Shape-invariant speech transformation with the phase vocoder

*Axel Röbel**

IRCAM - CNRS - STMS, France

axel.roebel@ircam.fr

Abstract

This paper proposes a new phase vocoder based method for shape invariant real-time modification of speech signals. The performance of the method with respect voiced and unvoiced signal components as well as some control strategies for the voiced/unvoiced balance of the transformed speech signals will be discussed. The algorithm has been compared in perceptual tests with an implementation of the PSOLA algorithm demonstrating a very satisfying performance. Due to the fact that the quality of transformed signals remains acceptable over a wide range of transformation parameters the algorithm is especially suited for real-time gender and age transformations.

Index Terms: shape invariant speech transformation, phase vocoder.

1. Introduction

The desire to modify speech signals such that the transformed signals keep a high degree of naturalness has triggered considerable research and development efforts. As a consequence there currently exist numerous algorithms that achieve high quality speech transformations. In many cases, however, high quality can be achieved only for limited transformations (e.g., time stretching and transposition factors in the range $[0.7, 1.4]$). In the present context we are interested to achieve real time speech signal transformation of age and gender characteristics, that is, a transformation of a man's voice into a woman's or into a girl's voice. Here pitch changes of a factor 2-3 as well as manipulation of the perceived vocal tract characteristics are generally necessary. In the following we will discuss a new algorithm that achieves comparatively high quality transformation for a wide range of transformation parameters and is therefore well suited for the transformations mentioned above.

Conceptually relatively simple approaches to speech transformation are the time domain algorithms "Synchronized Overlap-Add" or (SOLA) [1] and the "Pitch Synchronous Overlap-Add" (PSOLA) algorithm [2]. These algorithms can operate in real time. The first provides only time scale modifications and has to be combined with a re-sampling operation if transposition is required. The PSOLA algorithm on the other hand provides time and frequency scale modifications. Both require additional means for vocal tract filter (VCF) modification.

If VCF modifications are to be considered spectral signal representations are beneficial. An important model for spectral domain speech representation is the shape invariant sinusoidal model introduced in [3]. Recent variants are the "Harmonic and Noise Model" (HNM) [4] and the "Quasi Harmonic Model" (QHN) [5]. A phase vocoder based algorithm following very closely the algorithm proposed in [3] has been presented in [6]. The analysis/synthesis system TANDEM-STRAIGHT [7] is a recent improvement of the STRAIGHT system; however, it is not yet clear whether TANDEM-STRAIGHT will allow real time operation [8]. An inconvenience with these algorithms is the fact that they depend on precise description of the fundamental frequency and/or pitch markers (glottal closure instantants) to achieve high quality transformation.

In the present article we will present an augmented phase vocoder based SOLA algorithm that achieves high quality signal transformations for time and pitch scale manipulation for a wide range of scaling factors without requiring the fundamental frequency and/or pitch marks to be known. The proposed algorithm has been evaluated in a number of subjective listening tests that will be summarized below. It has been implemented in a real time speech transformation system [9] that is frequently used for composition and artistic sound manipulation often requiring extreme sound transformations. Using the spectral envelope estimator described in [10] the proposed algorithm achieves high quality gender transformation notably for transformations requiring pitch shifting upwards (e.g., man→woman) that in many cases have been evaluated to be indistinguishable from natural signals. The system performs real time sound transformation using only 10-20% of the CPU time of desktop computers (1.7GHz Pentium M) when using mono 44.1kHz speech signals. The latency of the algorithm is related to the fact that at least one analysis window needs to be present for analysis before the algorithm can start working.

The following article is organized as follows. In section **2** we introduce the shape invariant phase vocoder (SHIP) algorithm, and in section **3** we will discuss some properties of the algorithm using a simple sound example as well as the results of the perceptual evaluation of the algorithm. Finally, in section **4** we will present a summary and a short outlook.

2. Shape invariant processing in a modified phase vocoder

The standard phase vocoder performs signal transformation by means of modifying and moving the spectral frames of an short time Fourier transform (STFT) analysis of the sound to be transformed [11, 12]. The discrete Fourier transform (DFT) sequence representing the STFT of the input signal $x(n)$ using a length M analysis window $w(n)$ that is centered around the origin is given by

$$X_l(k) = \sum_n x(n)w(n - C_l)e^{\frac{-j2\pi kn}{N}}. \qquad (1)$$

Here $N \geq M$ is the DFT size and C_l is the window center for frame l. During transformation the spectral frames X_l are modified in content and position [12, 13] yielding output DFT sequence $\widetilde{X_l}$ that is then synthesized using overlap-add.

Whenever the STFT frames are time-shifted, which means that the synthesis frame position C'_l is different from C_l, the phases of the STFT have to be adapted to achieve coherent overlap-add of the sinusoidal components. Within the phase vocoder this phase adaptation (horizontal phase synchronization) is based on the observed phase evolution in all the bins of

*This work was supported in part by the french FEDER project Respoken

the original signal frames. Phases at position $C'(l)$ are obtained from phases at position $C'(l-1)$ as follows

$$I = C_l - C_{l-1} \quad (2)$$

$$\Theta_l(k) = \frac{[\arg(X_l(k)) - \arg(X_{l-1}(k)) - I\frac{2\pi k}{N}]_{2\pi}}{I} \quad (3)$$

$$\widetilde{\Phi_l(k)} = \widetilde{\Phi_{l-1}(k)} + (\Theta_l(k) + \frac{2\pi k}{N})(C'_l - C'_{l-1}). \quad (4)$$

Here Θ_l is the frequency difference between the center frequency at bin k that is obtained using the principal value $[]_{2\pi}$ of the observed and nominal expected phase in frame l. $\widetilde{\Phi_l(k)}$ is the phase in $X_l(k)$.

The phase update in the phase vocoder does not take into account the phase relations between the different sinusoids (vertical phase synchronization) and therefore, especially due to the recursive eq. (4), frequency estimation errors will result in a desynchronization of the different sinusoidal components. Note that this problems persists even if the state-of-the-art intra-sinusoidal phase synchronization method [12] is used. While the vertical de-synchronization of the sinusoidal components is perceptually uncritical for most musical signals, for speech signals it affects the perception of the underlying excitation pulses, and leads to an artifact that is generally described as missing clarity (phasiness) of the transformed voice. Following the terminology proposed in [3] we will denote the action of a transformation algorithm that preserves these inter-partial phase relations as *shape invariant processing*.

For the general case of polyphonic or in-harmonic sounds the horizontal (inter-frame) phase synchronization requires that frequencies of the different sinusoidal components are integrated over time as shown in eq. (4). For the special case of harmonic and monophonic sound signals phase relations are essentially periodic such that $\widetilde{\Phi_l(k)}$ can be obtained from the phases of the current unmodified frame according to

$$\widetilde{\Phi_l(k)} = \Phi_l(k) + \Omega_l(k)\Delta_n, \quad (5)$$

using a time shift Δ_n that will be determined below. Because Δ_n is generally very small (smaller than half the fundamental period of the sound segment under operation) and because the recursive structure of eq. (4) is avoided the vertical inter-partial phase synchronization is always maintained such that shape invariant processing is achieved.

2.1. Estimation of the optimal time shift

A coherent calculation of the optimal time shift can be obtained from the cross-correlation between the last synthesis frame $\widetilde{X_{l-1}(k)}$ and the current unmodified synthesis frame $X_l(k)$ that has been placed in position C'_l determined by the desired time stretching factor. Cross-correlation has been used as well to derive optimal placement in the time domain SOLA algorithm [1]. The phase vocoder based SOLA, however, does not require any adaptation of the position of the synthesis frame such that no compromise of the desired local time stretching factor has to be made. Moreover, we can constrain the cross-correlation to use only sinusoidal signal components such that the impact of the signal background noise during the estimation of the optimal overlap position is attenuated. This can be achieved by means of a spectral mask $S_l(k)$ retaining only spectral bins that constitute the spectral peaks related to sinusoidal signal components. Here we use a computationally efficient algorithm to establish the sinusoidal mask according to [14].

The cross-correlation sequence can be calculated in the spectral domain if $N \geq 2M$. If this condition is not fulfilled the complex signal spectrum can be interpolated prior to masking to double the DFT size N. Note that the spectral domain interpolation can be limited to the frequency range containing sinusoidal components and, therefore, its costs compared to the complete processing costs are relatively small even if a precise interpolation is performed.

For $N \geq 2M$ the cross-correlation sequence for the sinusoidal components denoted as $Z(n)$ and is given by

$$Z(n) = \sum_{k=0}^{N}((X_l(k)^* S_l(k)\widetilde{X_{l-1}}(k)S_{l-1}(k))e^{j\frac{kn}{N}2\pi}. \quad (6)$$

Here $X_l(k)^*$ represents the conjugate complex of $X_l(k)$. We note that the signal noise is masked by means of the spectral masks S such that the impact of the noise on the estimation of the optimal delay parameter is significantly reduced.

Under the assumption of a quasi stationary harmonic signal component and an analysis window $w(n)$ it can be shown that the cross-correlation sequence $Z(n)$ is locally quasi periodic with an amplitude envelope that follows approximately the auto correlation sequence of the analysis window. The objective is to find the maximum of the underlying periodic structure of the cross-correlation sequence removing as much as possible the effect of the analysis window. For a given frame offset between the synthesis frames $O_l = C'_l - C'_{l-1}$ the preferred time delay between the successive frames would be O_l. For this time delay we do not have to modify the phases of the synthesis frame because the synthesis frame will be placed at that position anyway. For modified signals we would like the delay to be as close as possible to O_l such that the changes to be applied to the phases of the synthesis frames are as small as possible. Accordingly, if P is the length of the signal period at the center of the current synthesis frame X_l we would like the time shift to stay within $O_l \pm P/2$.

If we denote the auto-correlation sequence of the analysis window as $Z_w(n)$ we can estimate the optimal time shift following the constraints discussed above by means of

$$N(n) = \max(Z_w(n), Z_w(D)) \quad (7)$$

$$Z'(n) = Z(n)/N(n) \quad (8)$$

$$T_l = \arg\max_n(Z'(n)N(n - O_l)) \quad (9)$$

The sequence $N(n)$ represents a normalization sequence that compensates the effect of $Z_w(n)$ on the cross-correlation sequence. This compensation should not be applied to the extreme ends of the $Z(n)$ because with only very few samples the local correlation may be very large without being significant. The $\max$ operation limits the compensation to the range that contains sufficient samples to prevent degeneration of the compensated cross-correlation. A sensible value of D can be derived from the constraint on the delay $T_l \leq O_l + P/2$. The sequence $Z'(n)$ represents the cross-correlation sequence after compensation of the systematic impact of the analysis window by means of $N(n)$. The optimal time delay T_l is given by the maximum of $Z'(n)N(n - O_l)$. The multiplication by $N(n-O_l)$ favors small time shifts with respect to the preferred time shift O_l. From T_l we get $\Delta_n = T_l - O_l$. The bias introduced by $N(n - O_l)$ can be removed by means of re-adjusting T_l to the local maximum in $Z'(n)$ that is closest to T_l.

A major advantage of the procedure is the fact that there is no need to know the fundamental frequency or the pulse positions of the signal to achieve synchronous overlap-add. A misclassification of sinusoidal components is uncritical as long as the maximum common divisor of the partial numbers of the detected sinusoidal components is 1.

To achieve pitch shifting with timbre preservation the algorithm described above has been combined with a resampling stage. The spectral envelope estimated according to [10] is prewarped prior to resampling.

2.2. Phase adaptation for aperiodic components

The procedure described so far achieves the synchronization of the sinusoidal components. For the correct treatment of the aperiodic signal components a number of additional comments are necessary. In the following we will discuss three classes of aperiodic signal components: transients, quasi-stationary noise (e.g., in fricatives, aspiration or whispered speech) and modulated noise (e.g., in voiced fricatives or breathy vowels).

Transient signal components can be handled without further means and with high quality according to [13]. Completely unvoiced signal segments are generally composed of quasi-stationary noise components. These segments do not require any specific shape invariant processing and can be treated with the standard phase vocoder algorithm. The modulated noises that are present in voiced signal components are considered perceptually important [4]. The modulation is synchronized with the glottal pulses and accordingly the delay estimated for maximizing the cross-correlation of the sinusoidal components will at the same time be a good candidate to align the envelope of the modulated noise. The amplitude modulation of the noise component will introduce an interdependency (correlation) between the phases of the spectrum at distant bins. Because these interdependencies will be reflected in the phase update equation eq. (5) the characteristic interdependencies in the phase spectrum will be preserved at least partly. As an example we display in fig. 1 the modulated noise in a speech signal that is obtained by band-stop filtering frequencies up to 5kHz. The comparison with the signal obtained from the time stretched version using a stretch factor 2 clearly shows that the noise components will indeed preserve a considerable part of the amplitude modulation.

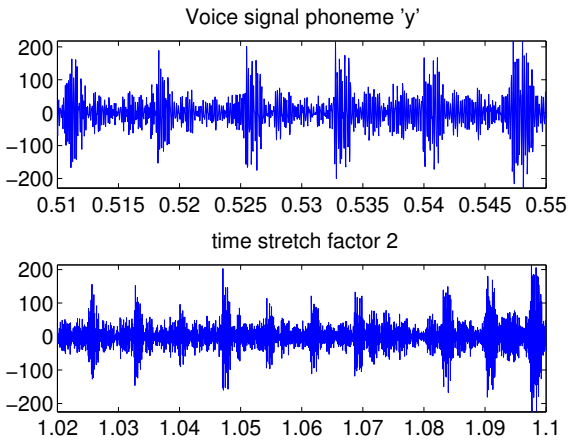

Figure 1: *high frequency ($>$ 5kHz) modulated noise of the original and time stretched speech signal containing phoneme 'y'*

2.3. Balancing unvoiced/voiced signal components

An important problem for speech transformation algorithms is the fact that signal transformations may transform signal components producing unnatural results. For the phase vocoder this problem is mostly related to the fact that time stretching of unvoiced signal components may change the perception of these components due to their increased time stability. Moreover we note that during transposition with preservation of the spectral envelope the change of the spectral distribution of the unvoiced and voiced parts of the excitation signal is often perceived as unnatural. Transposition downwards moves unvoiced components of the excitation signal into the formant regions, which is often perceived as very disturbing. Transposition upwards moves relatively clean sinusoids into frequency regions normally containing modulated noise. This is perceived as an unnatural metallic voice quality by human listeners. To counter these effects we would need signal processing strategies to randomize sinusoids and to transform unvoiced signal components into sinusoids.

A requirement for a solution of the problem is the detection of the frequency regions containing voiced and unvoiced signal components. The algorithm uses a simple two band model with time varying frequency boundary (voiced/unvoiced frequency boundary *VUF*) separating the 2 regions. There exist many approaches to estimate the VUF (e.g., [4]). In our case the procedure is simple because we can rely on the fact that we already detect sinusoidal components. To robustly detect the VUF we simply divide the spectrum into bands and compare the relative amount of sinusoidal energy per band to a threshold. The threshold depends on the bandwidth, the analysis window and other things, but can easily be derived experimentally using the sinusoidal peak detector on a pure white noise signal. Using the VUF as control there are two mechanisms that are used to readjust signal components:

Phase randomization: To preserve unvoiced signal components during time stretching we add a random uniformly distributed phase offset ($|\Delta p| < \alpha\pi$) to the phase of all spectral peaks above the VUF whenever the effective time stretching factor is larger than 1.1. This effectively destroys undesired sinusoidal components in the unvoiced frequency region. Note, however, that phase randomization is generally destroying the pitch synchronous modulation of the unvoiced signal components. Therefore, phase randomization needs to be used carefully. In the present implementation α is linearly increasing with frequency following

$$\alpha(f) = 0.3 + (f - VUF)/(20000Hz - VUF) \qquad (10)$$

where f and VUF are given in Hz.

SNR control: During transposition with preservation of the VCF the excitation signal is used in frequency regions that require a different balance between unvoiced and voiced signal energy. To adapt this balance we remix the sinusoidal and residual signal components. The residual is created in the spectral domain using the sinusoidal parameters (amplitude, phase, frequency and frequency slope [15]) that are estimated on the fly for each sinusoidal peak below the VUF. The sinusoidal peaks are then obtained from these parameters using precalculated sinusoidal peaks with normalized amplitude, phase and frequency and a grid of frequency slopes. The remixing is done for all sinusoidal peaks below the VUF. In the experiments discussed below we re-adjust the sinusoidal/residual balance by means of applying a constant factor γ to the residual part. For transposition down we use $\gamma_d = 0.1$ and for transposition up we use $\gamma_u = 1.4$. Note that γ is a control parameter that can be used to tune the voice characteristics and therefore there does not exist a unique correct setting.

3. Evaluation and discussion of results

The proposed shape invariant phase vocoder (SHIP) algorithm has been evaluated in a number of subjective tests comparing it to implementations of the PSOLA, HNM and STRAIGHT algorithms. In all tests HNM and STRAIGHT performed significantly worse than PSOLA and so, due to space constraints,

we will discuss only the results of the SHIP and PSOLA algorithms. The test covered transposition with timbre preservation (transposition factors 0.5 and 2) as well as time scaling (scaling factors 0.5 and 2). While exact timbre preservation is not desired for gender and age transformations we use this setting to be able to compare to the baseline PSOLA algorithm that is considered as quality reference. The extreme time stretching and compression will certainly not be used for the intended transformations; however, time stretching and compression is the fundamental operation of the SHIP algorithm that is used internally even if no time scaling is required. Therefore, we were interested to evaluate the quality of this transformation.

The transformed signals using originals from a male and female speaker have been evaluated by 19 individuals with and without professional background in sound processing. The individuals were asked to evaluate the degradation of the transformed sound signals on a 5 level scale containing the levels: 5 *perfectly natural*, 4 *minor artifacts*, 3 *small artifacts*, 2 *annoying artifacts*, 1 *inacceptable*. In fig. 2 we display the MOS level differences between the SHIP and PSOLA algorithms.

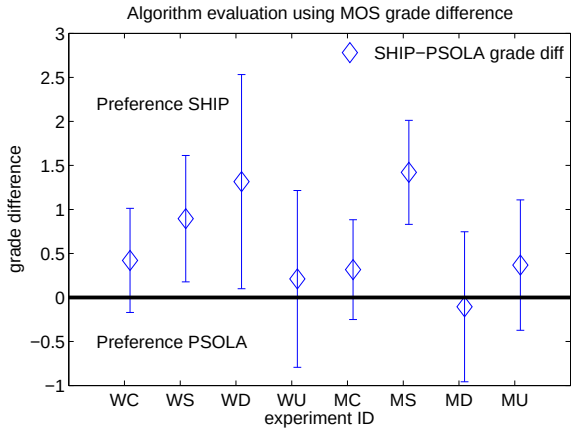

Figure 2: *Perceptive test results comparing PSOLA and SHIP algorithms described in the text. Given are the average MOS level differences and the standard deviation of the difference for transformation of woman 'W' and man 'M' signals using transposition up 'U' and down 'D' as well as time stretch 'S' and compression 'C'. Positive values indicate preference for SHIP.*

The results demonstrate that in most situations the SHIP algorithm performs considerably better than PSOLA. The improvement is significant especially for time scaling. SHIP achieves significant improvements when transposing the woman's voice down and when transposing the man's voice up. For the transposition of the woman's voice up or the man's voice down both algorithms are evaluted to have similar performance. This is partly due to the fact that for these 2 cases the transposition with timbre preservation is perceived to produce unnatural results anyway. We have to add that especially when transposing the man's voice down both algorithms clearly create artifacts. PSOLA creates a mechanical voice and SHIP a voice that is perceived as to have weakly synchronized excitation pulses. This is due to the fact that completely unvoiced excitation energy is moved into the voiced frequency region a situation that cannot be handled by the SNR adjustment.

4. Summary and Outlook

The present paper presents a new approach to shape invariant signal processing using a modified phase vocoder algorithm. The proposed algorithm can be understood as an implementation of the SOLA algorithm that uses the phase vocoder algorithm to achieve phase alignment. Compared to other speech transformation algorithms, the proposed algorithm shares the advantage of the SOLA system that it does not require an elaborate pre-analysis (pitch marks, F0). The algorithm is based on an inexpensive classification of sinusoidal and noise peaks that can be performed on the fly directly in the DFT frames [14].

Subjective evaluation has shown that the SHIP algorithm can significantly improve signal quality for extreme transformations. The main problem that is currently present in the SHIP algorithm is the fact that during transposition the characteristics of the excitation signal that is used to excite the formants may change, which can have a severe impact on the quality of the transformed speech. A signal operator that allows to change noise excitation energy into sinusoidal excitation is currently under investigation.

5. References

[1] S. Roucos and A. Wilgus, "High quality time-scale modification for speech," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1985, pp. 493–496.

[2] F. J. Charpentier and M. G. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1986, pp. 2015–2018.

[3] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 497–510, 1992.

[4] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[5] Y. Pantazis, O. Rosec, and Y. Stylianou, "On the properties of a time-varying quasi-harmonic model of speech," in *Proc. Interspeech*, 2008, pp. 1044–1047.

[6] J. Laroche, "Frequency-domain techniques for high-quality voice modification," in *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*, 2003.

[7] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proc. Conf. on ASSP (ICASSP'08)*, 2008, pp. 3933–3936.

[8] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara, "Implementatioin of realtime STRAIGHT speech manipulation system: Report on its first implementation," *Acoustic Science and Technology*, vol. 28, no. 3, pp. 140–146, 2007.

[9] IRCAM, "SuperVP-TRaX," `http://forumnet.ircam.fr/373.html?L=1`.

[10] A. Röbel, F. Villavicencio, and X. Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," *Pattern Recognition Letters, Special issue on Advances in Pattern Recognition for Speech and Audio Processing*, pp. 1343–1350, 2007.

[11] M.-H. Serra, *Musical signal processing*, chapter Introducing the phase vocoder, pp. 31–91, Studies on New Music Research. Swets & Zeitlinger B. V., 1997.

[12] 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, 1999.

[13] A. Röbel, "A new approach to transient processing in the phase vocoder," in *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*, 2003, pp. 344–349.

[14] M. Zivanovic, A. Röbel, and X. Rodet, "A new approach to spectral peak classification," in *Proc. of the 12th European Signal Processing Conference (EUSIPCO)*, 2004, pp. 1277–1280.

[15] M. Abe and J. O. Smith, "AM/FM rate estimation for time-varying sinusoidal modeling," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2005, pp. 201–204 (Vol. III).