A Shape-Invariant Phase Vocoder for Speech Transformation

Axel Roebel

A SHAPE-INVARIANT PHASE VOCODER FOR SPEECH TRANSFORMATION

*A. Röbel**

IRCAM-CNRS-STMS,
Analysis-Synthesis team,
Paris, France
axel(dot)roebel(at)ircam(dot)fr

ABSTRACT

This paper proposes a new method for shape invariant real-time modification of speech signals. The method can be understood as a frequency domain SOLA algorithm that is using the phase vocoder algorithm for phase synchronization. Compared to time domain SOLA the new implementation provides improved time synchronization during overlap add and improved quality of the noise components of the transformed speech signals. The algorithm has been compared in two perceptual tests with recent implementations of PSOLA and HNM algorithms demonstrating a very satisfying performance. Due to the fact that the quality of transformed signals stays constant over a wide range of transformation parameters the algorithm is well suited for real-time gender and age transformations.

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. Most algorithms, however, will degrade considerably if important transformations are requested. In our studies we found that time stretching or transposition of more than the factor 1.5 will often create annoying artifacts. While signal transformation by a factor of 1.5 is sufficient in many applications there are interesting applications that require transformation factors of more than 2. As an example we note gender and age transformations, e.g. transformation of a man's voice into a woman's or into a girl's voice, that require the pitch to change by factors of 2-3. 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. Because no critical pre-analysis is required, the algorithm works very well in real-time.

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

Many approaches to speech transformation are working in time domain using time domain signal models that represent the signal as a sequence of time domain pulses and noise. These are notably the "Synchronized Overlap-Add" or (SOLA) algorithm [1], the "Waveform Similarity Overlap-Add" or WSOLA algorithm [2] and the "Pitch Synchronous Overlap-Add" (PSOLA) algorithm [3]. The first 2 algorithms provide only time scale modifications and have to be combined with a re-sampling operation if transposition is required. They can easily be operated in real time. An important drawback is the fact that there is no distinction between noise and sinusoidal components during the overlap-add synchronization leading to relatively bad quality of noise components. The PSOLA algorithm on the other hand provides time and frequency scale modifications. The need to extract the individual excitation pulses, however, requires a robust pitch marking algorithm which for real-time operation is not straight-forward and imposes at least additional latency. The PSOLA algorithm will generally extract two pitch periods centered at the pitch marks and the segments are expected to represent the glottal pulse having passed through the vocal tract. Due to the decreased time span that is available for the representation of the vocal tract features the quality of PSOLA will generally degrade with increasing pitch.

On the other hand there are algorithms that are based on a sinusoidal signal representation. These have been influenced notably by the shape invariant sinusoidal model introduced in [4]. Recent variants are the "Harmonic + Noise Model" (HNM) [5, 6] as well as a phase vocoder based algorithm [7] or the wide band harmonic model [8]. Because these algorithms work in the frequency domain they allow advanced frequency domain transformations as for example the shifting and scaling of individual formats. An inconvenience is the fact that all these algorithms require the fundamental frequency and/or pitch marks to be known. Leading to increased latency and reduced robustness of the algorithms. Another frequency domain algorithm that is using pitch adaptive analysis windows is the STRAIGHT system [9]. The original version of the algorithm does not allow real time signal transformation, however, a real time version

of the algorithm has been presented recently [10]. Unfortunately the real time STRAIGHT system has been reported to entail reduced synthesis quality [10] and therefore we did not consider it for our real time speech transformation system.

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 in a rather wide range of scaling factors without requiring the fundamental frequency and/or pitch marks to be known. Accordingly, in conjunction with recent algorithms for spectral envelope estimation [11] the proposed algorithm allows us to achieve high quality gender transformation with reasonable latency.

The proposed algorithm has been evaluated in two subjective listing test that will be described below. It has been implemented in a real time speech transformation system that achieves good 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 algorithm has been implemented in form of a C++ library that performs real time sound transformation using only 10-20% of the CPU time of recent desktop computers 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 total latency is in the order of 6-8 periods of the minimum fundamental frequency of the input signal (1.5-2 analysis windows).

The following article is organized as follows. In section **2** we will briefly resume the basic SOLA algorithm, in section **3** we introduce the phase vocoder based SOLA algorithm, and in section **4** 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 **5** we will present a summary and a short outlook.

2. SOLA SPEECH TRANSFORMATION

The basic idea of SOLA algorithm is to cut the signal into overlapping frames and to displace the frames to achieve the desired time scaling. To avoid destructive overlap-add of displaced frames the frame positions are constrained to positions that maximize cross correlation between successive frames [1]. A consequence of the constraint is the fact that the effective time stretching factor will normally differ from the value that has been requested. Due to the fact that the signal cross correlation is done in the time domain, the voiced and unvoiced components will both influence the placement of the consecutive frames. Especially for analysis frames containing a rather weak voiced signal component the result suffers from sub optimal frame placement. Moreover, the unvoiced signal components in voiced signal frames will superimpose incoherently and therefore these components will be subject to cancellation.

3. 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 STFT analysis of the sound to be transformed [12, 13, 14]. The DFT sequence representing the STFT of the input signal $x(n)$ using the 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 that should be selected according to the transformation to be performed as explained e.g. in [12, 13, 14]. During transformation the spectral frames X_l are modified in content and position [15, 14, 16] yielding output DFT sequence Y_l. If the modified position of the frames is given by $C'(l)$ the resynthesis operation can be represented as follows

$$y_l(n) = \sum_{k=0}^{N} Y_l(k)e^{\frac{j2\pi kn}{N}}, \qquad (2)$$

$$y(n) = \frac{\sum_n w(n - C'(l))y_l(n - C'(l))}{\sum_n w^2(n - C'(l))}. \qquad (3)$$

Note that eq. (3) ensures optimal signal reconstruction in a mean squared error sense [17] even if there does not exist any signal $y(n)$ that produces an STFT $Y_l(k)$.

Whenever the STFT frames are time-shifted, which means $(C(l) \neq C'(l))$, the phases of the STFT have to be adapted to achieve coherent overlap add. Within the phase vocoder this phase adaptation is based on the observed phase evolution (frequency) in all the bins of the original signal frames as follows

$$I = C_l - C_{l-1} \qquad (4)$$

$$\Theta_l(k) = \frac{[\arg(X_l(k)) - \arg(X_{l-1}(k)) - I\frac{2\pi k}{N}]_{2\pi}}{I} \qquad (5)$$

$$\widetilde{\Phi_l(k)} = \widetilde{\Phi_{l-1}(k)} + (\Theta_l(k) + \frac{2\pi k}{N})(C'_l - C'_{l-1}). \qquad (6)$$

Here Θ_l is the frequency difference between the center frequency at bin k that is obtained using the principle value $[]_{2\pi}$ of the observed and nominal expected phase in frame l. $\widetilde{\Phi_l(k)}$ is the phase off the spectral frames after the phase update. Those frames will in the following be denoted as $\widetilde{Y_l(k)}$.

The original phase vocoder suffers from phase desynchronization of the individual bins related to a single sinusoidal peak which is due to frequency estimation errors on one hand and systematic errors of the phase evolution model that is used to modify the phases. The problem is addressed by the intra sinusoidal phase synchronization method presented in [15]. This method ensures phase synchronous modification of all bins within a single sinusoidal component and most of the time ensures a very high quality of the transformed signals. Note that the phase update in the phase vocoder shifts the sinusoidal phase keeping the amplitude envelope of the sinusoid - and notably also the analysis window - in its original position.

The phase update in the phase vocoder does not take into account the phase relations between the different sinusoids and therefore the frequency estimation errors will result in a desynchronization of the different sinusoidal components even if the intra sinusoidal phase synchronization method is used. The vertical desynchronization of the sinusoidal components is perceptually uncritical for most musical signals such that the phase vocoder algorithm (including all known enhancements) works well for nearly all musical signals.
For speech signals however, the vertical phase desynchronization between sinusoidal components affects the perception of the underlying excitation pulses and leads to an artifact that is generally described as missing clarity (or phasiness) of the transformed voice. Following the terminology proposed in [4] we will denote the action of a transformation algorithm that preserves these inter partial phase relations as *shape invariant processing*.

The desynchronization of the different sinusoidal components in the phase vocoder is due essentially to the same reason that also provokes the individual bins of a sinusoid to desynchronize if the intra peak phase synchronization is not especially enforced. For the general case of polyphonic or in-harmonic sounds the establishment of the phase coherence of subsequent frames (horizontal phase synchronization) requires that the frequencies of the different sinusoidal components are integrated over time as shown in eq. (6). Because the frequency of the individual bins is estimated independently the frequencies of the different partials will not be perfectly harmonic and the integration of the frequency estimation errors leads to the desynchronization of the sinusoidal components.

For harmonic and monophonic sound signals, however, there exists an alternative means to establish the (horizontal) phase synchronization of subsequent frames without destroying the (vertical) phase synchronization between the individual sinusoidal components. This alternate method uses the basic idea of the SOLA algorithm that is to achieve coherent overlap by means of simply adapting the placement of the consecutive frames such that the cross correlation of the synthesized frames is maximized. In the phase vocoder this can be achieved without adaptation of the position of the synthesis frames by means of adapting the phase of the bins constituting the sinusoidal components of the input signal according to

$$\widetilde{\Phi_l(k)} = \Phi_l(k) + (\Theta_l(k) + \frac{2\pi k}{N})\Delta_n. \qquad (7)$$

Here $\Phi_l(k)$ is the original phase of the input frame l in bin k and Δ_n is the time shift that has to be applied to the original signal to maximize phase alignement between the previous anbd the current synthesis frames. This time shift will be determined below. Because Δ_n is generally very small (smaller than half the fundamental period of sound segment under operation) and because the recursive structure of eq. (6) is avoided the vertical inter-partial phase synchronization is always maintained such that shape invariant processing is achieved. The intra peak phase synchronization discussed above can be used but is not as important as in the standard phase vocoder algorithm because time displacements are kept sufficiently small (smaller than half a fundamental period of the current signal) to avoid desintegration of the sinusoidal peaks.

Compared to the original time domain SOLA algorithm the phase vocoder based SOLA does not require any adaptation of the position of the synthesis frame and therefore, no modification of the local time stretching factor has to be made. Moreover, as explained below, we can constrain the cross correlation to use only sinusoidal signal components such that any effects of the signal background noise during the estimation of the optimal overlap position is significantly reduced.

So far we have introduced the basic principle of the new phase vocoder based shape invariant speech processing algorithm. In the following we will discuss two details that require a solution: the estimation of the optimal delay and the handling of the unvoiced or noise signal components.

3.1. Estimation of the optimal time shift

To coherently calculate the optimal time shift we would like to use the cross correlation between the last synthesis frame after phase adaptation has been applied $\widetilde{Y_{l-1}(k)}$ and the current unmodified synthesis frame $Y_l(k)$ that has been placed in the position $C'(l)$ following the desired time stretching factor. To avoid the impact of the signal noise during the estimation of the delay parameter we propose to restrict the cross correlation to the sinusoidal components of the respective frames. This can be achieved by means of a spectral mask $S_l(k)$ that retains only spectral bins that constitute the spectral peaks related to sinusoidal signal components. A very efficient and straight forward means to establish this mask is available in form of an algorithm for sinusoidal peak classification [18]. When this algorithm is used (a) the esti-

mation of the fundamental frequency is not required. Alternatively the mask can be generated by means of removing all bins outside the frequency range $[0.5 \cdot F0, N \cdot F0]$ (b) or even simpler by means of masking all bins outside a fixed frequency range (c). Best results are obtained with method (a) but methods (b) and (c) still provide an advantage compared to the time domain SOLA algorithm that always uses the complete signal spectrum to calculate the best frame position and that requires to actually move the frame to the required position compromising the local time stretch factor.

The cross correlation sequence can be calculated in the spectral domain if $N \geq 2M$. In the following presentation we will assume that this condition holds and will discuss the approach that handles the case $N < 2M$ later on. For $N \geq 2M$ the cross correlation sequence for the sinusoidal components will be denoted as $Z(n)$ and is given by

$$Z(n) = \sum_{k=0}^{N}((Y_l(k)^* S_l(k) \widetilde{Y_{l-1}}(k) S_{l-1}(k)) e^{j\frac{kn}{N}2\pi}. \quad (8)$$

Here $Y_l(k)^*$ represents the conjugate complex of $Y_l(k)$. 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 $w(n)$ the cross correlation sequence $Z(n)$ displays locally a quasi periodic evolution that has the auto correlation sequence of the analysis window superimposed. The local maxima of the cross correlation sequence can be used to determine the optimal time shift that is required for the new synthesis frame to optimally overlap with the previous frame. In the phase vocoder system the synthesis frames are placed at specific locations and therefore the maximum of $Z(n)$ should not be taken directly to determine the optimal delay time for the phase alignment of the frames Y_l and $\widetilde{Y_{l-1}}$.

If the frame offset between Y_l and Y_{l-1} is given by $O_l = C'_l - C'_{l-1}$ our interest is to find the appropriate maximum of the underlying periodic structure of the cross correlation sequence removing as much as possible the effect of the analysis window. We first note that for the unmodified signal the optimal time delay between the successive frames is 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. For modified signals we would like the delay be as close as possible to O_l such that the changes to be applied to the phases of the synthesis frames are minimized. If P is the length of the signal period at the center of the current synthesis frame Y_l we would like the time shift to stay within $O_l \pm P/2$.

If we denote the autocorrelation sequence of the analysis window with $Z_w(n)$ we can determine the approximate value of the optimal time shift following the constraints discussed above by means of

$$N(n) = \max(Z_w(n), Z_w(D)) \quad (9)$$

$$Z'(n) = Z(n)/N(n) \quad (10)$$

$$O'_l = \arg\max_n(Z'(n)N(n - O_l)) \quad (11)$$

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. Accordinlgy we determine a maximum absolut time offset D to be used. The $\max$ operation limits the compensation to the range that contains sufficient samples to prevent degeneration of the compensated cross correlation. The limiting value D to be used and accordingly $Z_w(D)$ is derived as follows:
Due to the fact that in the phase vocoder all sinusoidal peaks need to be sufficiently resolved we can assume $P <= P_{max} = M/4$. Moreover we will assume $O_l < O_{max} = M/3$. This limit is due to the fact that the correct estimation of the frequency to be used for the phase update mechanism in the phase vocoder may be compromised due to phase wrapping effects if O_l becomes too large. The present discussion suggests that compensation of $Z_w(n)$ in $Z(n)$ is only needed for $n < P_{max}/2 + O_{max}$ such that we can select $D = M/3 + M/8$. If other limits P_{max} and O_{max} are desired D can be adapted accordingly.

The sequence $Z'(n)$ represents then the cross correlation sequence after compensation of the systematic impact of the analysis window by means of $N(n)$. The optimal time delay is determined from $Z'(n)N(n - O_l)$ to favor small time shifts with respect to the nominal position O_l. The small bias of the time shift O'_l that is due to the multiplication with $N(n - O_l)$ can be removed by means of searching the local maximum of the compensated cross correlation sequence $Z'(n)$ in the direct neighborhood of O'_l. As a result we find the optimal bias free time delay $O_{l,opt}$, which can then be used to determine the displacement operator Δ_n that is used in eq. (7) to adapt the phase of the local frame.

As mentioned above the estimation of the optimal time shift takes into account only sinusoidal components and will therefore provide a precise phase alignment of the overlapping frames. Only the noise present in the sinusoidal peaks will effect the estimation of the time shift. A major advantage of the procedure is the fact that there is no need to know the fundamental frequency or pulse positions of the signal to achieve phase synchronous overlap add. A missclassification of some (or many) of the sinusoidal components will not significantly impact the result as long as the

maximum common divisor of the partial numbers of the detected sinusoidal components is 1. Accordingly the algorithm's robustness against miss-classifications of the sinusoidal components is relatively high.

To conclude the discussion of the time shift estimation we reconsider the case when $N < 2M$. In that case we use spectral domain interpolation of the complex signal spectrum prior to masking to double the DFT size N. In our implementation we use the a linear phase FIR filter designed according to the Kaiser window filter design method [19, ,chapter 7] requiring $60dB$ sidelobe rejection and a transition bandwidth of 7% of the target FFT size. Note that the spectral domain interpolation can be limited to the frequency range containing sinusoidal components and, therefore, its costs relative to the complete processing costs are relatively small ($< 10\%$).

3.2. 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. We first note that the aperiodic signal components in a speech signal can have very different properties. In the following we will discuss the three classes of aperiodic signal components introduced in [20]: 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 are very common in musical signals and they can be handled with very high quality in the phase vocoder [16]. No additional measures for speech signals are required. 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 standard phase vocoder algorithm that achieves comparably good quality for this kind of signals. The modulated noises that are present in voiced signal components are considered to be perceptually important for the fusion of voiced and unvoiced signal components [5]. While preservation of the amplitude modulation of the modulated noise component may seem to pose a difficult problem it turns out that the proposed algorithm is sufficient to achieve perceptually convincing preservation of the modulation of the modulated noise components. First we note that 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 phase adaptation procedure of the phase vocoder works reasonably well for noise signals ensuring that noise components do not cancel. The remaining question to be addressed is whether the phase update will preserve as well the amplitude modulation of the noise components related to the excitation pulse.

The experimental example displayed in fig. 2 shows that the noise components will in fact preserve an important part of the amplitude modulation. This can be explained as follows. The amplitude modulation of the noise component will introduce a interdependency (correlation) between the phase spectrum at distant bins in the noise spectrum. Because these interdependencies will be reflected in the phase update equations to be used in the phase vocoder the characteristic interdependencies in the phase spectrum will be preserved by the phase vocoder phase update procedure. As a result we can observe that the amplitude modulation of the noise signal remains present after the phase update.

A last problem related to unvoiced signal components is the fact that with increasing time stretching factors the unvoiced signal components are progressively transformed into noise with a tonal quality. This effect can be avoided by means of a small phase randomization of the unvoiced signal components. As a first step we establish a voiced/unvoiced frequency boundary (VUF) that separates voiced and unvoiced frequency bands similar to the maximum voiced frequency used in [5]. In our algorithm this frequency is obtained for each frame simply from the largest frequency enclosing all spectral peaks that are classified as sinusoidal. Whenever the effective time stretching factor is larger than 1.2 we add a random uniformly distributed phase offset ($|\Delta p| < 0.3\pi$) to the phase of all spectral peaks above the (VUF).

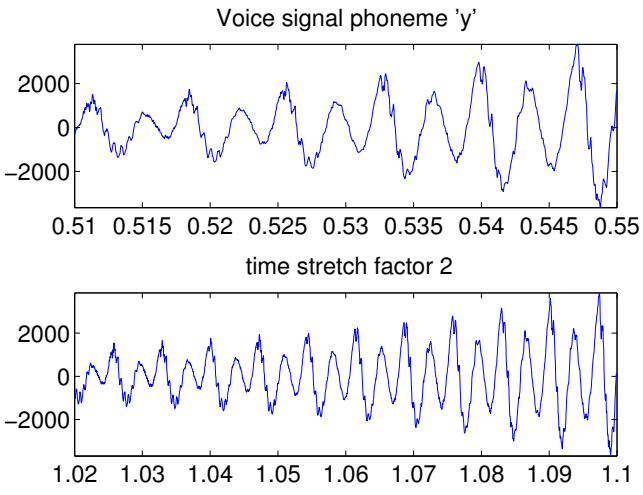

Figure 1: *signal waveform of original and time stretched speech signal containing phoneme 'y'*

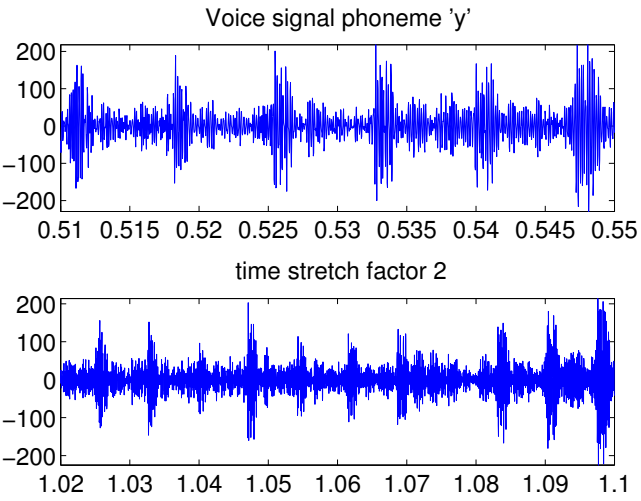

Figure 2: *original and time stretched speech signal containing phoneme 'y' after high pass filtering ($> 5kHz$). The original signal displays pitch synchronous noise modulation which is preserved after the transformation.*

4. EVALUATION AND DISCUSSION OF RESULTS

The proposed algorithm does achieve comparatively high signal quality for time scale modifications of more than a factor of 2. As an example we use here the results for time stretching a sentence of a french male speech signal by a factor of 2. A short segment of the original and transformed speech segment containing the phoneme 'y' are shown in upper and lower pane of fig. 1. The waveform clearly has preserved its original shape. Besides the fact that the speech rhythm is unnaturally slow the signal sounds very natural. The signals presented in fig. 2 show the same signal after band-stop filtering frequencies up to 5kHz. The band filtered original signal contains modulated high frequency noise. The time stretched signal clearly preserves a considerable part of the noise modulation as discussed in section **3.2**.

The algorithm presented in the present paper has been evaluated in two subjective tests comparing it to a recent implementation of PSOLA and HNM algorithms. The proposed shape invariant phase vocoder algorithm will be referred to as SHIP. The first test has been part of the French ANR Project VIVOS that was aiming at high level speech transformations (gender and age changes of the perceived speaker). The subjective test did evaluate transpositions with timbre preservation with transposition factors of 2, 1.4, 1/1.4, and 1/2. To achieve pitch shifting with timbre preservation an initial version of the algorithm described above has been combined with a resampling stage. The timbre preservation has been obtained by means of removing the estimated spectral envelope estimated according to [11] prior to resampling and reestablishing it afterward. The 10 subjects participating in the test were asked to evaluate the degradation of the natural quality of the transformed signal compared to the original signal on a 5 level scale containing the levels: 5 *not noticeable*, 4 *just noticeable*, 3 *noticeable and slightly disturbing*, 2 *disturbing but tolerable*, 1 *very disturbing*. The results are displayed in table (1).

Table 1: *Perceptive test results comparing PSOLA/HNM/SHIP. Given are the average quality levels (5=best and 1=worst) for the transposition up and down. Best results are displayed in bold*

	transposition up	transposition down
PSOLA	2.8 ± 1.2	$\mathbf{3.8 \pm 1}$
HNM	2 ± 1	1.7 ± 1
SHIP	$\mathbf{3.1 \pm 1}$	1.9 ± 1

We note first that the HNM based transformation was consistently evaluated to provide the lowest quality of all algorithms. The results show further that the proposed method was preferred for the transposition upwards but did not perform very well for transposition down. The standard deviation of about 1 quality level is explained by the fact that the people in the perceptive test are using various interpretations of the perceptual quality *disturbing* and *tolerable*.

Investigation into the weak performance of the SHIP algorithm when transposing downward revealed an explanation and a partial solution. The problem is related to the fact that pitch shifting downwards may move noise signal components that are present in the high frequency region down into the formants in the lower frequency range. Due to the change of the excitation quality in the formant frequency region the transformed voice may loose its original clarity. The same problem exists for transposition up when excitation energy that was located in frequency regions with low energy where the sinusoidal excitation is covered by background noise enters into formants. In this case again the excitation signal will loose its clarity as well, however, the case is expected to happen less frequently when transposing up. In the initial version of the implementation used in the test 1 the problem was aggravated by the fact that the VUF frequency was considered to be transposed such that phase randomization of the signal components that were above the VUF would finally affect the excitation signal at frequencies that are amplified by the formants after transposition.

Following the investigation a modified version was implemented that aimed to enhance the signal quality by means of changing the phase randomization procedure such that 1) the VUF is preserved and therefore phase randomization will never affect signal components that will be transposed into the formants, and 2) phase randomization is only applied if the signal transformation contains an effective

time stretching transformation as detailed at the end of section **3.2**. After improving the SHIP algorithm a second evaluation has been performed comparing only the PSOLA and SHIP algorithms. This time the signal transformation that were evaluated covered transposition with timbre preservation (transposition factors 0.5 and 2) as well as time scaling (with factors 0.5 and 2). The second test comprises only relatively extreme transformations that are considered difficult for state of the art speech transformation algorithms. The reduction of examples was helping to motivate the subjects to participate and concentrate on the task. Transformed signals based on a male and female speaker have been evaluated by 29 individuals with varying professional background. Results for the PSOLA and the modified SHIP algorithm are displayed in table (2).

Table 2: *Perceptive test results comparing PSOLA/SHIP using the modified SHIP phase randomization described in the text. Given are the average quality levels (5=best and 1=worst) for transposition up and down as well as time stretching and compression. Best results are displayed in bold.*

	all examples			
	transposition		time scale	
	up	down	compress	stretch
PSOLA	2.6 ± 0.8	2.6 ± 1	4.1 ± 0.9	1.9 ± 0.8
SHIP	$\mathbf{2.9 \pm 0.9}$	$\mathbf{2.8 \pm 1}$	$\mathbf{4.3 \pm 0.6}$	$\mathbf{2.9 \pm 0.9}$
	female speaker			
	transposition		time scale	
	up	down	compress	stretch
PSOLA	2.5 ± 0.8	$2. \pm 1$	$4. \pm 0.8$	1.9 ± 0.6
SHIP	$\mathbf{3.2 \pm 0.9}$	$\mathbf{2.8 \pm 1}$	$\mathbf{4.4 \pm 0.6}$	$\mathbf{2.6 \pm 0.9}$
	male speaker			
	transposition		time scale	
	up	down	compress	stretch
PSOLA	$\mathbf{2.8 \pm 0.9}$	$\mathbf{3.1 \pm 1.1}$	4.2 ± 1	2 ± 0.9
SHIP	2.7 ± 0.8	2.9 ± 1	$\mathbf{4.3 \pm 0.7}$	$\mathbf{3.2 \pm 0.9}$

Considering first the average quality over both genders we find that the small changes discussed above did lead to a significant improvement of the SHIP algorithm such that it is now outperforming the PSOLA algorithm in all cases.

If the results are broken down according to the gender of the speaker we find that the SHIP algorithm is always leading to the best signal quality - besides for transposition of the male speaker. While the perceived quality of the SHIP transposed signal is about the same for female and male speakers PSOLA transposed signals are loosing nearly one quality grade for female speakers such that the SHIP algorithm is able to clearly outperform the PSOLA algorithm when transposing female speakers. For time stretching the SHIP algorithm is outperforming the PSOLA algorithm on average by nearly one quality grade. For time compression the difference is less pronounced notably because time compression generates less artifacts in both algorithms such that signal quality is very high in both cases.

We conclude that the SHIP algorithm can significantly improve signal quality for extreme transformations. The main problem that is 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 perceived quality of the transformed speech. An appropriate operator that allows to change noise excitation into sinusoidal excitation and that will hopefully improve the result in at least some situations is currently under investigation.

5. 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. The frequency domain representation of the signal in the phase vocoder provides multiple advantages compared to the time domain SOLA algorithm: only the sinusoidal components will be used for synchronization between the consecutive frames such that the synchronization for segments with few sinusoidal components is improved, the use of the phase vocoder algorithm for phase spectrum adaptation reduces the cancellation of unvoiced signal components and preserves the noise modulation that is characteristic for breathy vowels or voiced fricatives, frequency domain treatments as for example independent modification of the excitation signal and the spectral envelope [11] or transposition by means of peak shifting[7] which would further reduce the latency.

Compared to other high quality speech transformation algorithms, the proposed algorithm shares the advantage of the SOLA system that it does not require an elaborate preanalysis (pitch marks, F0). The algorithm is based on a very cheap classification of sinusoidal and noise peaks that can be performed on the fly directly in the DFT frames [18].

The perceptual evaluation of the transformed signals reveals that the algorithm achieves better quality than a recent HNM implementation. Compared to PSOLA the algorithm achieves significantly better quality for time scale modification and transposition of female speech but cannot quite achieve the quality of the PSOLA algorithm when transposing male speech. On of the remaining problems related to pitch shifting operation has been determined and current research activities are investigating into a solution of the problem.

6. 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] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1993, pp. 554–558.

[3] 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.

[4] 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.

[5] 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.

[6] Y. Stylianou, *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*, Ph.D. thesis, Ecole Nationale Supèrieure des Télécommunications, Paris, France, 1996.

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

[8] J. Bonada, "Wide-band harmonic sinusoidal modeling," in *Proc. Inter. Conf. on Digital Audio Effects (DAFx)*, 2008, pp. pp. 265–272.

[9] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1997, vol. 2, pp. 1303–1306.

[10] 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.

[11] 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.

[12] M. Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

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

[14] J. Laroche and M. Dolson, "New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing and other exotic audio modifications," *Journal of the AES*, vol. 47, no. 11, pp. 928–936, 1999.

[15] 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.

[16] 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.

[17] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[18] 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.

[19] A. V. Oppenheim, R. W. Schafer, and John. R. Buck, *Discrete-Time Signal processing*, Prentice-Hall Intern., 2nd edition, 1995.

[20] G. Richard and C. Alessandro, "Analysis/synthesis and modification of the speech aperiodic components," *Speech Communication*, vol. 19, no. 3, pp. 221–244, 1996.