Instantaneous Harmonic Representation of Speech Using Multicomponent Sinusoidal Excitation

Elias Azarov, Maxim Vashkevich, Alexander Petrovsky

Department of Computer Engineering, Belarusian State University
of Informatics and Radioelectronics
6, P.Brovky str., 220013, Minsk, Belarus

azarov@bsuir.by, vashkevich@bsuir.by, palex@bsuir.by

Abstract

This paper introduces a framework for parametric speech modeling that can be used in various speech applications such as text-to-speech synthesis, voice conversion etc. In order to reduce impact of pitch variations the harmonic analysis is done in the warped time scale that is aligned with instantaneous pitch values. It is assumed that each harmonic has its own periodic excitation source that evolves in time and can be modeled as a sum of several sinusoidal components with close frequencies. The parameters of the excitation components are estimated using a modified instantaneous Prony's method. The proposed analysis/synthesis technique is compared with TANDEM-STRAIGHT.

Index Terms: harmonic representation of speech, speech morphing

1. Introduction

Wide-band speech modification (such as time-stretching, changing of pitch and spectral envelopes) is a challenging task that requires developing of rather sophisticated models. There are some very impressive tools for speech morphing like TANDEM-STRAIGHT [1,2] and AHOcoder [3] that perform morphing effects with a low level of audible artifacts. The success of these tools justifies applicability of harmonic representation as the main component in voiced speech modeling. There is a report of full-band speech modeling based solely on harmonic parameters [4].

Basically harmonic modeling represents speech as a sum of periodical (sinusoidal) components with slowly varying parameters. The frequencies of the components are multiples of the current pitch value and may change very rapidly. This is why accurate harmonic representation of wide-band speech is an extremely difficult task. One of the possible perceptually motivated solutions to the problem is to model high-frequency part of the spectra using stochastic signals, however this approach leads to some loss of natural sonorousness of the vowels. The second major challenge is to model voiced sounds with mixed excitation. In STRAIGHT an aperiodicity spectrogram is extracted that represents the ratio between aperiodic and periodic components in each frequency band. The output signal is a combination of two separate parts synthesized with different (aperiodic and periodic) excitations. Though the overall synthesis quality of STRAIGHT is very high the vowels are still synthesized with a touch of artificiality.

The main idea behind the approach presented in this paper is that the resynthesis quality of vowels can be improved if we find a consistent periodic model that can handle wide-band mixed excitations. We assume that the entire spectral band of voiced speech consists of harmonics and each of them has its own excitation source which can be modeled as a sum of sinusoidal components with close frequencies. Each harmonic is represented as a complex analytical signal and separated from others using a DFT-modulated filter bank. To ensure that each harmonic is placed in a separate channel of the filter bank we use adaptive time warping i.e. the time axis of the signal is scaled in a way that always keeps instantaneous pitch constant. The parameters of the excitation components (amplitude, frequency and initial phase) are estimated using a modified instantaneous Prony's method that matches subchannel signal's derivatives. The proposed model of voiced speech uses only periodic functions for excitation without any noise generation.

In the evaluation section of the paper the proposed analysis/synthesis framework (referred to as 'GUSLY') is compared with the state-of-the-art TANDEM-STRAIGHT model. The experiments show that GUSLY provides high-quality reconstruction of morphed speech and can be used in various wide-band speech applications.

2. Analysis / synthesis algorithm outline

2.1. Harmonic / noise separation

The model represents the signal as a sequence of parametric frames which are classified as either voiced or unvoiced. The voiced/unvoiced decision is made by an excitation detector based on analysis of spectral envelope shapes[1]. The modeling of unvoiced frames is made using random excitation (white noise) filtered with a filter that approximates the target power spectral density of the frame. Since the approach is quite well-known we skip its detailed description and throughout the rest of the paper we focus solely on processing of voiced frames.

2.2. Parametric representation of the signal

2.2.1. Harmonic plus noise model

Basically harmonic plus noise representation of speech can be considered as a sum of periodic and aperiodic parts:

$$s(n) = \sum_{k=1}^{K} A_k(n) \cos \varphi_k(n) + r(n), \qquad (1)$$

where $A_k(n)$ – instantaneous magnitude of the k-th harmonic, K – number of harmonics, $r(n)$ – noise part (sometimes referred to as residual) and $\varphi_k(n)$ is instantaneous phase of the k-th component. Instantaneous phase is related to

[1] The description of the detector is not given in the paper considering its relative insignificance (any robust voiced/unvoiced classifier could be used instead)

instantaneous normalized angular pitch frequency $f_0(n)$ as follows:

$$\varphi_k(n) = \sum_{i=0}^{n} f_0(i)k + \varphi_k(0),$$

where $\varphi_k(0)$ is the initial phase of k-th harmonic.

2.2.2. Harmonic model with multicomponent sinusoidal excitation

In order to obtain a pure periodic model we should represent the residual part in (1) using periodic functions. A known solution is to make updates of model parameters frequently enough so harmonics become able to model noisy signals [4]. This approach is quite suitable in the case where signal reconstruction is needed (no morphing is applied) otherwise it is prone to audible artifacts. Instead of increasing update rate we propose to add more components to the model namely to represent each harmonic as a sum of sinusoidal components.

Let us assume that the pitch frequency is constant. Then we can introduce the harmonic model with multicomponent excitation in the following way:

$$s(n) = \overbrace{\sum_{k=1}^{K} G_k(n)}^{envelope} \overbrace{\sum_{c=1}^{C} A_k^c(n)\cos(f_k^c n + \varphi_k^c(0))}^{excitation} \quad (2)$$

where $G_k(n)$ is a gain factor specified by the spectral envelope and C – number of sinusoidal components for each harmonic, f_k^c and $\varphi_k^c(0)$ – frequency and initial phase of c-th component of k-th harmonic respectively. Amplitudes are normalized in order to set the unit energy to each harmonic's excitation:

$$\frac{1}{2}\sum_{c=1}^{C}[A_c^k(n)]^2 = 1$$

for $k = 1, \dots, K$. To obtain periodic excitation i.e.

$$\cos(f_k^c n) = \cos(f_k^c (n + li))$$

for $i \in \mathbb{Z}$, $k = 1, \dots, K$ and $c = 1, \dots, C$ where l is number of pitch periods we confine frequencies f_k^c to a finite set of uniformly spaced values

$$f_k^c \in \left\{\frac{2\pi}{li}, 2\frac{2\pi}{li}, 3\frac{2\pi}{li}, \dots\right\}. \quad (3)$$

2.3. Analysis routine

The analysis routine schematically shown in figure 1 involves the following steps.

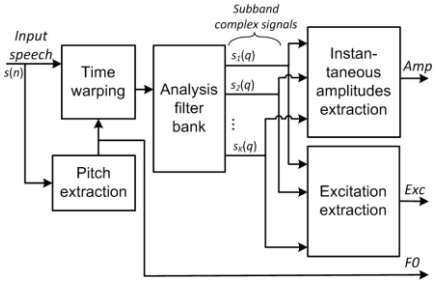

Figure 1: A*nalysis routine*

2.3.1. Pitch extraction

The instantaneous version of RAPT (Robust Algorithm for Pitch Tracking) [5] is used for pitch values extraction. The algorithm provides instantaneous estimates and is rather accurate for pitch modulations.

2.3.2. Time warping

Accurate estimation of model parameters requires separation of the signal into individual harmonics. Since the presented model assumes constant pitch the time axis of the signal is adaptively warped in order to eliminate pitch modulations [6]. Signal $s(n)$ is recalculated in new time moments m in such way that each period of pitch has equal number of samples N_{f0}. For every time sample $s(n)$ a phase mark $\phi(n)$ is associated using instantaneous pitch values $f_0(n)$:

$$\phi(n) = \sum_{i=0}^{n} f_0(i).$$

Thus new time moments m are obtained as:

$$m = \phi^{-1}(q/N_{f0}),$$

where q is sample index in warped time domain. The samples of the signal $s(q)$ are recalculated using sinc-interpolation. Figure 2 shows an example of voiced speech in time and warped-time domains.

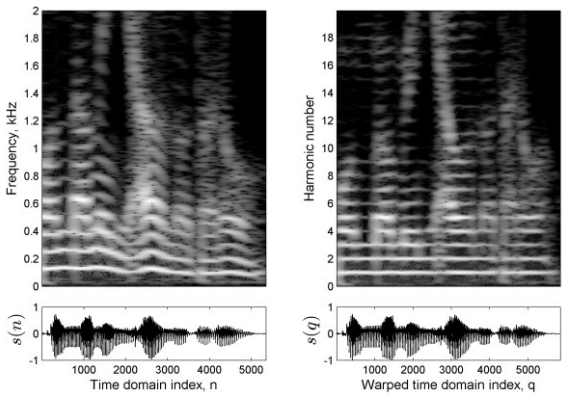

Figure 2: *Time-frequency and warped time-frequency representations of speech*

2.3.3. Subchannel amplitudes extraction

After applying time warping the required harmonic separation can be done effectively using a uniform DFT-modulated analysis filter bank with N_{f0} channels. According to the Nyquist–Shannon sampling theorem the maximum effective number of harmonics K is specified by the number of samples per period as $K = N_{f0}/2$. Center frequencies of the uniform N_{f0}-channel filter bank exactly correspond to integer multiplies of the constant pitch.

Instantaneous harmonic amplitudes are estimated from subband signals $s_k(q)$ as

$$A_k(q) = \sqrt{\mathrm{Re}^2(s_k(q)) + \mathrm{Im}^2(s_k(q))}$$

where Re and Im denote the real and imaginary parts respectively.

2.4. Synthesis routine

Figure 3 shows the synthesis routine that consists of the following steps: 1) the decimated excitation sequence is generated from the estimated excitation parameters using (2); 2) instantaneous amplitudes are recalculated according to the morphing task; 3) using the new (target) values of pitch excitation gain factors are recalculated; 3) the warped output signal is synthesized using a DFT-modulated synthesis filter bank; 4) the signal is unwrapped according to the target pitch contour.

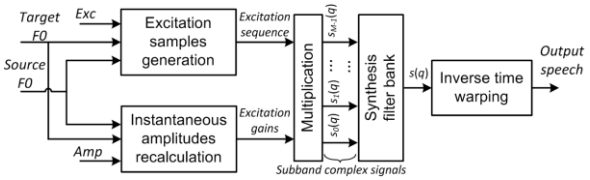

Figure 3: *Synthesis routine*

3. Extraction of sinusoidal excitation

3.1. Prony's method

According to Prony's method [7] discrete-time complex signal $s(n)$ is represented as a sum of dumped complex exponents:

$$s(n) = \sum_{k=1}^{p} h_k \, z_k^{n-1}$$

where p is the number of exponents, $h_k = A_k e^{j\theta_k}$ is an initial complex amplitude and $z_k = e^{\alpha_k + j f_k}$ is a time-dependent dumped complex exponent with dumping factor α_k and normalized angular frequency f_k. In order to estimate exact model parameters $2p$ complex samples of the signal are required. The solution is obtained using the following system of equations:

$$\begin{pmatrix} z_1^0 & z_2^0 & \dots & z_p^0 \\ z_1^1 & z_2^1 & \dots & z_p^1 \\ \vdots & \vdots & & \vdots \\ z_1^{p-1} & z_2^{p-1} & \dots & z_p^{p-1} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_p \end{pmatrix} = \begin{pmatrix} s(1) \\ s(2) \\ \vdots \\ s(p) \end{pmatrix} \qquad (4)$$

The required exponents $z_1, z_2, \dots, z_p$ are estimated as the roots of the polynomial

$$\psi(z) = \sum_{m=0}^{p} a(m) z^{p-m}$$

with complex coefficients $a(m)$ which are the solution of the system

$$\begin{pmatrix} s(p) & s(p-1) & \dots & s(1) \\ s(p+1) & s(p) & \dots & s(2) \\ \vdots & \vdots & & \vdots \\ s(2p-1) & s(2p-2) & \dots & s(p) \end{pmatrix} \begin{pmatrix} a(1) \\ a(2) \\ \vdots \\ a(p) \end{pmatrix} = \begin{pmatrix} s(p+1) \\ s(p+2) \\ \vdots \\ s(2p) \end{pmatrix}$$

and $a(0) = 1$. Each dumping factor α_k and frequency f_k are calculated using the following equations:

$$\alpha_k = \ln|z_k|, \quad f_k = \operatorname{atan}\left[\frac{\operatorname{Im}(z_k)}{\operatorname{Re}(z_k)}\right].$$

Using the extracted values of $z_1, z_2, \dots, z_p$ the system (4) is solved with respect to complex parameters $h_1, h_2, \dots, h_p$. From each of these parameters initial amplitude A_k and phase θ_k are calculated as:

$$A_k = |h_k|, \quad \theta_k = \operatorname{atan}\left[\frac{\operatorname{Im}(h_k)}{\operatorname{Re}(h_k)}\right].$$

3.2. Modified instantaneous Prony's method

The parameters of dumped exponents estimated using the original Prony's method are averaged over observation period $2pT$ where T is the sampling interval. It is possible to get a local moment-related modeling of the signal by matching its derivatives instead of adjacent samples. For a specified moment n we can require

$$s^{(d)}(n) = \left(\sum_{k=1}^{p} h_k \, z_k^{n-1} \right)^{(d)}$$

where (d) denotes derivative order from 0 to $p-1$. After differentiation with respect to n we get

$$s^{(d)}(n) = \sum_{k=1}^{p} h_k \, (\alpha_k + j f_k)^{d-1} z_k^{n-1}.$$

For the fixed moment $n = 1$ this can be written in the following simple form:

$$s^{(d)}(1) = \sum_{k=1}^{p} h_k \, y_k^{d-1}$$

where $y_k = \alpha_k + j f_k$. The equation leads to a system that is similar to (4), however the complex exponents are now expressed in terms of signal's derivatives that are related to the same specified moment of time:

$$\begin{pmatrix} y_1^0 & y_2^0 & \dots & y_p^0 \\ y_1^1 & y_2^1 & \dots & y_p^1 \\ \vdots & \vdots & & \vdots \\ y_1^{p-1} & y_2^{p-1} & \dots & y_p^{p-1} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_p \end{pmatrix} = \begin{pmatrix} s^{(0)}(1) \\ s^{(1)}(1) \\ \vdots \\ s^{(p-1)}(1) \end{pmatrix}. \qquad (5)$$

The required parameters can be extracted from system (5) just like from (4) except that dumping factor α_k and frequency f_k are calculated as:

$$\alpha_k = \operatorname{Re}(y_k), \; f_k = \operatorname{Im}(y_k).$$

3.3. Parameters extraction

At each specified moment of time excitation parameters are extracted using the modified instantaneous Prony's method. Each subchannel signal $s_k(q)$ of the analysis filter bank (see figure 1) that corresponds to a separate harmonic is represented as a sum of sinusoids with close frequencies. Figure 4 shows how excitation is actually modeled. On the left side of the figure the source signal is shown with an indicator of the moment where the excitation parameters are extracted. The signal on the right side is synthesized using the extracted parameters (we fix extracted pitch, amplitudes, initial phases and the envelope and use (2) for synthesis). The produced synthetic vowel with mixed excitation has the stochastic patterns (see harmonics 6-10) that very much resemble those in the source signal at the moment of parameters extraction.

In order to obtain a periodic excitation sequence with a specified period the frequencies of the components are quantized with the uniform frequency grid (3).

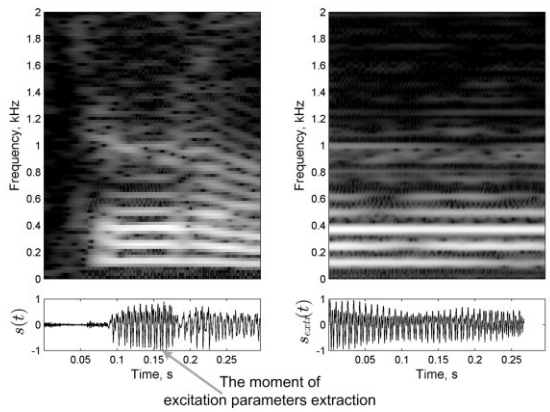

Figure 4: *Excitation modeling example*

4. Evaluation

The performance of the proposed model is compared to TANDEM-STRAIGHT [1] using subjective mean opinion score (MOS) measures. We use speech data of four speakers from the CMU ARCTIC database [8]: two male English speakers ('bdl' and 'rms') and two female English speakers ('clb' and 'slt'). The evaluation set consists of 10 sentences for each speaker. The speech morphing is performed using the described model (denoted as 'GUSLY') and the TANDEM-STRAIGHT model (denoted as 'T-S'). Some generated samples are available for download at http://dsp.tut.su/gusly_vs_straight.rar.

Twenty volunteers were asked to rate naturalness of the morphed speech in 1-to-5 scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad).

In the first experiment time stretching by factors of 1.5 (denoted as 'x 1.5') and 2.2 (denoted as 'x 2.2') is applied to speech. Figure 5 shows the average results of the first MOS test (male voices are denoted as 'm' and female voices as 'f'). We can see that proposed GUSLY method outperforms TANDEM-STRAIGHT when time stretch factor is equal to 1.5. However for stretch factor 2.2 GUSLY shows a lower performance which can be explained by emerging of sharp pre-echo of the transients.

In the second experiment pitch is increased by factors 1.2 (denoted as '↑ 1.2') and 1.9 (denoted as '↑ 1.9'). The evaluation results are presented in figure 6. For both male and female voices the results of GUSLY outperform those of TANDEM-STRAIGHT.

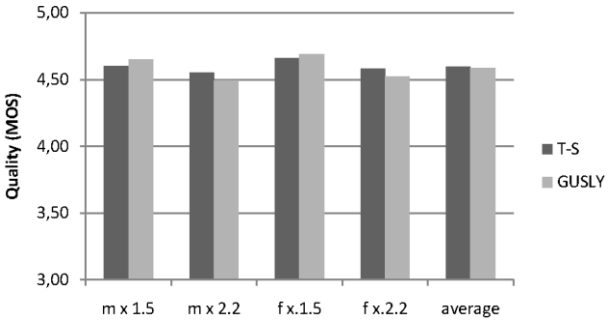

Figure 5: *Time stretching MOS evaluation*

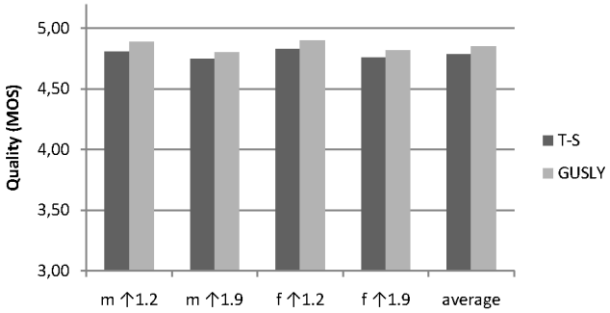

Figure 6: *Pitch increasing MOS evaluations*

In the third experiment pitch is decreased by factors 1/1.2 (denoted as '↓ 1/1.2') and 1/1.9 (denoted as '↓ 1/1.9'). Figure 7 shows that speech generated using GUSLY has slightly lower scores compared to TANDEM-STRAIGHT. This can be explained by the fact that the number of harmonics in the source signal is smaller than in the target and the excitations for high-frequency harmonics cannot be properly extracted.

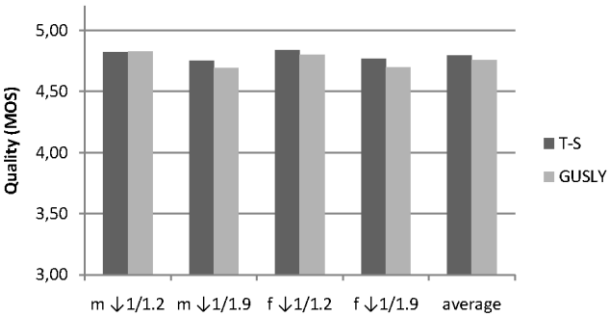

Figure 7: *Pitch decreasing MOS evaluations*

5. Conclusions

A speech modeling framework based on instantaneous harmonic parameters has been presented. The proposed model represents each harmonic of voiced speech as a sum of sinusoidal components multiplied by a gain factor. The instantaneous parameters of the components are extracted using the modified Prony's method. The processing is made in the warped time domain specified by instantaneous pitch contour. The subjective comparison with TANDEM-STRAIGHT shows that the proposed harmonic model can effectively represent wide-band voiced speech with mixed excitations and produce high-quality morphing effects.

6. Acknowledgements

The authors are grateful to professor Hideki Kawahara for the up to date TANDEM-STRAIGHT implementation that has been used for performance evaluations. The authors would like thank the IT-Mobile company for their support in implementing GUSLY as a part of voice conversion web services[1].

[1] The speech processing framework presented in the paper is a part of voice conversion web services 'CloneVoice' and 'CloneAudioBook'. Both services are to become available for users on August 2013 at: http://clonevoice.com/en.

7. References

[1] Kawahara H., Takahashi T., Morise M. and Banno H. "Development of exploratory research tools based on TANDEM-STRAIGHT," *Proc. APSIPA*, Japan Sapporo, Oct. 2009.

[2] Kawaahra H., Nisimura R., Irino T., Morise M., Takahashi T., and Banno B., "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," *Proc. ICASSP*, Taipei, Taiwan, April 2009.

[3] Erro D., Sainz I., Navas E., Hernaez I., "Improved HNM-based Vocoder for Statistical Synthesizers," *Proc. INTERSPEECH*, Florence, Italy, Aug. 2011.

[4] Degottlex, G., Stylianou, Y., "A full-band adaptive harmonic representation of speech," *Proc. INTERSPEECH*, Portland, Oregon, USA, Sep. 2012.

[5] Azarov, E., Vashkevich, M., and Petrovsky A., "Instantaneous pitch estimation based on RAPT framework," *Proc. EUSIPCO*, Bucharest, Romania, Aug. 2012.

[6] Gade S., Herlufsen H., Konstantin-Hansen H., Wismer H.J., "Order tracking analysis," Bruel & Kjaer Technical Review, 1995.

[7] Marple, S.L. "Digital spectral analysis: with applications" NJ, USA, Prentice-Hall, 1987.

[8] Kominek, J., and Black A., "The CMU ARCTIC speech databases for speech synthesis research," Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-LTI-03-177 http://festvox.org/cmu arctic/, 2003.