VOICE TRANSFORMATIONS THROUGH INSTANTANEOUS COMPLEX FREQUENCY MODIFICATIONS

Magdalena Kaniewska

Multimedia Systems Department, Gdansk University of Technology
11/12 Narutowicza St., 80-233 Gdansk, Poland
phone: + 48 59 347 19 67, fax: +48 58 347 11 14, email: emka@sound.eti.pg.gda.pl
web: www.multimed.org

ABSTRACT

The paper presents the possibilities of altering human voice by modifying instantaneous complex frequency (ICF) of the speech signal. The proposed algorithm is based on the factorization of a band-limited, analytic signal into two factors: one fully characterized by its envelope and the other with positive instantaneous frequency (PIF). ICFs of both factors are estimated and modified for different sound effects. The algorithm is tested on short speech utterances and the results of these experiments are presented in the paper.

1. INTRODUCTION

The aim of presented in the paper work was to alter human voice by modifying instantaneous complex frequency (ICF) of the speech signal. ICF is a concept defined by Hahn [4] and based on the concept of instantaneous frequency (IF), which was firstly introduced with respect to FM modulations, widely used in telecommunications. Later, however, it became popular also in other applications, e.g. biomedical engineering, seismology, radiolocation, oceanography, underwater acoustics or sound processing. The main reason for applying IF in these fields is the nonstationarity of the signals used there. It is also the main reason for using IF in speech processing. The use of IF for speech analysis has been already widely investigated, e.g. in [2-3], [9-10], [12]. It was mainly used for estimating pitch as well as formants' frequencies of speech. There have also been attempts of using IF for changing human voice [13], [15]. However, all the algorithms, that can be found in literature require processing speech in sub-bands, which considerably increases their complexity. The method described in this paper overcomes this problem. It is simple and fast and therefore can be used in real-time applications.

The algorithm of voice transformation is based on ICF modification. Before ICF estimation and modification analytic representation of real signal is obtained and a bi-factorization algorithm is applied. The two obtained factors are minimum-phase envelope and phase signal with positive instantaneous frequency . ICFs of both factors are estimated and modified in order to obtain different sound effects. The proposed modifications are all simple operations that introduce minimum delay. The entire algorithm can be reversed, i.e. one can obtain the signal from its ICF, multiplying the factors of bi-factorization restores the original analytic signal and then the real-valued signal can be calculated as the real part of the analytic signal. Of course after modifying ICF one obtains a transformed, different sounding voice.

The details of the algorithm are presented in the next Section along with the description of the proposed modifications. Section III presents the results of experiments conducted for the recordings of polish vowels and short sentences.

2. ALGORITHM DETAILS

The main blocks of the proposed algorithm are signal bi-factorization, ICF estimation and ICF modification. Before that, the complex representation of the real speech signal has to be computed. If we consider an arbitrarily modulated AM-FM signal $x(t) = a(t)\cos\varphi(t)$ with instantaneous amplitude $a(t)$, and instantaneous phase $\varphi(t)$, its analytic equivalent $u(t)$ is

$$u(t) = x(t) + jH_T\{x(t)\} = a(t)e^{j\varphi(t)} \tag{1}$$

where $H_T\{\ \}$ stands for the ideal Hilbert transformer [4]. For the purposes of this algorithm a complex Hilbert filter is used instead of the Hilbert transformer. This will be explained further in the paper.

2.1 Signal bi-factorization

The second block of the algorithm is the signal bi-factorization. This concept takes its origin in the theory of spectral factorization described by Oppenheim et al. [11].

Taking advantage of the time-frequency dualism we can use the concept of spectral factorization to factorize a band-limited analytic signal $u(t)$. As the result of the bi-factorization we obtain two also analytic signals: minimum-phase and all-phase ones.

$$u(t) = a_{mp}(t)\gamma_{pif}(t) \tag{3}$$

where $a_{mp}(t)$ is the minimum-phase signal and $\gamma_{pif}(t)$ is a phase signal with positive IF, called the all-phase signal

$$a_{mp}(t) = a(t)\exp(j\varphi_{mp}(t)) \quad (4)$$

$$\gamma_{pif}(t) = \exp(j\varphi_{pif}(t)) \quad (5)$$

In (4) and (5) $\varphi_{mp}(t)$, $\varphi_{pif}(t)$ are the instantaneous phases of $a_{mp}(t)$ and $\gamma_{pif}(t)$ respectively. Instantaneous amplitude of $a_{mp}(t)$ is equal to the instantaneous amplitude of $u(t)$ and $\varphi_{mp}(t)$ and $\ln a(t)$ are a pair of Hilbert transforms. In order to compute both factors we first have to compute $\varphi_{mp}(t)$ as

$$\varphi_{mp}(t) = H_T\{\ln a(t)\} \quad (6)$$

Having that, $\gamma_{pif}(t)$ is obtained simply by dividing $u(t)$ by $a_{mp}(t)$. Bi-factorization is described in detail in [8].

Factorization of speech signal into minimum-phase and all-phase factors was already proposed by Kumaresan and Rao [9]. However, their method of computing factors is not based on their definitions. Instead, linear prediction in spectral domain is used and speech is processed in sub-bands.

2.2 ICF estimation

ICF is based on the concept of IF. The most widely used and accepted definition of IF was presented by Gabor (cited in [1]), followed later by Ville (cited in [1]). They defined IF of a real signal, $x(t) = a(t)\cos\varphi(t)$ as

$$f(t) = \frac{1}{2\pi}\frac{d}{dt}\arg u(t) = \frac{1}{2\pi}\frac{d\varphi}{dt} \quad (7)$$

One can also write the definition of instantaneous angular frequency

$$\omega(t) = \frac{d\varphi(t)}{dt} \quad (8)$$

Analogically, Hahn defined ICF of a complex signal $u(t)$ as the first derivative of the instantaneous complex phase $p(t)$. The definition of $p(t)$ given by Hahn [4] is

$$p(t) = \ln u(t) \quad (9)$$

The imaginary part of instantaneous complex phase is the instantaneous phase $\varphi(t)$, while the real part is the log-envelope of $u(t)$, $\lambda(t) = \ln|u(t)|$. ICF is then given by

$$s(t) = \frac{dp(t)}{dt} = \frac{du(t)}{dt}\frac{1}{u(t)} \quad (10)$$

If we investigate the real part, Re( ), of ICF we will see that

$$\mathrm{Re}(s(t)) = \frac{a'(t)}{a(t)} = \sigma(t) \quad (11)$$

It is a measure of the relative speed of the changes of $a(t)$ [4]. The imaginary part, Im( ), of ICF is

$$\mathrm{Im}(s(t)) = \frac{d}{dt}\arg u(t) = \omega(t) \quad (12)$$

This gives us

$$s(t) = \sigma(t) + j\omega(t) \quad (13)$$

a generalization of the time-independent complex frequency $s = \sigma + j\omega$ known from the circuit and signal theory and used to define the Laplace transformation [4]. Since in practice we deal with discrete-time signals, an estimation of discrete ICF is needed. For computational purposes, a good estimation is [14]

$$s[n] = \mathrm{Ln}\frac{u[n]}{u[n-1]} \quad (14)$$

The accuracy of this estimation depends on the sampling rate. The more the signal is oversampled, the better estimation we obtain. Using (14) we can estimate ICFs of discrete-time minimum-phase and all-phase signals. Knowing that:

1) $u[n] = a_{mp}[n]\gamma_{pif}[n]$
2) $|u[n]| = |a_{mp}[n]| = a[n]$
3) $|\gamma_{pif}[n]| = 1$

and considering (11)-(13) we can write that

$$s[n] = s_{mp}[n] + s_{pif}[n] \quad (15)$$

$$s_{mp}[n] = \sigma[n] + j\omega_{mp}[n] \quad (16)$$

$$s_{pif}[n] = j\omega_{pif}[n] \quad (17)$$

$$\omega[n] = \omega_{mp}[n] + \omega_{pif}[n] \quad (18)$$

where $s_{mp}[n]$ and $s_{pif}[n]$ are ICFs of $a_{mp}[n]$ and $\gamma_{pif}[n]$ respectively, and $\omega_{mp}[n]$ and $\omega_{pif}[n]$ are their IFs. It is also worth emphasizing here that $\omega_{pif}[n]$ is always positive.

2.3 Proposed modifications

Four ICF modifications are proposed for altering the human voice [6]. All of them are simple operations of adding/subtracting or multiplying/dividing. When choosing the parameters of modifications it is important to limit their ranges in a manner that the spectrum of the obtained signal is not shifted to the negative frequencies and does not exceed the range of human hearing. The proposed modifications are:

1) Scaling ICF of the all-phase signal (multiplying or dividing $s_{pif}[n]$ by a constant value c)

2) Shifting ICF of the all-phase signal (adding vector $\omega_c[n]$ to $\omega_{pif}[n]$, only the imaginary part of ICF can be shifted since the real part equals zero and it should stay this way)

3) Scaling real part of ICF of minimum-phase signal (multiplying $\sigma[n]$ by c)

4) Scaling imaginary part of ICF of minimum-phase signal (multiplying $\omega_{mp}[n]$ by c)

After modifying ICF a new speech signal is synthesized by computing new minimum-phase and all-phase signals from their ICFs, using the inverse of formula (14), and then multiplying them [6].

3. EXPERIMENTS

Figure 1 presents the block diagram of the proposed voice modification algorithm.

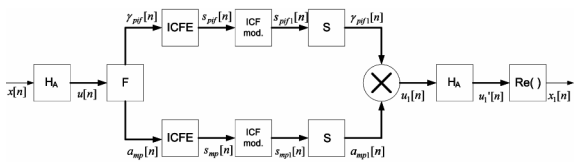

Figure 1 – Block diagram of the modification algorithm.

The consecutive blocks of the diagram are:

- H_A – complex Hilbert filter
- F – factorization block;
- ICFE – ICF estimator;
- ICF mod. – ICF modification block;
- S – modified signal synthesis;
- Re( ) – real part;

Complex Hilbert filter is used for computing the complex representation of the real signal. $u[n]$ is computed as a convolution of $x[n]$ with the impulse response of a bandpass complex Hilbert filter $h_A[n]$

$$u[n] = h_A[n] * x[n] \tag{19}$$

This filter limits the band of the signal to a desired range, 200-8000 Hz, so that all significant formants are maintained. It is also applied at the end of the algorithm to the modified signal, since some of the proposed modifications may produce a non-analytic signal. Using the Hilbert filter once again guarantees that the signal obtained after modifications is analytic.

The proposed algorithm was tested for speech recordings – polish vowels as well as short sentences spoken by men and women. Sampling frequency was 48 kHz.

3.1 Analysis

Figure 2 presents the results of bi-factorization and ICF estimation for vowel /a/. In the first column waveforms of the analytic equivalent of the speech signal (real part), as well as its minimum-phase and all-phase factors (real and imaginary parts) are plotted. The second column contains periodograms of the above mentioned signals. In the third column waveforms of IFs of the signals are plotted.

Looking at the periodograms we can see that the minimum-phase signal maintains in a certain degree the formant structure of the original signal. The spectra of $u[n]$ and $a_{mp}[n]$ are similar - the more minimum-phase $u[n]$ is, the more similar the spectra are. For pure minimum-phase signal they are identical [8], but the spectrum of $s_{mp}[n]$ is shifted to zero on the frequency axis. The spectrum of all-phase signal is cumulated around 400-1000 Hz. It is also the range of power accumulation in $u[n]$ periodogram. The more minimum-phase $u[n]$ is, the narrower spectrum $\gamma_{pif}[n]$ has. IF of $u[n]$, $\omega[n]$, is the sum of $\omega_{mp}[n]$ and $\omega_{pif}[n]$. As it can be seen, the mean value of $\omega_{mp}[n]$ is zero, while the mean values of $\omega[n]$ and $\omega_{pif}[n]$ are equal (about 570 Hz). The IF of all-phase signal is smooth and reflects the trend in $\omega[n]$ waveform.

3.2 Modification results

In this paragraph changes of speech signal obtained by applying proposed modifications are shortly described. More detailed description can be found in [6]. To illustrate the obtained results, plots for six different modifications applied to a polish sentence spoken by a male are presented in figure 3. It can be seen that the proposed modifications do not change temporal structure of the sentence (the modified sentences are in synchronization with the original one). We can also see that trajectories of fundamental frequency do not change after the modifications (although, as it will be further noticed, the perceived pitch does change), so the intonation of the sentence remains unchanged.

3.2.1 Scaling ICF of all-phase signal

Scaling ICF of all-phase signal results foremost in the change of mean value of IF. As a consequence, signal's spectrum is shifted on the frequency axis - to the right, if the scaling factor is greater than 1, or to the left if it is less than 1. Moreover, the higher the factor is, the less distinct the formants are. For $c > 2.5$ the formant structure practically disappears and the intelligibility of speech is considerably affected.

The change of voice is mainly in its pitch. One can hear that the pitch is higher or lower for $c > 1$ and $c < 1$ respectively. The timbre of the voice also changes. However, the higher c is (or closer to zero for $c < 1$) the less natural the voice is.

3.2.2 Shifting IF of all-phase signal

Shifting IF of all-phase signal results only in the change of the mean value of IF of the speech signal and in the shift of the speech spectrum on the frequency axis, to the right or to the left, for $\omega_c[n]$ grater or less than zero respectively. The

best results of voice change are obtained for $\omega_c[n]$ equal to the multiple of fundamental frequency of speech signal. The pitch is then simply higher or lower. Therefore, a pitch estimation algorithm described by Kaniewska [7] was added. It is a real-time, pipeline algorithm. Pitch is estimated for every sample of speech. The obtained trajectory of fundamental frequency is then used for shifting $\omega_{pif}[n]$.

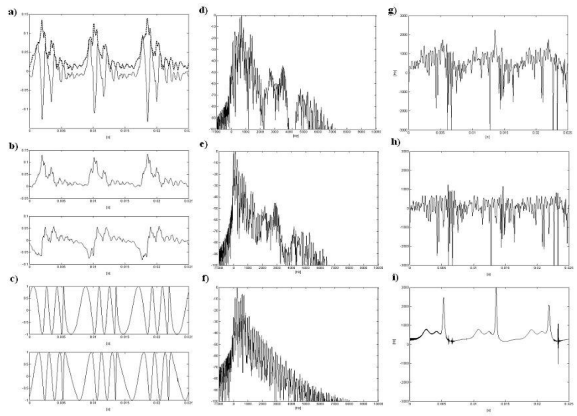

Figure 2 –Bi-factorization of vowel /a/ (3 periods): waveforms (a-c) and periodograms (d-f) of $u[n]$, $a_{mp}[n]$ and $\gamma_{pif}[n]$; waveforms of $\omega[n]$ (g), $\omega_{mp}[n]$ (h) and $\omega_{pif}[n]$ (i).

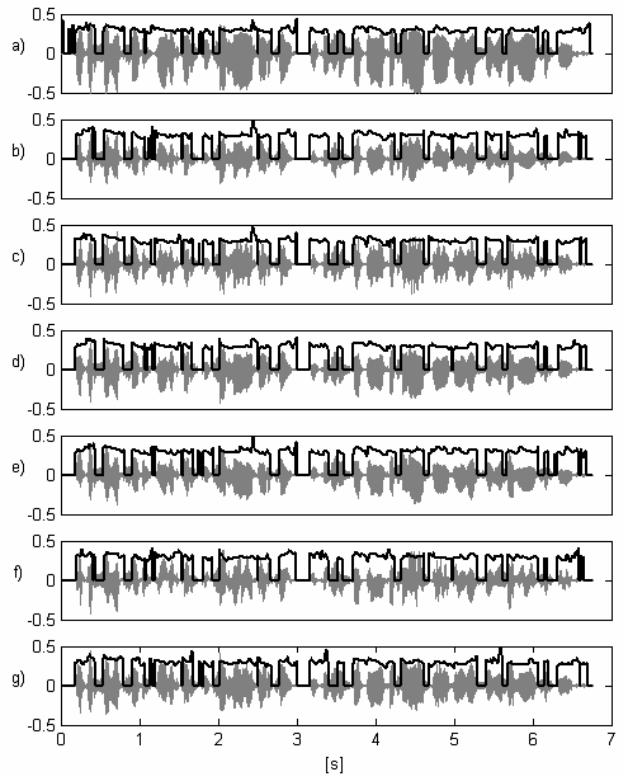

Figure 3 – Speech waveforms (grey line) and estimated pitch (black line) for original speech (a), speech modified by scaling IF of all-phase signal with c=1.05 (b) and c=0.95 (c), speech modified by scaling IF of minimum-phase signal with c=1.5 (d) and c=0.5 (e), speech modified by shifting down (f) and up (g) IF of all-phase signal by estimated fundamental frequency.

3.2.3 Scaling real part of ICF of minimum-phase signal

By scaling $\sigma[n]$ we change only the signals envelope. This change is nonlinear, since the envelope of the modified signal is the original envelope raised to the power of *c*. Linear modifications can be applied directly on $a[n]$ [6]. We can observe two aspects of this modification: change in the sound's level and in its dynamic range. For $c > 1$ the dynamic range (understood as $\lambda_{\max} - \lambda_{\min}$) is bigger, for $c < 1$ it is smaller. For *c* close to 3 the compression is so high that the quietest sounds become inaudible, which influences not only the sound quality, but also the speech intelligibility.

3.2.4 Scaling IF of minimum-phase signal

This modification does not cause the change of the mean value of the speech signal's IF, therefore there is no shift of the signal's spectrum. However, the formant structure of the speech signal changes. For smaller *c* further formants are attenuated. One can also set *c* to a negative value, but for $c < -1$ higher frequencies are excessively attenuated.

The changes of voice are more delicate than for the modification of $\omega_{pif}[n]$. Only the timbre changes, the pitch remains unchanged. The most natural sounding voice is obtained for $c < 3$. For higher *c* the voice becomes cartoon-like and for $c > 5$ the intelligibility is considerably affected.

3.2.5 Combinations of modifications

The simple modifications described above can be combined for obtaining different voice changes simultaneously. First of all one can combine $\omega_{pif}[n]$ scaling and shifting. Another possible combination is the simultaneous change of $\omega_{pif}[n]$ and $\omega_{mp}[n]$. In the case when $\omega_{pif}[n]$ is increased by scaling or shifting (or both) we can scale $\omega_{mp}[n]$ with factor $c < 1$ in order to attenuate high frequencies and make the transformed voice more natural. Scaling $\omega_{mp}[n]$ with factor $c > 1$ makes the sound more bright so we can use it when we decrease $\omega_{pif}[n]$.

3.3 Listening tests

The results of the speech modifications were evaluated in listening tests. The modifications chosen for the tests were: scaling imaginary part of ICF of minimum-phase signal, scaling imaginary part of ICF of all-phase signal, shifting imaginary part of ICF of all-phase signal, combination of scaling and shifting imaginary part of ICF of all-phase signal and combination of shifting imaginary part of ICF of all-phase signal and scaling imaginary part of ICF of minimum-phase signal. The modifications were applied to two male and two female voices. The evaluated set contained 20 recordings, original voices were among them. They were played in random order. The participants of the test listened to the set twice. While the first listening, they were to evaluate the naturalness of the voices on a scale form 1 to 5 with 0.5 step (5 being a completely natural voice). During second listening they were to evaluate the overall quality of

the heard speech. No parameters were defined for this test, the listeners could freely decide what they understand by "quality". The scale was the same as in the first step, 5 being the best quality.

The first conclusion drawn from the tests for speech naturalness was that voices with lower pitch were evaluated higher. The mean score for modified male voices was for no modification lower than 3, while the lowest mean score for a female voice modification was 1.7.

The highest mean scores were given to voices modified by scaling imaginary part of ICF of minimum-phase signal (they varied from 3.5 to 5, higher for a lower scaling factor). However, these modifications introduced the most delicate changes to human voice (the speaker is fully recognizable). The same concerned modifications obtained by scaling imaginary part of ICF of all-phase signal by $0.9 < c < 1.1$.

The modification that gave more significant voice changes was shifting imaginary part of ICF of all-phase signal by value equal to fundamental frequency. The mean score for these modifications varied from 2 to 4.5 when shifted up and from 1.7 to 3.5 when shifted down. Higher score was obtained when shifting was combined with scaling imaginary part of ICF of minimum-phase signal or imaginary part of ICF of all-phase signal (from 2.5 to 4.5, higher for male voices). Further improvements can also be achieved when only voiced parts of speech are modified. The voiced/unvoiced classification can be done sample-by-sample, based on the value of lowpass filtered $\omega_{pif}[n]$ (this value is higher for unvoiced parts). The results of this improvement, however, were not yet subjected to listening tests.

The second stage of tests showed that for about 75% of listeners naturalness of voice is correlated with the perceived speech quality, i.e. the recordings that were evaluated as not natural got also lower score in quality test. The quality was also evaluated lower when the factor of scaling imaginary part of ICF of minimum-phase signal was higher than 2 (this modification introduced audible clicking).

The results of the tests showed that shifting imaginary part of ICF of all-phase signal by value equal to fundamental frequency combined with scaling imaginary part of ICF of minimum-phase signal or imaginary part of ICF of all-phase signal can be used to significantly change human voice and maintain, in high degree, its naturalness. This modification could be used in speaker depersonalization algorithms.

4. CONCLUSIONS

It was shown in the paper that ICF can be used for changing human voice, its pitch and timbre. Moreover factorizing speech signal into minimum-phase and all-phase factors and then modifying ICFs of the factors individually produces different sound effects. The algorithm cannot be used for imitating a certain target speaker. Further research will include using the most natural sounding modifications for speaker depersonalization.

The advantage of the proposed method over other algorithms that use IF is that it is fast and can operate in a sample-by-sample mode. Only Hilbert filters introduce delay, but it can be minimized when the filters are optimally designed. This is a great advantage if the algorithm is used in real-time applications, e.g. speaker depersonalization.

REFERENCES

[1] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal – part 1: fundamentals", *Proceedings to the IEEE*, vol. 80, pp. 520-538, 1992.

[2] F. Gianfelici et al., "AM-FM Decomposition of Speech Signals: An Asymptotically Exact Approach Based on the Iterated Hilbert Transform", in Proc. *13th Workshop on Statistical Signal Processing*, Bordeaux, France, 2005.

[3] M. Grimaldi and F. Cummins, Speaker identification using instantaneous frequencies, *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 6, pp. 1097-1111, 2008.

[4] S.L. Hahn, *Hilbert Transforms in Signal Processing*, Artech House, 1995.

[5] E. Hermanowicz and M. Rojewski, "Pitch Shifter Based on Complex Dynamic Representation Rescaling and Direct Digital Synthesis", *Bulletin of The Polish Academy of Sciences*, vol. 54, no.4, pp. 499-504, 2006.

[6] M. Kaniewska, "Human voice modification using instantaneous complex frequency", *in Proc. 128th AES Convention*, London, May 2010.

[7] M. Kaniewska, "Instantaneous complex frequency for pipeline pitch estimation", unpublished.

[8] M. Kaniewska, "On the use of instantaneous complex frequency for analysis and modification of simple sounds", *Ph. D. Research in Microelectronics and Electronics*, pp. 340-343, 2009.

[9] B. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications", *J. Acoust. Soc. Am.*, vol. 105, pp. 1912-1924, 1999.

[9] P. Maragos and A. Potamianos, "Speech Formant Frequency and Bandwidth Tracking Using Multiband Energy Demodulation", *J. Acoust. Soc. Am.*, vol. 99, pp. 3795–3806, 1996.

[11] A.V. Oppenheim, R.W. Schafer and J.R. Buck J.R., *Discrete-time signal processing*, Prentice Hall, 1989.

[12] K.K. Paliwal and B.S. Atal, "Frequency-related representation of speech", in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 2003.

[13] A. Potamianos, *Speech processing Applications using an AM-FM modulation model*, PhD. Thesis, Harvard University, 1995.

[14] M. Rojewski, „Nowa definicja i bezbłędna estymacja dyskretnej zespolonej pulsacji chwilowej", *X Krajowe Sympozjum Telekomunikacji*, Bydgoszcz, 1994.

[15] J. Timoney and T. Lysaght, "EPS models of AM-FM vocoder output for new sounds generations", in *Proc. COST G-6 Conf. Digital Audio Effects (DAFX-01)*, Limerick, Ireland, 2001.