14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

THEORETICAL AND EXPERIMENTAL BASES OF A NEW METHOD FOR ACCURATE SEPARATION OF HARMONIC AND NOISE COMPONENTS OF SPEECH SIGNALS

Laurent GIRIN

ICP (Speech Communication Lab.), INPG/Univ. Stendhal/CNRS UMR 5009
46 av. Félix Viallet, 38040 Grenoble, France
Phone: +33 476 57 47 15, fax: +33 476 57 47 10, email : girin@icp.inpg.fr
Web : http://www.icp.inpg.fr/~girin/

ABSTRACT

In this paper, the problem of separating the harmonic and aperiodic (noise) components of speech signals is addressed. A new method is proposed, based on two specific processes dedicated to better take into account the non-stationarity of speech signals: first, a period-scaled synchronous analysis of spectral parameters (amplitudes and phases) is done, referring to the Fourier series expansion of the signal, as opposed here to the typically used Short-Term Fourier Transform (STFT). Second, the separation itself is based on a low-pass time-filtering of the parameters trajectory. Additionally to presenting the theoretical basis of the method, preliminary experiments on synthetic speech are provided. These experiments show that the proposed method has the potential to significantly outperform a reference method based on STFT: Signal-to-error ratio gains of 5 dB are typically obtained in the presented experiments. Conditions to go beyond the theoretical framework towards more practical applications on real speech signals are discussed.

1. INTRODUCTION

Speech signal components belong to two great classes because of the different possible voice sources: on the one hand, *harmonic* (H) components are generated by the vibration of the vocal folds, and on the other hand *aperiodic* (i.e., *noise* (N)) components are generated by fricative, plosive or aspiration noise sources [1]. Since the H/N sources can be simultaneous, these components are often mixed together in the acoustic realization of speech. For a given sound, the contribution of the respective components can be quantified by estimating a harmonics-to-noise (power) ratio (HNR) [2]–[4]. Such HNR is a useful measure for speech quality characterization, or for the diagnostic of pathological voices. Further, the complete *separation* of the H/N components from mixed-source signals is a major challenge in many speech (and also music [5]) processing applications [6]–[9]. It aims at obtaining two separate signals from the original speech: an estimated completely voiced signal and an estimated completely unvoiced signal, such that the sum of the two is equal to the original signal. Thus, the two estimated signals can be separately analyzed, modeled and modified, especially for synthesis [10][11], coding [12][13], and the study of fundamental mechanisms of speech production [8].

Several methods have been proposed in the literature for HNR estimation [2]–[4], and H/N separation [5]–[7]. Frequency-domain methods are almost all based on the use of the Short-Time Fourier Transform (STFT) for analysis/ synthesis: grossly speaking, dominant peaks of the spectrum are assumed to correspond to the harmonics, and “irregular” or inter-peak regions of the spectrum are assumed to correspond to the noise components. Now, such approach (as well as other time-domain methods such as in [2]) is limited by a crucial factor: speech signals are locally quasi-stationary and not strictly stationary. This means that both harmonics and noise components continuously evolve with time, more or less slowly. Therefore, significant differences generally occur from one period to the next, for both kinds of components, and it is a major difficulty in accurate H/N separation not to consider the evolution of the harmonics as part of the noise components [4]. However, in the literature, analysis frames generally include several periods of signal, and the analysis/synthesis process, e.g. using STFT, is intrinsically an averaging process that does not accurately capture the differences between the successive periods within the frame, but that rather extracts average characteristics across these periods and identify them to frame-wise constant harmonic components.

In this paper, we propose a new H/N separation method that aims to focus at the period scale, in order to accurately track and restitute the evolution of signal parameters from one period to the next. The H/N separation results from a filtering of the parameter trajectories. Thus the method is called PS-SPTF for Period-Scaled Spectral Parameters Trajectory Filtering. It is both time- and frequency-domain since it refers to the Fourier series expansion of *each signal period* (instead of the usual STFT approach). Foundations of such approach can be found in the previous work of Murphy [4] where HNR estimation was based on an averaging of successive values of complex Fourier series coefficients. The new contribution of the study is that we work with more “phenomenological” (real) phase and amplitude parameters, and that a complete H/N separation is conducted by using a linear filtering process on the global time-trajectory of these parameters, before resynthesizing separate signals from the filtered parameters.

The paper is organized as follows. The H/N separation method is presented in Section 2. Test methodology is given in Section 3, including the generation of test synthetic signals and the presentation of a reference STFT method derived from [7][8]. Results and perspectives are given in Sections 4 and 5.

2. THE PS-SPTF METHOD

2.1. General principle

We consider a mixed voiced-unvoiced signal of interest, which is made of K (pseudo-)periods $s_k(n)$. K is generally quite larger than the usual size of pitch-scaled method frames (e.g. four periods in [3][7]), since we lead a global process on the whole signal of interest. Each one of these K periods is separately decomposed as a sum of harmonically related cosine functions, referring to the Fourier series expansion of real signal:

$$s_k(n)=\sum_{i=1}^{I} A_i^k \cos\left(i\omega_0^k n+\theta_i^k\right) \qquad k = 1 \text{ to } K \qquad (1)$$

Thus, the whole processed signal is represented by I sets[1] of K amplitudes A_i^k and offset phases θ_i^k (i = 1 to I, k = 1 to K), plus one set of fundamental frequencies ω_0^k, k = 1 to K.

For a pseudo-periodic signal, the evolution of the amplitudes and offset phases from one period to the next must be quite "slow" or "smooth" because of the deterministic nature of the signal. On the contrary, aperiodic/noise components have a random nature, and the associated spectral parameters (especially the phases) should vary greatly from one period to the next[2]. Since the parameters are actually extracted from the mixed voiced-unvoiced signal, their trajectories typically exhibit a "smooth/slowly-evolving background", assumed to be due to the pseudo-periodic components, corrupted by additive-like noise, assumed to be due to the aperiodic components. Therefore, retrieving the harmonic signal from the mixed-source signal is made by retrieving the smooth background trajectory of the parameters and identifying it to the trajectory of the harmonic components parameters. This is done by low-pass filtering the time-trajectory of the parameters. The estimated harmonic signal is then generated by applying Eq. (1) using the filtered parameters instead of the measured (unfiltered) parameters. Eventually, the estimated aperiodic signal is generated by subtracting the former to the mixed signal. It is of primarily importance to note that the proposed filtering technique is a sliding adaptive averaging that follows and respects the period-scale dynamics of the parameters, as opposed to Murphy's technique and also to the global averaging on the entire analysis frame (including several periods of signals) resulting from STFT techniques. The proposed scheme attempts to retrieve the true trajectories of the harmonic parameters from the measures corrupted by the noise components, and, as opposed to signal reconstruction schemes based on inverse STFT, harmonic signals that evolve from one period to the next are reconstructed by the proposed method.

2.2. Technical Details

Parameters analysis: The proposed method assumes that the mixed-source speech signal to decompose is previously segmented into successive periods. In this paper, experiments are conducted on synthetic signals (see Sections 3 and 4; the justification for such choice is provided). Thus, pitch instants and periods length are exactly known. In this first examination of the proposed new method, the available exact values are used for the analysis process. In the case of real speech signals, different methods can be used to automatically estimate the pitch-marks. Clearly, the accuracy of the method strongly depends on the accuracy of the pitch-mark values. We do not deal with this specific point in this paper, because we focus on the basic principle of the H/N separation and we test first the feasibility of the new approach before possibly go further. Note that the problem of the influence of pitch-marking accuracy is more largely discussed in Section 5 and solutions to overcome this difficulty are provided. Thus, in this study, for each period k, the fundamental frequency ω_0^k is directly given by the inverse of the period length. Then, given ω_0^k, the amplitudes A_i^k and offset phases θ_i^k are estimated by using the procedure given by George and Smith in [15]. The estimation is based on an iterative minimum mean square error (MMSE) fitting of the harmonic model of Eq. (1) with the signal and it has been shown to provide very accurate parameter estimation with very low computational cost.

Phase regularization: Offset phase measures are provided modulo 2π. Since we want to extract information from the phase time-trajectories, we must first assume that no 2π-jump artificially corrupts their "natural" behavior. For this purpose, a regularizing "wrapping" process along the time axis is applied on each phase trajectory: it consists in successively adding or suppressing 2π to each phase value if this process results in a decrease of the variance of the phase trajectory vector. Since the background trajectory of the spectral parameters evolves with time, the variance is calculated using a sliding window of a few periods (typically four periods can be used). Several passes may be needed to ensure that no 2π-jump has escaped the regularizing process. Eventually, this process leads to perfectly regularized (but still noisy) phase trajectories.

Parameters filtering: As explained in Section 2.1, the next step and heart of the process is the low-pass filtering of the spectral parameters (amplitudes and phases) trajectories. Pilot tests have shown that a large set of very simple filters (*i.e.*, FIR, reduced number of coefficients) provide similar results. In the experiments presented in this paper, we used a 10-coefficients FIR filter with digital cut-off frequency of 0.1 resulting from the basic windowing method with a rectangular window. It is applied with zero-phase forward-backward filtering, so that the filtered and unfiltered parameters are kept synchronized, and so are the separated harmonic/noise and original mixed-source signals (remind that the noise signal is estimated by subtraction of the estimated (resynthesized) harmonic signal to the original signal in the time domain).

Amplitude re-estimation: In practice, it was observed that the corruption of parameter trajectories due to noise components was generally more pronounced for amplitude parameters than for phase parameters. This may be due to the quite low values for the amplitudes of middle-to-high rank harmonics for most voiced speech sounds. Therefore, the method was refined with a second-pass estimation of the amplitude parameters, after the filtering of the phases: for each period and each harmonic, the amplitude is re-estimated given the filtered offset phase value. This is done by a simplified

[1] For simplicity, the fixed maximum number of harmonics I corresponds to the minimum value of ω_0 over the K periods. Amplitudes of the harmonics that overcome the Nyquist limit are set to zero.

[2] See, e.g., [13] for more arguments on the deterministic vs. random nature of pitch-synchronous phase parameters for voiced and unvoiced components, and see, e.g., [14] for an application to unvoiced speech synthesis.

version of the previously used MMSE fitting between the harmonic model and the signal, where the phase is now fixed and only the amplitude must be calculated. The re-estimated amplitudes are then filtered with the low-pass filter.

3. TEST METHODOLOGY

3.1. Synthetic signals generation

Synthetic mixed-source signals are generated so that the "true" harmonic and noise parts are available for evaluation of the method. This is a standard methodology, largely adopted in the literature (e.g [4][7][9])., at least as a first step before applying the methods on real speech signals. The reason for this is that the synthetic harmonic and noise signals must be separately available for the calculation of objective and accurate separation measures such as the signal to error ratios (SER) that are in use in the following (see Sub-section 3.3). Direct application of any separation method on real speech signals can only be assessed by subjective listening tests, or by applying *a posteriori* "harmonicity measures" or "noiseness measures" on the separated signals. In the presented experiments, we do not use real speech signals, but informal listening test were conducted on the synthetic data, additionally to SER measurements (see Section 4).

The synthetic signals consist in different versions of the sustained ($K = 300$) vowels /a/ from a "male voice" and /i/ from a "female voice", sampled at 48kHz. The generation of these signals follows the usual methodology used in previous studies (e.g. [4][7]). A train of glottal flow pulse following the cosine-based model of Rosenberg [16] is used as the harmonic source. Random white Gaussian noise is used to simulate the noise source. It is possibly modulated by the amplitude of the glottal pulse train to take into account speech production considerations and increase naturalness [8][11]. Both sources are used as input into a digital all-pole filter that models the vocal tact. This filter results from 50-order LP analysis of a real signal from a male speaker for /a/ and a female speaker for /i/. The mixed signals are obtained by summing the two resulting filtered (centered) signals with different HNRs within the range –10 to 30 dB. Note that adequate pre-emphasis and lip-radiation first-order filters are used to fit the resulting mixed-source synthetic spectrum with the one of the real signal and ensure better natural sounding. Also, although sampled at 48kHz, the signals are band-limited by a 8 kHz low-pass filtering, so is the H/N separation process.

In order to assess the robustness of the H/N separation method on non-stationary (and closer to natural) signals, prosody is integrated by modulation of the fundamental frequency of the glottal source according to:

$$\omega_0^k = \alpha + \beta \cos\left(2\pi \frac{3k}{K}\right) + \gamma \frac{k^2}{K^2} \qquad (2)$$

The cosine term ensures three cycles of ω_0 contour, and the quadratic term ensures a fast raise at the end of the vowel. In Section 4, results are reported for experiments conducted with fixed fundamental (i.e., for /a/, $\alpha = 130$, $\beta = \gamma = 0$; for /i/, $\alpha = 280$, $\beta = \gamma = 0$), "normal intonation" values (i.e., for /a/, $\alpha = 130$, $\beta = 10$, $\gamma = 20$; for /i/, $\alpha = 250$, $\beta = 10$, $\gamma = 20$), and "exaggerated intonation" values (i.e., for /a/, $\alpha = 110$, $\beta = 30$, $\gamma = 100$; for /i/, $\alpha = 200$, $\beta = 30$, $\gamma = 200$) (all values are given in Hz).

3.2. STFT-based reference method

For comparative assessment of our method, we implemented the Pitch-Scaled Harmonic Filter (PSHF) method of Jackson and Shadle [7][8]. This method was chosen because i) it is well representative of methods based on STFT-analysis/synthesis ii) it is quite simple to implement compared to other methods (e.g. [9]) iii) its assessment on synthetic signals using SER measures provided an objective reference (see Section 4). Its principle is to calculate successive STFT spectra of exactly four periods of the mixed-source signal, so that the harmonic peaks are expected to be located every four bins and can be easily isolated. Thus, four periods of the estimated harmonic signal are given by inverse STFT of the comb-filtered spectrum, and the complete estimated harmonic signal is reconstructed by weighted overlap-add between successive local estimations. Subtracting this signal to the mixed-source signal provides the estimated noise signal.

3.3. SER measures

Assessment of the H/N separation is given by signal-to-error ratio (SER) which can be calculated for both harmonic and noise signals estimation. Let denote SER_H the power ratio between the harmonic part of the signal, and its difference with the estimated harmonic signal. Similarly, let denote SER_N the power ratio between the noise part of the signal, and its difference with the estimated noise signal. Since the estimated noise signal is obtained by subtracting the estimated harmonic signal to the mixed-source signal, the two SER measures are redundant: $\text{SER}_\text{H} = \text{SER}_\text{N} + \text{HNR}$. Thus, in the following, we only present SER_N (denoted simply SER) results, since it was found to be almost constant across HNRs in [7][8].

4. RESULTS

4.1. SERs

Fig. 1 gathers the SERs obtained on the test vowels, with both PS-SPTF and PSHF methods, and for the three ω_0 contours. The major results are the following:

- Both methods provide remarkably stable results across a large range of HNRs: SERs are almost constant from –10 to around 15 dB HNR for all cases except for /a/ with exaggerated intonation. For the PSHF method, the SER is around 5 dB (from 5 to 5.4 dB within the –10 to 15 dB HNR range, depending on the conditions) and this result is highly coherent with the results of [7][8], where such typical stable value of 5 dB was reported.
- The performances obtained with the new PS-SPTF method largely outperform this 5 dB reference. Within the –10 to 15 dB HNR range, values are all around 9.5 dB for /a/ (except for the exaggerated intonation) and around 10.5 dB for /i/ (at least for the normal and exaggerated intonation; Quite surprisingly, a slightly lesser value of 10 dB is obtained when the fundamental is fixed). Thus, the PS-SPTF generally provides a 4 to 5.5 dB improvement compared to the PSHF method, depending on the conditions. A typical separation result can be observed on the signals plotted in Fig. 2.
- Performances of both methods drop when HNR is over 15 dB, and the greater is the intonation variation, the greater is the SER deterioration. This is not surprising, since the smaller is the noise part of the signal, the more difficult it is to separate from the harmonic part. Also, increasing non-

stationarity makes the task more difficult for both methods. It can also be remarked that the two vowels, "male" /a/ and "female" /i/, exhibit quite different robustness to these degradations, but a discussion on the phonetic factors of influence is beyond the scope of this paper. Now, even in difficult conditions, the advantage of the PS-SPTF method over the PSHF method remains always greater than 4 dB, except for /i/ with exaggerated intonation at 25–30 dB HNR, where "only" 3.7 and 3.1 dB gains are obtained. For other conditions, the PS-SPTF gain over PSHF is typically 5 dB, and it can even significantly overcome this value: e.g., for /a/ with ω_0 fixed, a gain of 8 dB is obtained at 30 dB HNR.

- Finally, it can be noted that the results obtained with or without modulation of the noise source by the amplitude of the glottal source were always very similar in our experiments. Thus, only the results without the modulation were presented in Fig. 1. Complementarily, in the example of separation given in Fig. 2, the noise signal is modulated. These results seem to indicate that both methods are quite robust regarding possible non-stationarity of the noise source. This point is of high interest for further study on real speech signals and is to be further investigated.

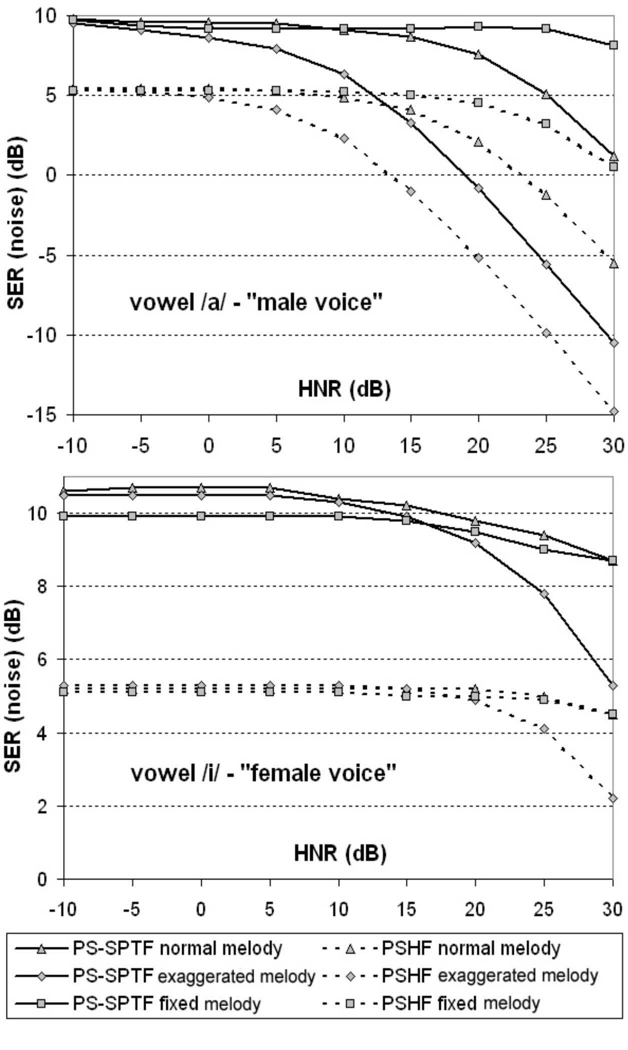

Figure 1 – SER (of noise signal) as a function of HNR.

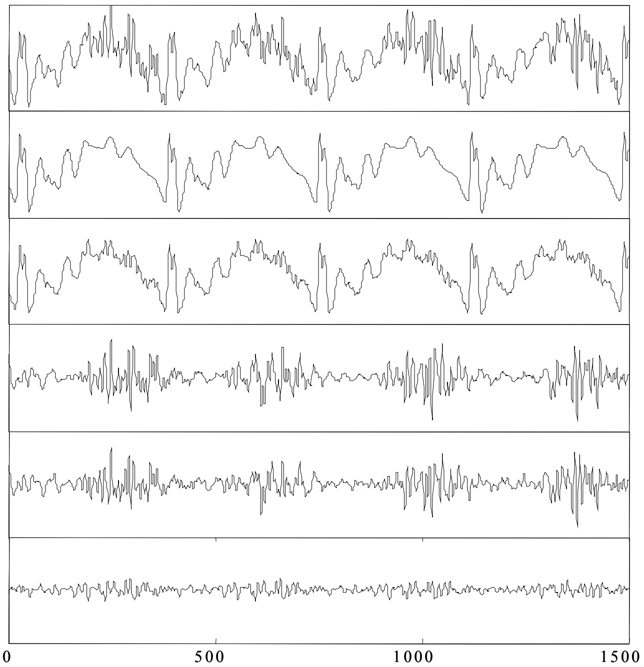

Figure 2 – Example of H/N separation with the PS-SPTF method: segment of vowel /a/ with modulated noise source and HNR = 0 dB. From top to bottom: mixed-source signal, true and estimated harmonic part, true and estimated noise part, estimation error (difference between true and estimated signal). Y-axis scale is arbitrary but consistent within different signals; For this example, we obtain SER = 9 dB.

4.2. Informal listening tests

Listening tests confirm the good performances of the PS-SPTF method, and the improvement compared to the PSHF method. For medium-to-high HNRs (*i.e.,* say 0 to 30 dB), the harmonic signal estimated with the PS-SPTF method is generally perceptually undistinguishable from the true harmonic signal, whereas there often remains a significant amount of noise in the harmonic signal estimated with the PSHF method. The filtering of the spectral parameters ensures quite smooth trajectories and leads to a "highly harmonic" estimated signal. In contrast, the PSHF method can suffer from the fact that values sampled every four bins of a STFT spectrum are not bound to actually correspond to harmonic peaks if the signal is non-stationary. For low HNRs, the separated signals are generally of lower quality, *i.e.* perceptually moved away from the true harmonic and noise signals, with different quality for the PS-SPTF and the PSHF methods. Note that all tested mixed, true, and separated signals are available online at the following URL: www.icp.inpg.fr/~girin/HNS/HNS_demo.zip. The reader is invited to make its own perceptual judgment.

5. DISCUSSION

Within the "ideal framework" of the presented study, the proposed PS-SPTF method was shown to provide a large improvement compared to a recently published reference method based on STFT analysis/synthesis (typically 10 dB vs. 5 dB SER). It appears from this preliminary theoretical and experimental approach that the filtered parameters trajectories can be associated to slow wave-shape variations, as a part of the pseudo-periodic components. Although preliminary, these encouraging results must be taken carefully because of the already mentioned expected dependency of the method to the accuracy of the pitch-mark estimation. Especially, phase measures are expected to be significantly corrupted by pitch-mark inaccuracies, much more than amplitude measures (although amplitude measures may suffer from consequent fundamental frequency bias). And the higher is the harmonic rank, the higher is the corruption, since phase variation is directly related to frequency time-integration. However, this limitation must be strongly alleviated by at least two points:

- First, the low-pass filtering of the spectral parameters may provide intrinsic compensation for such additional noise; In other words, the filtering may be useful to remove both the noise due to noise components of speech, and the noise due to the analysis inaccuracies. This is true as long as the amount of total noise does not prevent the emerging of the background shape for phase trajectories. Further investigation is needed to clarify this point. Especially, the influence of automatic pitch-mark estimation must be studied, and also we must analyze the interactions of both (measure and speech) noise sources.
- Second, the *offset phase* that are estimated and filtered in this study may be replaced by *absolute phase* values, i.e. the phase values resulting from the time-integration of frequency values. Indeed, the trajectory of absolute phases can be reconstructed from measures taken arbitrarily in time. As opposed to the regularization process of Section 2.2 for offset phases, the reconstruction of absolute phase trajectories requires phase measures unwrapping [17], a somehow dual procedure which is quite simple to implement (actually, in many analysis/synthesis system based on the sinusoidal model, absolute phases are considered as time functions, see e.g., [5][17). Thus the estimation of a smooth absolute phase trajectory from noisy measures is expected to lead to an equivalent result, with the strong advantage of not depending on measure instant, such as the pitch-marks used in the present study (however, the analysis window length should remain close to the signal period to accurately capture the evolution of the signal). Note that the smoothing of absolute phases trajectories can be obtained by a similar filtering process, and also by alternative approaches such as the long-term modeling proposed in [18].

These two points constitute the kernel of our current works. Obviously, the second point will build on [18]. They are expected to provide a significant step toward realistic implementation of the method. Beyond these points, future work may more generally concern:

- Application of the method on synthetic signals that simulate complex non-stationarities of speech, such as jitter, shimmer, evolution of the vocal tract (see e.g. [4][7]) and possibly other types of complex ω_0 variations.
- Application on real speech signals. Especially, signals that exhibit a significant amount of both H/N components and complex H/N components interactions, such as voiced fricatives, will be of strong interest.
- Comparison with other methods must be assessed (e.g., [9], or an adaptation of the HNR estimation method of [4] to provide complete H/N separation).

6. REFERENCES

1. Stevens, K. N. (1998). *Acoustic phonetics*, MIT Press, Cambridge, MA.
2. Qi, Y. & Hillman, R. E. (1997). Temporal and spectral estimations of harmonic-to-noise ratio in human voice signals, *J. Acoust. Soc. Am.*, 102(1), 537-543.
3. Muta, H., Baer, T., Wagatsuma, K., Muraoka, T. & Fukuda, H. (1988). A pitch-synchronous analysis of hoarseness in running speech, *J. Acoust. Soc. Am.*, 84(4), 1292-1301.
4. Murphy, P. (1999). Perturbation-free measurements of the harmonics-to-noise ratio in voice signals using pitch-synchronous harmonic analysis, *J. Acoust. Soc. Am.*, 105(5), 2866-2880.
5. Serra, X. & Smith, J. (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition, *Comp. Music J.*, 14(4), 12-24.
6. Stylianou, Y. (1996). Decomposition of speech signals into a deterministic and a stochastic part, *Proc. Int. Conf. Spoken Language Proc.*, Philadelphia, PA, USA.
7. Jackson, P. & Shadle, C. (2001). Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech, *IEEE Trans. Speech Audio Proc.*, 9(7), 713-726.
8. Jackson, P. & Shadle, C. (2000). Fricative noise modulated by voicing, as revealed by pitch-scaled decomposition, *J. Acoust. Soc. Am.*, 108(4), 1421-1434.
9. Yegnanarayana, B., d'Alessandro, C. & Darsinos, V. (1998). An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Trans. on Speech and Audio Processing*, 6(1), 1-11.
10. Richard, G. & d'Alessandro, C. (1996). Analysis, synthesis and modification of the speech aperiodic component, *Speech Communication*, 19, 221-244.
11. Hermes, D. J. (1991). Synthesis of breathy vowels: Some research methods, *Speech Communication*, 10, 497-502.
12. Makhoul, J., Viswanathan, R., Schwartz, R. & Huggins, A. (1978), A mixed-source model for speech compression and synthesis, *J. Acoust. Soc. Am.*, 64(6), 1577-1581.
13. Kang, G. & Everett, S. (1985). Improvement of the excitation source in the narrow-band linear prediction vocoder, *IEEE Trans. Acoust. Speech Sig. Proc.*, 33(2), 377-386.
14. Macon, M. W. & Clements, M. A. (1997). Sinusoidal modeling and modification of unvoiced speech, *IEEE Trans. Speech and Audio Proc.*, 5(6), 557-560.
15. George, E. & Smith, M, Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model, *IEEE Trans. Speech and Audio Proc.*, **5**(5), 1997, pp. 389-406.
16. Rosenberg, A. E., Effect of glottal pulse shape on the quality of natural vowels, *J. Acoust. Soc. Am.* 49(2), 583-590, 1971.
17. R. J. McAulay & T. F. Quatieri, Speech analysis/ synthesis based on a sinusoidal representation, *IEEE Trans. Acoust. Speech and Signal Proc.*, 34(4), 1986, pp. 744-754.
18. L. Girin, M. Firouzmand & S. Marchand, “Long term modeling of phase trajectories within the speech sinusoidal model framework,” *Proc. Int. Conf. on Speech & Language Proc.,* Jeju, 2004.