methods of envelope estimation and inverse filtering are considered here. The first method employs linear predictive all-pole modelling of the envelope. Standard autocorrelation method of parameter estimation using Levinson-Durbin recursion was applied. To counteract the tendency to overfit formants to single harmonic conponents, spectral smoothing technique (SST) [7] was applied, and invertibility of the autocorrelation matrix was enforced by adding white noise component at -40 dB to boost the diagonal. Envelope whitening and restoration was done by simple filtering using, respectively, the moving average filter $A(z)$ and the autoregressive filter $1/A(z)$. The method is comupationally efficient and real time operation is easy to achieve.

In the second method, True Envelope estimation was applied [5]. Experiments have shown that this approach yields smoother envelope transitions between successive windows. The envelope is estimated, removed and finally restored using homomorphic analysis. The estimation is a recursive process which iteratively updates the amplitude spectrum and recalculates the envelope, and thus a computationally demanding task. The process can be considerably accelerated using the correction weighting procedure described in [5]. However, with the available implementation, it was only possible to reliably execute only 2 correction steps within real time, whereas typically

5-6 iterations are needed for full convergence. Despite this, the suboptimal envelope estimate turned out to be good enough for practical use.

2.2. Phase vocoder transformation

The input to the phase vocoder is the residual signal obtained in one of the two inverse filtering procedures described. The signal is essentially spectrally flat - except for the fine harmonic structure in voiced segments which is retained. Phase vocoder transformation thus outputs a signal which is also spectrally flat except for possible harmonic structure, which is scaled in frequency to match the desired pitch.

The phase vocoder changes pitch in two combined steps. The first step interpolates the amplitude and differential phase spectra to change the duration of the signal, without affecting the frequency content. This step is accompanied by a matching time-scale interpolation that compensates the time-scale modification, and causes the frequency content to be rescaled.

To achieve real time performance, the implementation of the phase vocoder used in this work is based on fast Fourier transformation, and both steps are performed in the frequency domain [8]. The time-scale interpolation is achieved by changing the DFT order from N_A (analysis window size) to N_S (synthesis window size). This involves either extension or truncation of the DFT coeefficients, depending on whether the pitch is downscaled or upscaled.

The chosen orders N_A, N_S should factor into small primes in order to minimize the computational cost of the FFT operation. As a result, real time pitch conversion can only be done for a limited subset of conversion ratios. In practice, the number of possible source-target FFT orders is sufficient to make it possible to closely approximate any desired pitch transposition ratio and meet the phase vocoder constraints.

2.3. Bandwidth preservation

An important advantage of the described framework in comparison with the method described in [5] is that the full frequency envelope information is retained. This fact is used here to improve voice quality.

When pitch is downscaled using the phase vocoder, the frequency range is shrunk, leaving the upper end of frequency band empty. Using the information about the spectral envelope, it is possible to artificially extend the signal's band. The simple solution, implemented in this work within the phase vocoder, is to fill the spectrum with white noise when the DFT order needs to be extended (which happens in the case of pitch downscaling). By setting the amplitude to the mean level of the residual spectrum and randomly generating the phase (with symmetry enforced to make the inverse DTF real), a transformed residual is obtained that retains its bandwidth. The restoration of the formant structure then produces speech which retains better quality compared to the case without bandwidth extension in which especially the fricatives are impaired. The value of the described bandwidth extension method was confirmed by formal listening tests [9].

3. Conclusions

In this demonstration, a system was presented for pitch shifting with envelope preservation able to work in real time. The proposed bandwidth extension method helps prevent quality breakdown that occurs when downscaling pitch.

The described solution can be used as a building block of real time voice conversion. This is the long-term goal pursued by the author, with encouraging preliminary results [9].

4. References

[1] J. Flanagan and R. Golden, "Phase vocoder," *Bell System Technical Journal, The*, vol. 45, no. 9, pp. 1493–1509, Nov 1966.

[2] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, May 1999.

[3] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, vol. 11, Apr 1986, pp. 2015–2018.

[4] Y. Stylianou, "Modeling speech based on harmonic plus noise models," in *Nonlinear Speech Modeling and Applications*, ser. Lecture Notes in Computer Science, G. Chollet, A. Esposito, M. Faundez-Zanuy, and M. Marinaro, Eds. Springer Berlin Heidelberg, 2005, vol. 3445, pp. 244–260.

[5] A. Roebel and X. Rodet, "Efficient Spectral Envelope Estimation and its application to pitch shifting and envelope preservation," in *International Conference on Digital Audio Effects*, Madrid, Spain, Sep. 2005, pp. 30–35, cote interne IRCAM: Roebel05b.

[6] P. N. Petkov and W. B. Kleijn, "Improving the phase vocoder approach to pitch-shifting," in *Proceedings of Interspeech*, Antwerp, Belgium, August 2007, pp. 1985–1988.

[7] Y. Tohkura, F. Itakura, and S. Hashimoto, "Spectral smoothing technique in parcor speech analysis-synthesis," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 6, pp. 587–596, Dec 1978.

[8] M. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 243–248, Jun 1976.

[9] M. Lenarczyk and A. Janicki, "Voice conversion with pitch alteration using phase vocoder," in *Proceedings of Interspeech (submitted)*, Stockholm, Sweden, August 2017.