INSTANTANEOUS HARMONIC ANALYSIS FOR VOCAL PROCESSING

Elias Azarov and Alexander Petrovsky
Department of Computer Engineering,
Belarusian State University of Informatics and
Radioelectronics
Minsk, Belarus
palex@bsuir.by

ABSTRACT

The paper considers the application of instantaneous harmonic analysis to a real-time vocal processing system for pitch, timbre and time-scale modifications. The analysis technique is based on narrow band filtering using special analysis filters with frequency-modulated impulse response. The main advantage of the technique is high accuracy of harmonic parameters estimation that provides adequate harmonic/noise separation and artifact free implementing of voice modifications. The processing methods described in the paper are based on the harmonic+noise model.

1. INTRODUCTION

The present paper describes an approach to voice processing by means of the harmonic+noise model that considers a signal as a sum of two periodic (harmonic) and residual (noise) parts. The periodic part can be efficiently described as a sum of sinusoids with slowly varying amplitudes and frequencies, and the residual part is assumed to be irregular noise signal. This representation was introduced in [1] and since then has been profoundly studied and significantly enhanced. The model provides good parameterization of both voiced and unvoiced frames and allows using different modification techniques for them. It insures effective and simple voice processing in frequency domain. However the crucial point there is accuracy of harmonic analysis. The harmonic part of the signal is specified by sets of harmonic parameters (amplitude, frequency and phase) for every instant of time. A number of methods have been proposed to estimate these parameters. The majority of analysis methods assume local stationarity of amplitude and frequency parameters within the analysis frame [2-3]. It makes the analysis procedure easier, but, on the other hand, degrades parameters estimation and periodic/residual separation accuracy.

Some good alternatives are methods that make estimation of instantaneous harmonic parameters. The notion of instantaneous frequency was introduced in [4,5], the estimation methods have been presented in [4-9]. The aim of the current investigation is to study applicability of the instantaneous harmonic analysis technique described in [8,9] to a real time voice processing system for making voice effects (such as pitch, timbre and time-scale modifications). The analysis method is based on narrow band filtering by analysis filters with closed form impulse response. It has been shown [8] that the analysis filters can be adjusted in accordance with pitch contour in order to get adequate estimate of high order harmonics with rapid frequency modulations. The technique presented in this work has the following improvements:

- Simplified closed form expressions for instantaneous parameters estimation;
- Pitch detection and smooth pitch contour estimation;
- Improved harmonic parameters estimation accuracy.

The analysed signal is separated into periodic and residual parts and then processed through modification techniques. Then the processed signal can be easily synthesized in time domain at the output of the system. The harmonic+noise representation significantly simplifies the processing stage.

As it is shown in the experimental section the combination of the proposed analysis, processing and synthesis techniques provides good quality of signal analysis, modification and reconstruction.

2. SYSTEM OVERVIEW

Voice processing system can be divided into three main functional blocks: harmonic analysis, processing and synthesis (fig.1).

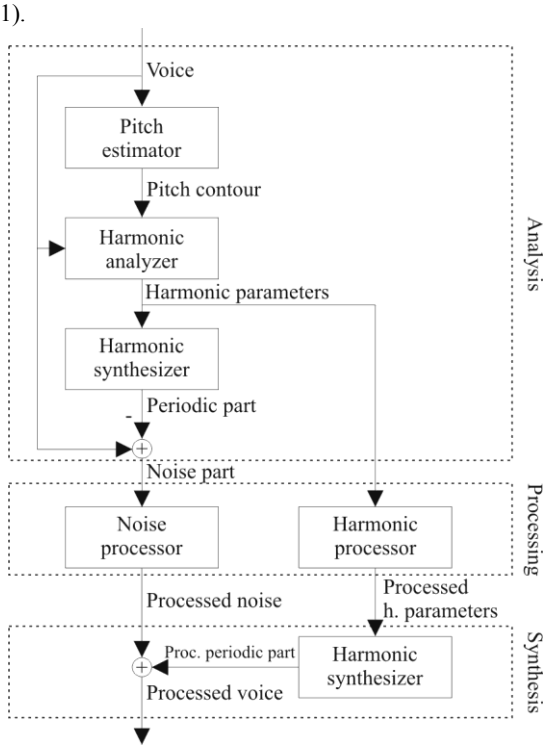


Figure 1: *General voice processing scheme.*

Harmonic analysis implies pitch and harmonic parameters estimation and periodic/noise separation of the source signal. In other words the periodic part of the source is presented in frequency domain as a set of sinusoidal parameters (amplitude, frequency, phase) apart from the noise part. All the vocal modifications are made in the processing block. Both the noise and periodical parts are processed separately. Noise can be processed in time domain while the periodical part is processed in harmonic domain that simplifies vocal effects implementation.
The output signal is the sum of the processed noise and the periodical part that is synthesized using processed harmonic parameters.

3. INSTANTANEOUS HARMONIC ANALYSIS

3.1. The Harmonic Model

The sinusoidal model assumes that the periodical part of the signal $s(n)$ can be expressed by the following formula:

$$s(n) = \sum_{k=1}^{K} \mathrm{MAG}_k(n)\cos\varphi_k(n), \quad (1)$$

where $\mathrm{MAG}_k(n)$ - the instantaneous magnitude of the k-th sinusoidal component, K is the number of components and $\varphi_k(n)$ is the instantaneous phase of the k-th component. Instantaneous phase $\varphi_k(n)$ and instantaneous frequency $f_k(n)$ are related as follows:

$$\varphi_k(n) = \sum_{i=0}^{n} \frac{2\pi f_k(i)}{F_s} + \varphi_k(0), \quad (2)$$

where F_s is sampling frequency and $\varphi_k(0)$ is the initial phase of the k-th component. The harmonic model states that frequencies $f_k(n)$ are integer multiples of fundamental frequency $f_0(n)$ and can be calculated as:

$$f_k(n) = kf_0(n). \quad (3)$$

Since voiced speech has harmonic structure the harmonic model is often used in speech coding. Accurate estimation of parameters $\mathrm{MAG}_k(n)$, $f_k(n)$ and $\varphi_k(0)$ is the fundamental problem of harmonic analysis. In the case of a monocomponent periodic signal a number of methods can provide good results [4,5]. However, multicomponent signals (like music or speech) are much more complex subjects for analysis and require special content and application-dependent methods. One of the most effective general approaches is to use adaptive filtering in order to pick out single components and then process them separately.

3.2. Analysis Filter

The proposed analysis method is based on the filtering technique that provides direct parameters estimation [8]. In voiced speech harmonic components are spaced in frequency domain and each component can be limited there by a narrow frequency band. Therefore harmonic components can be separated within the analysis frame by filters with non-overlapping bandwidths. These considerations point to the applicability and effectiveness of the filtering approach to harmonic analysis.

The analysis filter, used in this work has the following features [8]:

- Filtering in an arbitrary bandwidth;
- The impulse response is described by a closed form expression (a continuous function of bandwidth border frequencies);
- Estimation of the instantaneous parameters directly from the output signal;
- Implicit time warping (impulse response adjustment according to frequency modulations of pitch);
- Continuous and smooth contours of estimated parameters $\mathrm{MAG}_k(n)$ and $f_k(n)$.

The impulse response of the analysis filter $h(n)$ with pass band specified by center frequency $F_c(n)$ and half of the bandwidth F_Δ can be written in the following form:

$$h(n) = \begin{cases} 1, & n = 0 \\ \dfrac{\frac{F_s}{n\pi}\cos\left(\frac{2\pi}{F_s}\varphi_c(n,i)\right)\sin\left(\frac{2\pi n}{F_s}F_\Delta\right)}{2F_\Delta}, & n \neq 0 \end{cases} \quad (4)$$

where

$$\varphi_c(n,i) = \begin{cases} \sum_{j=n}^{i} F_c(j), & n < i \\ -\sum_{j=i}^{n} F_c(j), & n > i. \\ 0, & n = i \end{cases} \quad (5)$$

Filter output $s_{F_c,F_\Delta}(n)$ can be calculated as the convolution of the source signal $s(n)$ and $h(n)$, which can be expressed as the following sum:

$$s_{F_c,F_\Delta}(n) = \\ = \sum_{i=0}^{N-1} \frac{s(i)F_s}{2\pi(n-i)F_\Delta}\cos\left(\frac{2\pi}{F_s}\varphi_c(n,i)\right)\sin\left(\frac{2\pi(n-i)}{F_s}F_\Delta\right), \quad (6)$$

where N is filter order. The expression can be rewritten as a sum of zero frequency components:

$$s_{F_c,F_\Delta}(n) = A(n)\cos(0n) + B(n)\sin(0n), \quad (7)$$

where

$$A(n) = \\ = \sum_{i=0}^{N-1} \frac{s(i)F_s}{2\pi(n-i)F_\Delta}\sin\left(\frac{2\pi(n-i)}{F_s}F_\Delta\right)\cos\left(\frac{2\pi}{F_s}\varphi_c(n,i)\right), \quad (8)$$

$$B(n) = \\ = \sum_{i=0}^{N-1} \frac{-s(i)F_s}{2\pi(n-i)F_\Delta}\sin\left(\frac{2\pi(n-i)}{F_s}F_\Delta\right)\sin\left(\frac{2\pi}{F_s}\varphi_c(n,i)\right). \quad (9)$$

Thus, considering (7) – (9) the expression (6) is a magnitude and frequency-modulated cosine function:

$$s_{F_c,F_\Delta}(n) = \mathrm{MAG}(n)\cos\big(\varphi(n)\big). \quad (10)$$

Its instantaneous magnitude $\mathrm{MAG}(n)$, phase $\varphi(n)$ and frequency $f(n)$ can be calculated as:

$$\mathrm{MAG}(n) = \sqrt{A^2(n) + B^2(n)}, \quad (11)$$

$$\varphi(n) = \arctan\left(\frac{-B(n)}{A(n)}\right), \quad (12)$$

$$f(n) = \frac{\alpha(n+1) - \alpha(n)}{2\pi} F_s. \quad (13)$$

Filter output $s_{F_c,F_\Delta}(n)$ can be converted into analytical signal $s^a_{F_c,F_\Delta}(n)$ in the following way (i denotes the imaginary unit):

$$s^a_{F_c,F_\Delta}(n) = A(n) + iB(n). \quad (14)$$

Instantaneous sinusoidal parameters are available at every instant of time within the analysis frame. The bandwidth specified by $F_c(n)$ and F_Δ should cover the frequency of periodic component that is being analyzed. In fig.2 an example of parameters estimation is shown. The frequency contour of the harmonic component is covered by the filter pass band that is specified by the center frequency contour $F_c(n)$ and the bandwidth $2F_\Delta$.

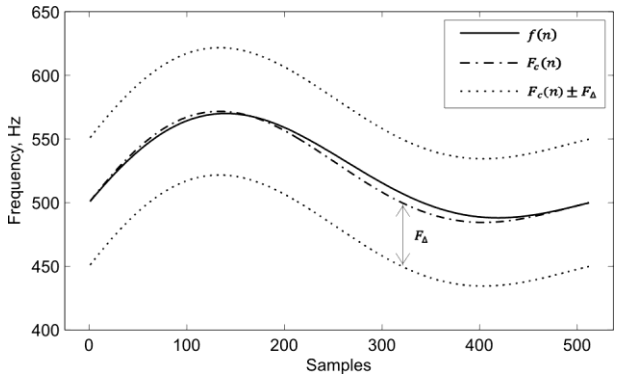


Figure 2: *Analysis filter ($N = 512$).*

Center frequency contour $F_c(n)$ is adjusted within the analysis frame providing narrow band filtering of frequency-modulated components.

3.3. Harmonic Parameters Estimation

It is assumed that voice frames can be either voiced or unvoiced. In voiced segments the periodical constituent prevails over the noise, in unvoiced segments the opposite takes place and therefore any harmonic analysis is unsuitable in that case. In the proposed analysis framework voiced/unvoiced frame classification is carried out using pitch detector. The harmonic parameters estimation procedure consists of the two following stages:

- Initial fundamental frequency contour estimation;
- Harmonic parameters estimation with fundamental frequency adjustment.

In voiced speech analysis, the problem of initial fundamental frequency estimation comes to finding a periodical component with the lowest possible frequency and sufficiently high energy. Within the possible fundamental frequency range (in this work it is defined as $[60,1000]$ Hz) all periodical components are extracted and then the suitable one is considered as the fundamental (fig.3). In order to reduce computational complexity the source signal is filtered by a low-pass filter before the estimation. The component extraction procedure involves iterative frequency recalculation with a predefined number of iterations. At every step the bandwidth of each filter is adjusted in accordance with the calculated frequency value in order to position energy peak in the centre of the band. At the initial stage the frequency range of the analyzed signal frame is covered by overlapping bandwidths $B_1, \dots, B_h$ (where h is the number of frequency bands) with central frequencies $F_C^{B_1}, \dots, F_C^{B_h}$ respectively.

At every step the respective instantaneous frequencies $f^{B_1}(n_c), \dots, f^{B_h}(n_c)$ are estimated at the instant that corresponds to the centre of the frame n_c. Then the central bandwidth frequencies are reset $F_C^{B_x} = f^{B_x}(n_c)$ before the next iteration. After all energy peaks are located the final sinusoidal parameters are estimated. Amplitude, frequency and phase are calculated using expressions (11) – (13). During the iterative energy peak location process some of the filter bands may locate the same component. Duplicated parameters are discarded by comparison of the centre band frequencies of $F_C^{B_1}, \dots, F_C^{B_h}$. Fundamental frequency contour should be long enough. To avoid estimation errors that may be caused by short-term components (that apparently are transients or noise and should be taken to residual) parameters are tracked from frame to frame. The frequency and amplitude values of adjacent frames are compared, providing adequate harmonic components estimation. If it is not possible to find a proper continuous pitch contour the corresponding frames are classified as unvoiced.

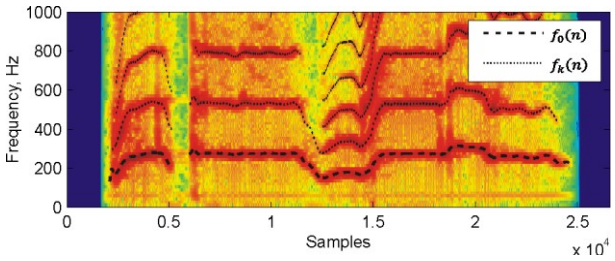


Figure 3: *Pitch estimation ($F_s = 8kHz$).*

Having fundamental contour estimated it is possible to calculate filter impulse responses aligned to the fundamental frequency contour. Central frequency of the filter band is calculated as the instantaneous frequency of fundamental multiplied by the number k of the correspondent harmonic $F_C^k(n) = kf_0(n)$. The procedure goes from the first harmonic to the last, adjusting fundamental frequency at every step. The fundamental frequency recalculation formula can be written as follows:

$$f_0(n) = \sum_{i=0}^{k} \frac{f_i(n)\mathrm{MAG}_i(n)}{(i+1)\sum_{j=0}^{k}\mathrm{MAG}_j(n)}. \quad (15)$$

The fundamental frequency values become more precise while moving up the frequency range. It allows making proper analysis of high order harmonics with significant frequency modulations. Harmonic parameters are estimated using expressions (11) – (13). After parameters estimation the periodical part of the signal is synthesized by formula (1) and subtracted from the source in order to get the noise part.

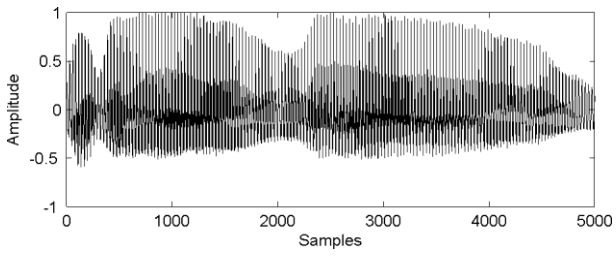
4. VOICE MODIFICATIONS

The harmonic analysis described in the previous section results in a set of harmonic parameters and residual signal that are the in-

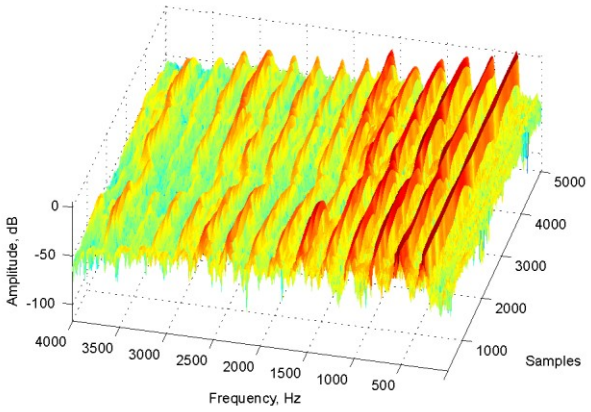
puts of voice processing block of the system. Many voice processing techniques require pitch and spectral envelope estimation in order to modify and synthesize voiced frames.

4.1. Spectral envelopes estimation

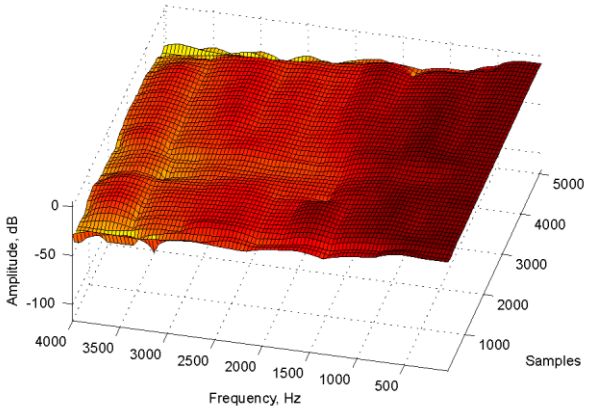
Instantaneous spectral envelopes can be estimated from the instantaneous harmonic amplitudes and the fundamental frequency obtained at the analysis stage. The linear interpolation can be used for this purpose.



a)



b)



c)

Figure 4: *Spectral envelopes estimation: a) source signal; b) spectrogram; c) spectral envelopes*

In fig.4. spectral envelopes estimation is illustrated. The source voiced segment (fig.4.a,b) is analyzed by the harmonic analysis technique and then the instantaneous spectral envelopes are interpolated from the obtained amplitude values and pitch contour (fig.4.c). The set of frequency envelopes can be considered as a function $E(n, f)$ of two parameters (sample and frequency).

4.2. Pitch shifting

Pitch shifting procedure affects only the periodic part of the signal that can be synthesized as follows:

$$s(n) = \sum_{k=1}^{K} E\left(n, \bar{f}_k(n)\right) \cos\bar{\varphi}_k(n). \qquad (16)$$

Phases of harmonic components $\bar{\varphi}_k(n)$ are calculated according to a new fundamental frequency contour $\bar{f}_0(n)$:

$$\bar{\varphi}_k(n) = \sum_{i=0}^{n} \frac{2\pi \bar{f}_k(i)}{F_s} + \bar{\varphi}_k^{\Delta}(n). \qquad (17)$$

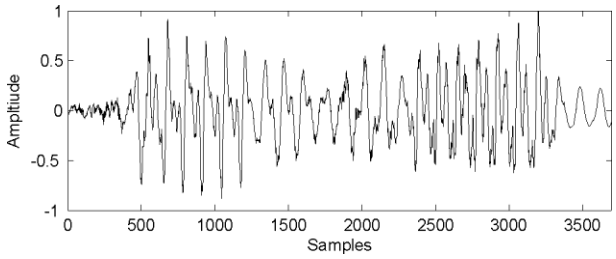
Harmonic frequencies are calculated by formula (3):

$$\bar{f}_k(n) = k\bar{f}_0(n). \qquad (18)$$

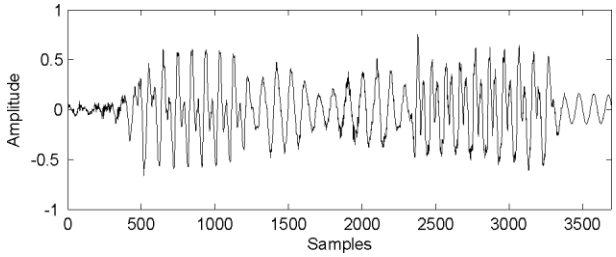
Additional phase parameter $\bar{\varphi}_k^{\Delta}(n)$ is used in order to keep the original phases of harmonics relative phase of the fundamental:

$$\bar{\varphi}_k^{\Delta}(n) = \varphi_k(n) - k\varphi_0(n). \qquad (19)$$

In fig.5 is presented a result of pitch shifting. The source signal is a recorded female voice sampled at 22,05kHz (fig.5.a). The fundamental of the source signal (150-180 Hz) was raised through analysis-processing-synthesis procedure to 200-250 Hz (fig.5.b).



a)



b)

Figure 5: *Pitch shifting: a) source signal; b) synthesized signal with increased pitch*

As long as described pitch shifting does not change spectral envelope of the source signal and keeps relative phases of the harmonic components, the processed signal has a natural sound with completely new intonation.

4.3. Timbre modifications

The timbre of the voice is defined by the spectral envelope function $E(n,f)$. If we consider the envelope function as a matrix

$$E = \begin{pmatrix} E(0,0) & \cdots & E\left(0,\frac{F_s}{2}\right) \\ \vdots & \ddots & \vdots \\ E(N,0) & \cdots & E\left(N,\frac{F_s}{2}\right) \end{pmatrix}, \quad (20)$$

then any timbre modification can be expressed as a conversion function $C(E)$ that transforms the source envelope matrix E into a new matrix $\bar{E}$:

$$\bar{E} = C(E). \quad (21)$$

The conversion function should be chosen in accordance with the target application of the system. In voice conversion systems estimation of $C(E)$ is usually based on a codebook that is formed by training procedure. The recordings of both source and target speakers are analyzed simultaneously in order to build up correspondence between spectral envelopes of them.
In this subsection a simple training and conversion technique is described. The training sets E_s and E_t are the spectral envelope matrixes estimated from the source and target recordings respectively. The recordings have the same content and are synchronized in time domain. Let us assume that the conversion function $C(E)$ can be written in the following matrix form:

$$C(E) = EK. \quad (22)$$

Then in order to make conversion from E_s to E_t the matrix K should minimize the conversion error R:

$$R = \min_K |E_s K - E_t|. \quad (23)$$

The training procedure can be easily implemented via solution of the corresponding system of linear equations.
Thus the conversion codebook is the matrix K and the converted spectral envelope can be simply calculated as:

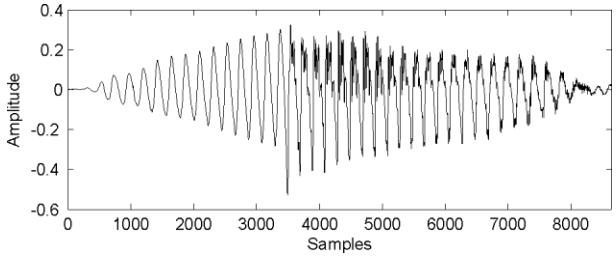
$$\bar{E} = EK. \quad (24)$$

Experiments have shown that this conversion approach can be very efficient, provided that sets E_s and E_t are long enough and accurately synchronized.

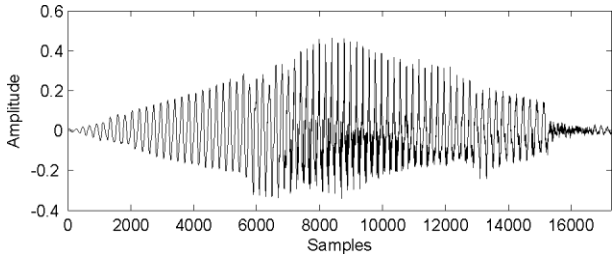
4.4. Time-scale modifications

Since the periodic part of the signal is expressed by harmonic parameters it is easy to synthesize the periodic part slowing down or stepping up the tempo. Amplitude and frequency contours should be interpolated in the respective moments of time and then the output signal can be synthesized using expressions (1) or (16).
The noise part is parameterized by spectral envelopes and then time-scaled as described in [10].

Separate periodic/noise processing provides high quality time-scale modifications without audible artifacts. In fig.6. a time-scale modification result is presented. The synthesized signal is slowed down two times.



a)



b)

Figure 6: *Time-scale modification: a) source signal; b) slowed down signal*

5. PITCH MODIFICATION EXPERIMENT

In this section an example of vocal processing is shown. The concerned processing system is aimed at pitch shifting in order to assist a singer in real-time.

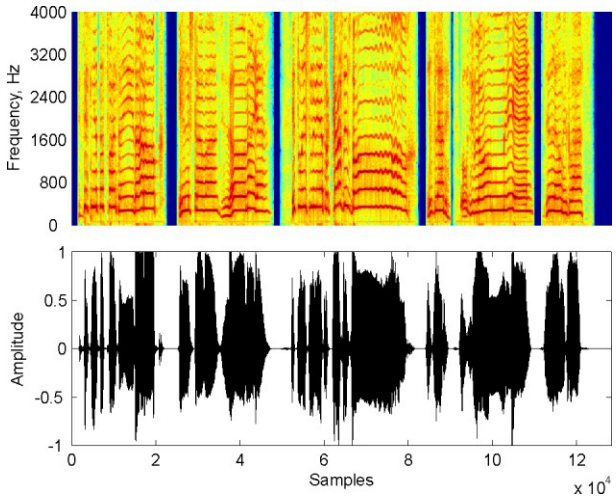


Figure 7: *Reference signal.*

The voice of the singer is analyzed by the proposed technique and then synthesized with pitch modifications to assist the singer to be in tune with the accompaniment. The target pitch contour is predefined by analysis of a reference recording. Since only pitch contour is changed the source voice maintains its identity. The output signal however is damped in regions where the energy of the reference signal is low in order to provide proper synchronization with accompaniment. The reference signal is shown in fig.7, it is a recorded male vocal. The recording was made in a studio with a low level of background noise. The fundamental frequency contour was estimated from the reference signal as described in section 3. As can be seen from fig.8 the source vocal has different pitch and is not completely noise free (it was recorded in conditions that are close to a real working environment of the system).

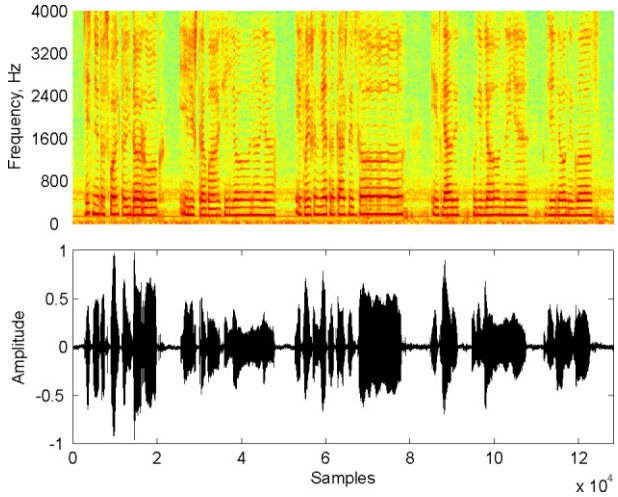


Figure 8: *Source signal.*

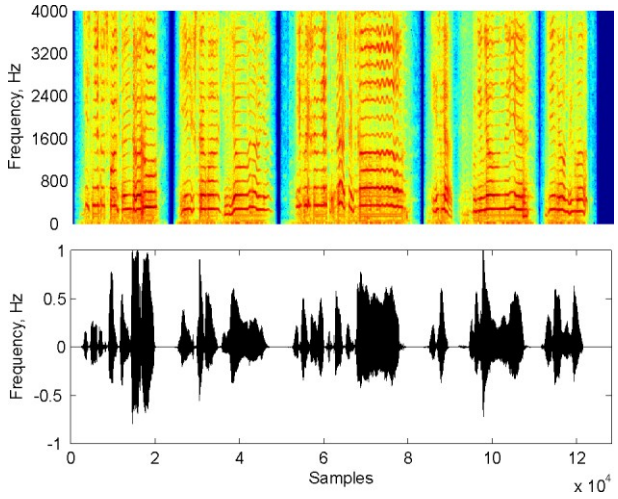


Figure 9: *Processed signal.*

The source signal was analyzed using proposed harmonic analysis and then the pitch shifting technique was applied as described in 4.2. The synthesized signal with pitch modifications is shown in fig.9. As can be seen the output signal contains the pitch contour of the reference signal, but still has timbre and energy of the source voice. The noise part of the source signal (including background noise) remained intact.

6. CONCLUSIONS

The harmonic+noise model can be applied to voice processing systems. It provides efficient signal parameterization in the way that is quite convenient for making voice effects such as pitch shifting, timbre and time-scale modifications. The practical application of the proposed harmonic analysis technique has shown encouraging results. The described approach might be a promising solution to harmonic parameters estimation in voice processing systems.

7. ACKNOWLEDGMENTS

This work was supported by the Belarusian republican fund for fundamental research under the grant T08MC-040 and the Belarusian Ministry of Education under the grant 09-3102.

8. REFERENCES

[1] R.J. McAulay, T.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation" *IEEE Trans. On Acoustics, Speech and Signal Process.*, vol. 34, no. 4, pp.744-754, 1986.

[2] A.S. Spanias, „Speech coding: a tutorial review", *Proc. of the IEEE,* vol. 82, no. 10, pp. 1541-1582, 1994.

[3] X.Serra, "Musical Sound Modeling with Sinusoids plus Noise" in *Musical Signal Processing* (C. Roads, S. Pope, A. Picialli, and G. De Poli eds.), Swets & Zeitlinger Publishers, 1997, pp. 91-122.

[4] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal", *Proceedings of the IEEE,* vol. 80, 4, pp. 520-568, 1992.

[5] Maragos P., Kaiser J. F., Quatieri T. F., "Energy Separation in Signal Modulations with Application to Speech Analysis", *IEEE Trans. on Signal Process.*, vol. 41, no. 10, pp. 3024-3051, 1993.

[6] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in *Proc. ICASSP*, 1995, pp. 756–759.

[7] T. Abe, M. Honda, "Sinusoidal model based on instantaneous frequency attractors", *IEEE Trans. on Audio, Speech, and Language processing*, vol. 14, no. 4, pp. 1292-1300, July 2006.

[8] E. Azarov, A. Petrovsky, M. Parfieniuk, "Estimation of the instantaneous harmonic parameters of speech", in *Proc. of the 16th European Signal Process. Conf., (EUSIPCO-2008)*, Lausanne, 2008.

[9] Azarov I., Petrovsky A, "Harmonic analysis of speech", *Speech technology*, no. 1, pp. 67-77, Moscow, 2008, (in Russian).

[10] Levine S., Smith J., "A Sines+Transients+Noise Audio Representation for Data Compression and Time/Pitch Scale Modifications", *AES 105th Convention* (San Francisco, CA, USA), Preprint 4781, September 1998.