Speech to chant transformation with the phase vocoder

Axel Roebel[1], *Joshua Fineberg*[2]

[1] IRCAM-CNRS-STMS, Paris, France
[2]Havard University, Boston, USA
roebel@ircam.fr, fineberg@fas.harvard.edu

Abstract

The technique used for this composition is a semi automatic system for speech to chant conversion. The transformation is performed using an implementation of shapeinvariant signal modifications in the phase vocoder and a recent technique for envelope estimation that is denoted as True Envelope estimation. We first describe the compositional idea and give an overview of the preprocessing steps that were required to identify the parts of the speech signal that can be used to carry the singing voice. Furthermore we describe the envelope processing that was used to be able to continusously transform the orginal voice of the actor into different female singing voices.

Index Terms: envelope estimation, speech transformation, chant synthesis.

1. Introduction

Perhaps one of the most tantalizing aspects of new technology is its potential for rendering things that ought not be able to exist. This is a particular challenge in music because it requires the sonic objects to seem both real and impossible simultaneously: If they sound artificial the sense of impossibility disappears.

The particular context of the composition being presented in the following is a dance piece being written by one of us (JF) based on Lolita by Vladimir Nabokov. The work is conceived like an opera, but one that occurs completely within the mind of the narrator. All the *sung* voices heard by the audience are the result of computer transformations of the narrator's spoken voice: they are simply manifestations of his speech/writing. In this way, the piece is a truly *imaginary* opera: It is the *opera* imagined in the mind of the narrator.

One of the compositional goals were the possibility to construct sound signals that support the understanding that the actors mind changes between reality and imagination. In the imaginary parts the actor invents female persons which, in the performance, should be supported by the fact that the actors voice is gradually transformed into female voices.

The final goal is an complete automatic transformation system that allows to transform the actors voice in real time into chant with the syllables automatically assigned to the notes in the composed score. In its current state, the system is not yet capable to achieve the automatic alignment of the spoken syllables and the score. We are in the process to adapt the score alignment system presented in [1] for this purpose. Therefore, for the premier performance of the first part of the opera the actor has been recorded before the performance and the speech signal has then been labeled into the segments corresponding with the notes of the score by hand.

The composed score requires extreme transposition of up to 3 octaves up and 1 octave down. In the process of the alignment of the speech signal to the score the vocal syllables are going to be stretched (depending on the actors articulation speed) by factor up to 8-10. Because signal transposition with preservation of the spectral envelope does in fact preserve the speaker identity we make use of an additional female voice, recorded with singing articulation with approximately constant fundamental frequency to create a natural envelope data base that will be used to replace the spectral envelope of the original actor.

In the following we will explain some technical details and furthermore explain the complete processing chain that has been used to create the sound file example.

2. Technical details

The underlying speech signal model is the standard source filter model using the true envelope estimator described in [2, 3]. The True envelope estimator achieves a cepstral representation of the spectral envelope that automatically selects the important spectral peaks to be enveloped according to the cepstral order. The considerable advantage of the True envelope estimator is the fact that the optimal model order can be deduced from the fundamental frequency of the sound signal [4]. The optimal order is given by

$$O_{opt} = \frac{R}{2F_0}, \tag{1}$$

where R is the sample rate of the sound signal and F_0 is the fundamental. The optimal order is simply deduced from the fact that the basis function of the highest cepstral coefficient does not allow the spectral mode to represent the space in between the harmonic grid of the signal spectrum. The eq. (1) assumes that the vocal tract filter is sampled regularly by the excitation signal. This formula reflects the fact that the information that is available about the vocal tract filter is limited by the fact that the transfer function is sampled by the harmonic structure of the excitation signal.

Note however, that for real world sound signals the sampling grid is not completely regular. At frequency $w = 0$ a sample point will be missing due to the fact that the signal transmission chain does remove all constant values. Accordingly, as proposed in [4] we use a 2 step approach to envelope estimation where the first step is the estimation of an envelope using an order $O_{opt}/2$. The results of this envelope estimation are used to fill the artificial hole of the envelope at frequency $w = 0$ by means of adding an artificial peak with amplitude given by the estimated envelope at that position. Then, in the second step the optimal order O_{opt} can be used to estimate the final envelope with all available details.

2.1. Transposition

Signal transposition in the phase vocoder can be obtained either in the spectral domain, by means of shifting the spectral peaks,

or in the time domain, by means of re-sampling the signal and apply a time stretch for compensating the change of the time duration [5]. The main advantage of the first technique is the fact that it's computational demands do not depend on the transposition. The disadvantage is, however, that the time precision is determined by the window duration.

The f0 estimation technique that has been used to determine the original pitch of the speech signal is the yin algorithm as described in [6]. To obtain an artifact free transposition that neutralizes the original pitch contour we tried two different implementations of the time variant re-sampling. The first approach was using piecewise constant transpositions and the second approach was using a piecewise linearly interpolated re-sampling function. In the experimental evaluation we found that the approach with piecewise constant transposition required a frame step size of approximately 1.5ms while the piece wise linearly interpolated transposition achieved an equivalent quality with frame step size of about 6ms. Because the inverse frame step size will directly affect the computational costs of the transformation in the phase vocoder we selected the linearly interpolated transposition despite its increased algorithmic complexity.

3. System overview

The whole transformation system consists of the following steps:

- create envelope data base:
  record of female singer for the complete set of phrases to be transformed. The singer should use relatively low and constant pitch to convey as much information about the spectral envelope as possible.
  The recording is labeled and for each segment a time dependent target pitch and target position and a time dependent mixing coefficient for the male and female envelopes are specified following the score of the composition.
- create actor recording:
  The actors voice is recorded for the same segments. In the current state of the system the actors voice needs to be labeled as well. The automatic alignment of the actors voice with the females singer voice will hopefully soon be possible. The fundamental frequency of the actors voice is estimated for the complete set of recordings.
- transform actors voice :
  The speech signal of the actor is aligned to the target positions and target pitch using time stretching and transposition in the phase vocoder at the same time the phase vocoder implementation allows us to scale the original envelope and to add a scaled version of the spectral envelope of the female database.

4. Conclusions

We described a speech to chant conversion system that is currently developed with goal to be used in a theater play that will be performed in 2008. As a preview of the complete performance the first part of the play has been performed in 2006. The voice transformation system that has been developed until then did achieve results with rather convincing quality.

At the current state a semi automatic transformation can be achieved that is robust with respect to transformation artifacts for transpositions that extend up to 3 octaves up. For transposing down we observe a significant loss in the quality of the transformed signal already for transpositions of about 500cents. We are currently investigating into this problem.

5. References

[1] X. Rodet, J. Escribe, and S. Durigon, "Improving score to audio alignment: Percussion alignment and precise onset estimation," in *Proc Int. Conf on Computer Music (ICMC)*, 2004, pp. 450–453.

[2] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Proc. of the 8th Int. Conf. on Digital Audio Effects (DAFx05)*, 2005, pp. 30–35.

[3] F. Villaviciencio, A. Röbel, and X. Rodet, "ImprovingLPCspectralenvelopeextractionofvoicedspeechby trueenvelopeestimation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2006, pp. 869–872 (Vol. I).

[4] A. Röbel, F. Villaviciencio, 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*, 2007, accepted for publication.

[5] 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.

[6] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Journal Acoust. Soc. Am.*, vol. 111, no. 4, 2002.