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

# A NOVEL VOICED SPEECH ENHANCEMENT APPROACH BASED ON MODULATED PERIODIC SIGNAL EXTRACTION

*Mahdi Triki*[†], *Dirk T.M. Slock*[*]

[†] CNRS, Communication Systems Laboratory
[*] Eurecom Institute
2229 route des Crêtes, B.P. 193, 06904 Sophia Antipolis Cedex, FRANCE
Email: {triki,slock}@eurecom.fr

## ABSTRACT

Most of the existing speech coding and speech enhancement techniques are based on the AR model and hence apply well to unvoiced speech. These same techniques are then applied to the voiced case as well by extrapolation. However, voiced speech is very structured so that a proper approach allows to go further than for unvoiced speech. We model a voiced speech segment as a periodic signal with (slow) global variation of amplitude and frequency (limited time warping). The bandlimited variation of global amplitude and frequency gets expressed through a subsampled representation and parameterization of the corresponding signals. Assuming additive white Gaussian noise, a Maximum Likelihood approach is proposed for the estimation of the model parameters and the optimization is performed in an iterative (cyclic) fashion that leads to a sequence of simple least-squares problems. Particular attention is paid to the estimation of the basic periodic signal, which can have a non-integer period, and the estimation of the amplitude signal with guaranteed positivity.

## 1. INTRODUCTION

Speech enhancement can be described as the processing of speech signals to improve one or more perceptual aspects of speech, such as overall quality, intelligibility for human or machine recognizers, or degree of listener fatigue. The need for enhancing speech signals arises in many situations in which the speech either originates from some noisy location or is affected by the noise over the channel or at the receiving end. In the presence of background noise, the human auditory system is capable of employing effective mechanisms to reduce the effect of noise on speech perception. Although such mechanisms are not well understood at the present state of knowledge to allow the design of speech enhancement systems based on auditory principles, several practical methods for speech enhancement have already been developed. Several reviews can be found in the literature [1, 2, 3].

In this study, it is assumed that i) only the degraded speech signal is available, and ii) that the noise is additive and uncorrelated with the speech signal. Under theses assumptions, if the statistics of the clean signal and the noise process are explicitly known, enhancement could be optimally accomplished using the estimator which minimizes the expected value of the distortion measure between the clean and the estimated signals [3]. In practice, however, these statistics are not explicitly available, and should be estimated. Hence, the above theoretical approach can be applied as a two-step procedure in which the statistics of signal and noise are first estimated, and then used together, with currently available distortion measures, to solve the problem of interest. The optimality of the two-step enhancement approach depends on the specific estimators used for the unknown statistics. For example, non-parametric spectral estimation techniques can be used to estimate both the noise and noisy-speech spectrum. Then, a frequency-domain Wiener filter is constructed, which is then used to obtain the clean speech estimate. This leads to the well-known, Spectral Subtraction technique [4]. Spectral subtraction has been one of the relatively successful DSP methods due to its implementation simplicity and its capability of handling noise non-stationarity to some extent. However, one major problem with this method is the annoying non-stationary ”musical” background noise associated to the enhanced speech.

A tractable alternative of non-parametric spectral estimation is provided by parametric modeling of the probability density of the sources (speech and noise). Enhancement based on the estimation of all-pole speech parameters in additive white Gaussian noise was investigated by Lim and Oppenheim [5], and later for a colored noise degradation by Hansen and Clements [6]. They propose an iterative algorithm in which we iterate AR coefficients estimation, and Wiener filtering (based on parametric spectrum estimate). Spectral constraints based on the AR modeling [7], or on the HMM phoneme class partition [8], are proposed to increase the technique performance.

Another useful class of speech signal models, for speech recognition and enhancement, are Hidden Markov Models (HMM). Enhancement methods that are based on stochastic models (HMM's) have been most successful as they model both clean speech and noise, and accommodate the non-stationarity of speech and noise with multiple states connected with transition probabilities in a Markov chain [9].

However, the nature of the human speech dictates that not every short segment can be treated in the same fashion. In fact,

[*] Eurécom's research is partially supported by its industrial partners: BMW, Bouygues Telecom,Cisco Systems,France Télécom, Hitachi Europe, SFR, Sharp, ST Microelectronics,Swisscom, Thales. The research reported herein was also partially supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content - K-Space.

speech segments can be classified in terms of the sounds they produce [10]. Basically, there are two sound categories: i) Unvoiced sounds, such as the /s/ in 'soft', are created by air passing through the vocal tract without the vocal cords vibrating. They exhibit low signal energy, no pitch, and a frequency spectrum biased towards the higher frequencies of the audio band, ii) Voiced sounds, such as /AH/ in 'and', are created by air passing through the glottis causing it to vibrate. And contrarily to unvoiced speech, voiced speech has greater signal energy, a pitch, and a spectrum biased towards the lower frequencies. In order to take advantage of the voicing in the glottal source signal, we propose modelling voiced sounds as a periodic signal with a global amplitude and phase modulation; and to take into account this structure to denoise the voiced segment.

This paper is organized as follows. In section 2, the global modulation model is presented. The speech enhancement procedure will then be derived in section 3. Performance of the algorithm is evaluated in Section 4, and finally a discussion and concluding remarks are provided in section 5.

## 2. GLOBAL MODULATION MODEL FOR VOICED SPEECH SIGNAL

In the sinusoidal model, the signal is modeled as a sum of evolving sinusoids:

$$s(n) = \sum_{k=0}^{P} A_k(n) \cos\left(\theta_k(n)\right) \quad . \tag{1}$$

where $\theta_k(n)$ represents the instantaneous phase of the $k^{th}$ partial. As the voiced speech signal is quasi-periodic, $\theta_k(n)$ can be decomposed into

$$\theta_k(n) = 2\pi k n f_0 \; + \; 2\pi\varphi_k(n) \tag{2}$$

where $k$ is the harmonic index, $f_0$ denotes the pitch frequency (normalized by the sampling frequency), and $\varphi_k(n)$ characterizes the evolution of the instantaneous phases around the $k^{th}$ harmonic; and can be assumed to be low-frequency.
The Global Modulation assumption implies that all harmonic amplitudes evolve proportionally in time; and that the instantaneous frequency of each harmonic is proportional to the harmonic index:

$$\left\{ \begin{array}{l} A_k(n) = A_k \; A(n) \\ 2\pi\varphi_k(n) = 2\pi k \; \varphi(n) \; + \; \Phi_k \end{array} \right. \quad . \tag{3}$$

In summary, we model a voiced speech signal as the superposition of harmonic components with a global amplitude modulation and time warping (that can be interpreted in terms of phase variations):

$$\begin{array}{ll} y(n) & = \; s(n) \; + \; v(n) \\ & = \; \sum_k A_k(n) \; \cos\left(2\pi k n f_0 + 2\pi\varphi_k(n)\right) + v(n) \\ & = \; A(n) \sum_k A_k \cos\left(2\pi k f_0 \left(n + \frac{\varphi(n)}{f_0}\right) + \Phi_k\right) + v(n) \end{array}$$

where

- $v_n$ is an additive white Gaussian noise.
- $A(n)$ represents the amplitude modulating signal. It allows an evolution of the signal power.
- $\varphi(n)$ denotes the phase modulating signal (that can be interpreted in terms of time warping). The time warping focuses on the time evolution of the instantaneous frequency.

In [11], we have expressed the time warping in terms of an interpolation operation over a basic periodic signal. In matrix form, the noisy voiced speech signal can be written as:

$$Y = \underbrace{A\,F\theta}_{=\,S} + V \tag{4}$$

where :
- $Y = [y(1) \cdots y(N)]^T$, represents the observation vector
- $S = [s(1) \cdots s(N)]^T$, represents the signal of interest
- $V = [v(1) \cdots v(N)]^T$, denotes the noise vector
- $\theta = [\theta(1) \cdots \theta(\lceil T \rceil)]$, characterizes the harmonic signature over essentially one period
- $A = diag[A(1) \cdots A(N)]$, represents the global amplitude modulation signal
- $F$ is an $N \times \lceil T \rceil$ interpolation matrix characterizing the time warping. See [11] for a detailed description.

Note that the previous model can be interpreted in terms of long-term prediction. Long-term prediction is typically used for voiced-speech coding. The most basic long-term predictor is the one tap filter given by

$$s_p\,(n) \;=\; G\,s\,(n - T) \tag{5}$$

where $s(n)$ is the input signal, $s_p(n)$ is the predicted signal, $T$ is an integer value, and $G$ is a gain. In [13], the authors propose a long-term scheme using fractional delay. They show that this technique enables a more accurate representation of the voiced speech and achieves an improvement of synthetic quality for female speakers. Our model generalizes the previous approach by allowing tracking (slow) variations of gain and fractional delay (global amplitude and frequency modulation variations). Such an approach enables, not only a good tracking of the signal of interest, but also the rejection of signals having a different structure (white noise, PC noise, car noise, and human voice...), especially if the spectrum of this colored noise is concentrated in different frequency regions than the voiced speech.

Remark also that the described extraction technique models, and takes advantage of the correlation between the different partials. And contrary to classical sinusoidal modeling techniques, it does not any assumption on the value of $P$ (in (1)). Implicitely, $P$ is the maximum integer such that $f_0 P < \frac{1}{2}$ (the sampling frequency satisfy the Nyquist-Shannon sampling theorem).

## 3. SPEECH ENHANCEMENT TECHNIQUE

The proposed enhancement algorithm (figure 1) is based on a different treatment of the voiced and unvoiced speech components. The processing steps are discussed in the following sections.

### 3.1. Enhancement Stage

*3.1.1. Voiced speech extraction*

As the voiced speech signal is assumed to be quasi-periodic (following (4)), it can be written as

$$\widehat{S} = \widehat{A}\;\widehat{F}\widehat{\theta}$$

The previous model is linear in $\theta$, $A$, or $F$ (separately), $F$ being parameterized nonlinearly.

As the noise is assumed to be a white Gaussian signal, the Maximum Likelihood (ML) approach leads to the following least-squares problem:

$$\min_{A,F,\theta} \|Y - AF\theta\|^2 \tag{6}$$

where $A$ and $F$ are parameterized in terms of subsamples. Trying to estimate all factors jointly is a difficult nonlinear problem. However, The estimation can easily be performed iteratively (as in [11, 12]).

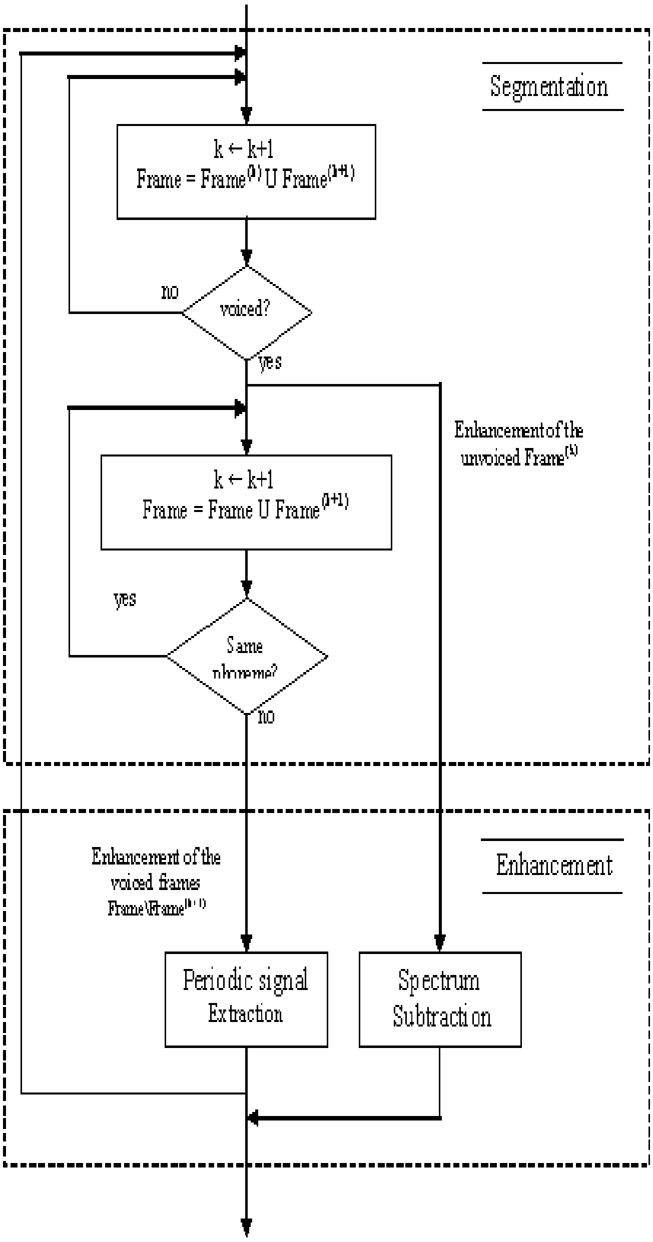
Segmentation
k ← k+1
Frame = Frame$^{(l)}$ U Frame$^{(l+1)}$
no
voiced?
yes
Enhancement of the unvoiced Frame$^{(l)}$
k ← k+1
Frame = Frame U Frame$^{(l+1)}$
yes
Same phoneme?
no
Enhancement
Enhancement of the voiced frames Frame\Frame$^{(l+1)}$
Periodic signal Extraction
Spectrum Subtraction

**Fig. 1**. Speech Enhancement Technique

*3.1.2. Unvoiced speech extraction*

In our preliminary experiments, the well-known spectral subtraction is employed to the unvoiced speech segments, for simplicity [4, 9]. In this conventional method, a frequency-domain Wiener filter is constructed from the speech and noise spectral estimates at each time frame, which is then used to obtain a clean speech estimate. The noisy signal power spectral density ($P_{yy}$) is estimated (by a Periodogram technique) using the observed signal of the current frame. Whereas the estimate of the noise spectrum ($P_{vv}$) is updated during periods of non-speech activity. The tracking of the noise spectrum can be performed, also, on voiced frames (using the noise estimate $\widehat{v} = y - \widehat{s}$). Finally, enhanced speech is reconstructed by Wiener filtering in the frequency domain:

$$\widehat{S}(w) = H(w)Y(w) \tag{7}$$

where $H(w) = \left(\frac{|\widehat{P}_{yy} - \widehat{P}_{vv}|}{\widehat{P}_{yy}}\right)^{\frac{1}{2}}$ denotes the estimated square root of the Wiener filter.

### 3.2. Segmentation stage

The segmentation of the speech signal, i.e. classification of speech into voiced/unvoiced frames, is a crucial issue to ensure the performance of the Enhancement stage. In fact, the estimation accuracy of the quasi-periodic signal, as well as the spectrum of the noisy speech, depends on the speech frame length. On the other hand, the time resolution of these parameters is only as fine as the window length, itself. Since a speech signal is strongly non-stationary, it is not always possible to find a constant frame length giving a good tradeoff between estimation and localization accuracy.

There is a vast literature on speech segmentation with applications to speech analysis, synthesis, and coding [14, 15]. In some speech applications, the digital signal processing techniques are augmented by linguistic constraints or may be ”supervised” by a human operator. However, manual phonetic segmentation is very costly and requires much time and effort. Automatic segmentation methods utilize from energy and zero crossings for silence and/or endpoint detection, to much more sophisticated spectral analysis methods for detecting changes in the speech spectrum. Each of these methods monitors one or more indicators, such as energy, number of zero crossings, pitch period, prediction error energy, or a spectral distortion measure, to detect significant changes.

Note that here the segmentation stage is not designed for recognition or classification applications. Its purpose is just to identify frames having similar spectrum characteristics (essentially spectrum envelope, and periodicity); such that they can be treated together. This motivates the choice of a distance criterion based on the energies of the extracted signal and the noise,

$$D = \max_{T} \frac{\sigma^2_{\widehat{s}_T} + \sigma^2_v}{\sigma^2_y} \tag{8}$$

where:

- $\widehat{s}_T$ is the quasi-periodic signal with a period $T$ extracted as described is section 3.1.1.

- $\sigma^2_{\widehat{s}_T}$, $\sigma^2_v$, and $\sigma^2_y$ represent, respectively, the power of the extracted quasi-periodic signal, the noise and the received signal.

As we have seen in section 3.1.1, for a given period $T$, the proposed extraction algorithm approximates the projection of the noisy signal onto the subspace spanned by the set of $T$-periodic signals with low-pass amplitude and phase modulations. Thus, if the received signal corresponds to a unique voiced phoneme, $\exists T \;/\; \sigma^2_{\hat{s}_T} + \sigma^2_v \approx \sigma^2_y$, then $D \approx 1$. However, if the received signal corresponds to an unvoiced phoneme ($\forall T \;\; \sigma^2_{\hat{s}_T} \approx 0$), or if it contains more than one phoneme ($\exists T_1 \neq T_2 \;/\; \sigma^2_{\hat{s}_{T_1}} \neq 0,\; \sigma^2_{\hat{s}_{T_2}} \neq 0$), we have $1 > D \rightarrow \frac{\sigma^2_v}{\sigma^2_y}$.
Consequently, the distance $D$ seems to be suitable for our application.

The proposed segmentation procedure is described in figure 1. The main idea to split speech signal into 10 ms frames; then use of the distance $D$ to group together frames belonging to the same voiced phonemes.

## 4. EXPERIMENTAL RESULTS

We now introduce some tests to evaluate the performance of the proposed speech enhancement scheme. The sampling rate is 8 kHz. A synthetic Gaussian white noise is added to speech signal. We first see the performance of the proposed scheme on a speech signal with relatively high SNR (SNR = 20 dB) in figure 2. In the figure 2.(b), we superpose curves of the extracted voiced signal, and the envelope of the original (noise free) signal. Obviously, the quasi-periodic model holds (with a good accuracy) for the voiced speech segments.

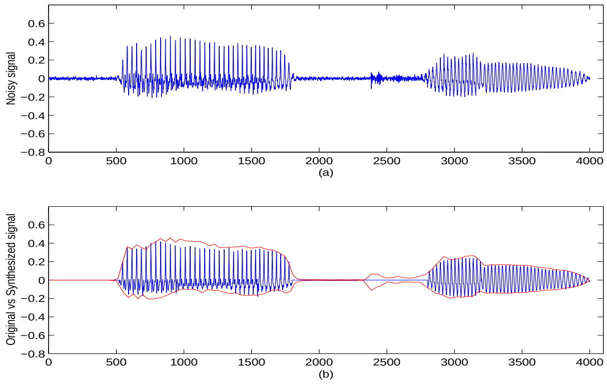

**Fig. 2**. Noisy speech, extracted voiced speech, and noisefree signal envelope (SNR=20dB)

We then test the proposed scheme in a very noisy environment (SNR = 0 dB) (figure 3). In this second set of simulations, we treat only voiced frames (as spectral subtraction gives poor results); unvoiced frames are set to zero. Remark that in a noisy environment, the speakers have a tendency to stretch voiced phonemes (Lombard effect ). We observe that the quasi-periodic characteristic is robust to the additional noise, and allows speech enhancement in a very noisy environment.
Furthermore, we consider a global measure of signal-to-noise ratio ($SNR_{out}$) as an objective evaluation criterion through this work

$$SNR_{out} = 10 \log \frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N} \left(s(n) - \hat{s}(n)\right)^2}$$

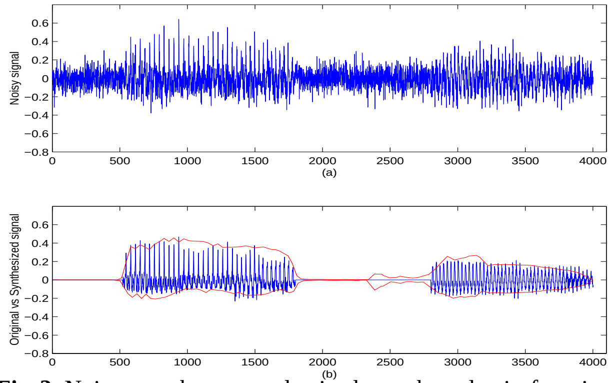

**Fig. 3**. Noisy speech, extracted voiced speech, and noisefree signal envelope (SNR=0dB)

which is consistent with previous enhancement studies [8, 9]. Figure 4 plots curves of the averaged output SNR (evaluated by Monte-Carlo techniques) for our proposed scheme and the classical spectral subtraction technique [4, 9].

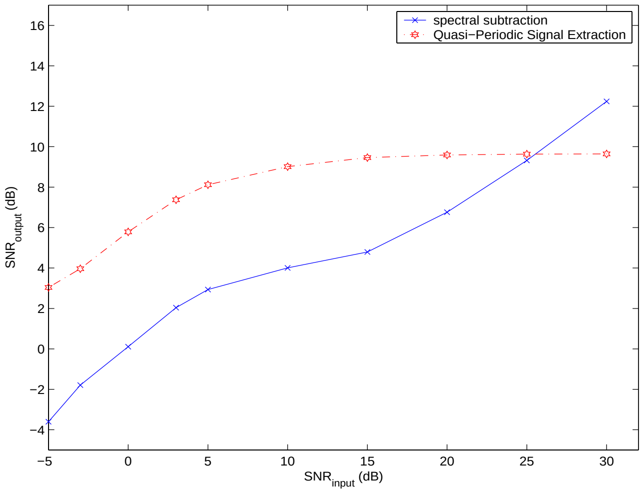

**Fig. 4**. Comparison of our proposed scheme and the spectral subtraction technique for white noise corrupted speech signal.

The output SNR has straightforward interpretation; and it can provide indications of the perceived audio quality in some cases [16]. Unfortunately, the output SNR shows a limited correlation with perceived speech quality. Therefore, some speech quality assessment algorithms try to include explicit models of the human auditory perception system. The ITU P.862 PESQ (Perceptual Evaluation of Speech Quality [18, 19]) is one of the most recently introduced methods, that is found implemented in many commercially available testing devices and monitoring systems [17].
Figure 5 plots curves of the averaged PESQ criterion (evaluated by Monte-Carlo techniques) for our proposed scheme and the classical spectral subtraction technique.

As can be observed in the previous graphs, the proposed scheme outperforms the spectral subtraction in low to high SNR regions. However, at very high SNR, the achievable output SNR of the proposed method is saturated due to approximation error in the periodicity model.
Remark that in our simulations, the noise spectrum is assumed to

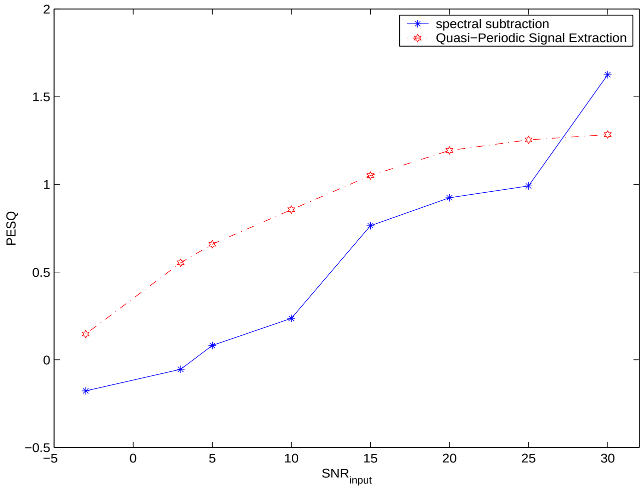

**Fig. 5**. Comparison of our proposed scheme and the spectral subtraction technique for white noise corrupted speech signal.

be known. It could be estimated during silence periods. Note that knowledge of the noise spectrum is required for spectral subtraction but not for the modulated periodic signal extraction. Nevertheless, the performance of this last technique is affected by the color of the noise. In this respect, a white noise will tend to lead to worse results than a colored noise (PC noise, car noise, human voice), especially if the spectrum of this colored noise is concentrated in different frequency regions than the voiced speech.

## 5. CONCLUSIONS

This paper has introduced a new speech enhancement technique based on quasi-periodic signal extraction. The proposed enhancement algorithm is based on a differential treatment of the voiced and unvoiced speech components. Unvoiced frames are treated using the well-known spectral subtraction technique. For voiced frames, we have considered the periodic signal model with a slow global amplitude and phase variation. The model parameters estimation is performed in an iterative (cyclic) fashion that leads to a sequence of simple least-squares problems. Simulations show that the enhancement technique achieves quite good performance (specially in very noisy environments).

## 6. REFERENCES

[1] J.S. Lim, Ed.“Speech Enhancement,” *Englewood Cliffs*, NJ: Prentice-Hall, 1983.

[2] D. O’Shaughnessy. “Enhancing speech degraded by additive noise or interfering speakers,” *IEEE Communications Magazine*, Vol. 27, Issue 2, pp. 46-52, Feb. 1989.

[3] Y. Ephraim. “Statistical model based speech enhancement systems,” *In Proc. of the IEEE*, Vol. 80, No. 10, pp. 1526-1555, Oct. 1992.

[4] J. Ortega-Garcia, J. Gonzalez-Rodriguez. “Overview of speech enhancement techniques for automatic speaker recognition,” *In Proc. of Int. Conf. on Spoken Language Processing*, Vol. 2, pp. 929-932, 1996.

[5] J. Lim, A. Oppenheim. “All-pole modeling of degraded speech,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 26, Issue 3, pp. 197-210, June 1978.

[6] J.H.L. Hansen, M.A. Clements. “Enhancement of Speech Degraded by Non-White Additive Noise,” *Technical Report DSPL-85-6*, Georgia Institute of Technology, Aug. 1985.

[7] J.H.L. Hansen, M.A. Clements. “Constrained iterative speech enhancement with application to speech recognition,” *IEEE Trans. on Signal Processing*, Vol. 39, Issue 4, pp. 795-805, April 1991.

[8] J.H.L. Hansen, L.M. Arslan. “Markov model-based phoneme class partitioning for improved constrained iterative speech enhancement,” *IEEE Trans. on Speech and Audio Processing*, Vol. 3, Issue 1, pp. 98-104, Jan. 1995.

[9] H. Sameti, H. Sheikhzadeh, L. Deng, R.L. Brennan. “HMM-based strategies for enhancement of speech signals embedded in nonstationary noise,” *IEEE Trans. on Speech and Audio Processing*, Vol. 6, Issue 5, pp. 445-455, Sept. 1998.

[10] J.A. Marks.“Real time speech classification and pitch detection,” *In Proc. of Southern African Conf. on Communication and Signal Processing*, pp. 1-6, June 1988.

[11] Mahdi Triki, Dirk T.M. Slock. “Periodic Signal Extraction with Global Amplitude and Phase Modulation for Music Signal Decomposition,” *In Proc. of Int. Conf. on Acoustic, Speech, and Signal Processing*, Vol. 3, pp.233-236, March 2005.

[12] Mahdi Triki, Dirk T.M. Slock. “Multi-channel mono-path periodic signal extraction with global amplitude and phase modulation for music and speech signal analysis,” *In Proc. of IEEE Workshop on Statistical Signal Processing*, pp.7782, July 2005.

[13] J.S. Marques, I.M. Trancoso, J.M. Tribolet, L.B. Almeida. “Improved pitch prediction with fractional delays in CELP coding,” *In Proc. of Int. Conf. on Acoustic, Speech, and Signal Processing*, Vol. 2, pp. 665-668, April 1995.

[14] L. Ta-Hsin, J.D. Gibson. “Speech analysis and segmentation by parametric filtering,” *IEEE Trans. on Speech and Audio Processing*, Vol. 4, Issue 3, pp. 203-213, May 1996.

[15] D.T. Toledano, L.A.H. Gomez, L.V. Grande. “Automatic phonetic segmentation,” *IEEE Trans. on Speech, and Audio Processing*, Vol. 11, Issue 6, pp. 617-625, Nov. 2003.

[16] S. Voran. “Objective estimation of perceived speech quality - part I: Development of the measuring normalizing block technique,” *IEEE Trans. on Speech, and Audio Processing*, Vol. 7, Issue 4, pp. 371-382, July 1999.

[17] A.E. Conway. “Output-based method of applying PESQ to measure the perceptual quality of framed speech signals,” *In Proc. of IEEE Wireless Communications and Networking Conf.*, Vol. 4, pp. 21-25, March 2004.

[18] A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra. “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” *In Proc. of Int. Conf. on Acoustic, Speech, and Signal Processing*, Vol. 2, pp. 749-752, May 2001.

[19] *ITU-T Recommendation P.862*, “Perceptual Evaluation of Speech Quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone network and speech codecs,” 2001.