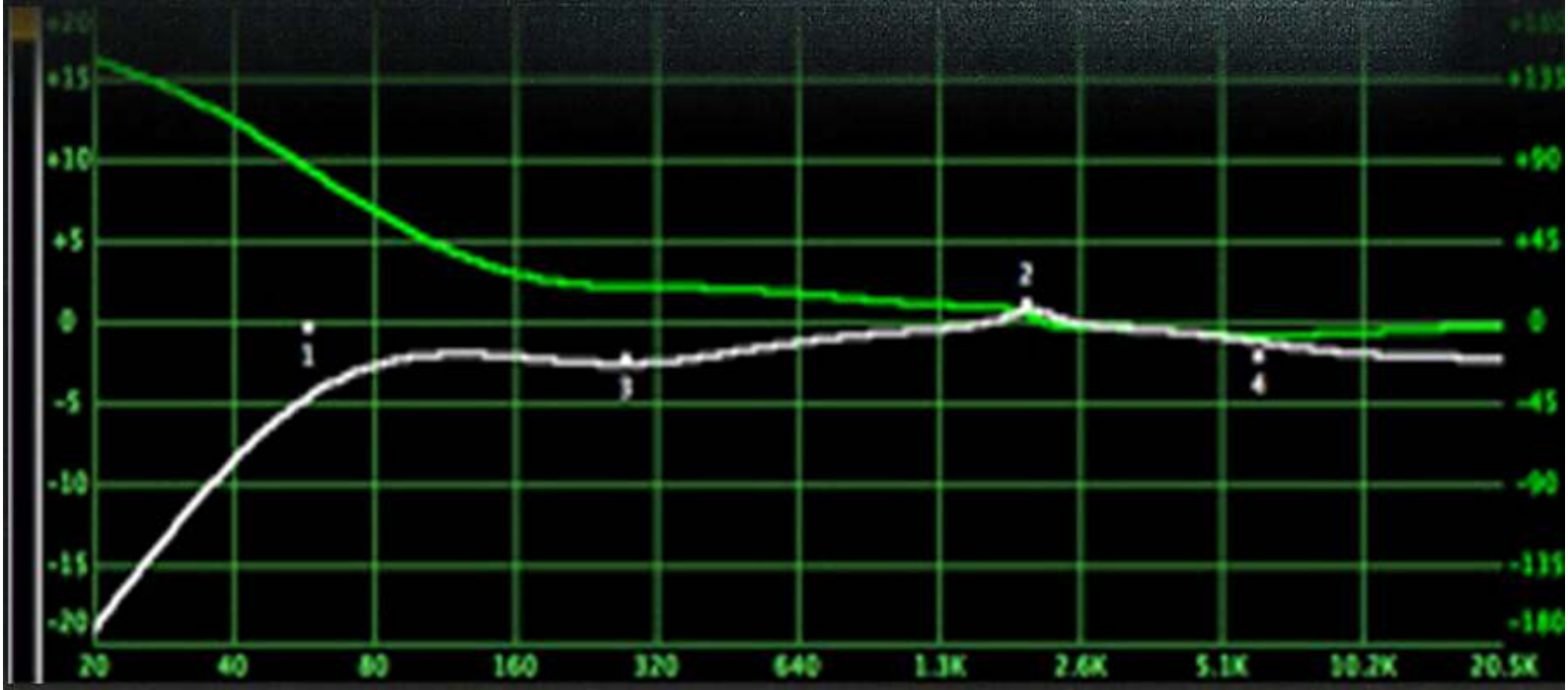
Signal Processing:
An International Journal (SPIJ)

ISSN : 1985-2339

VOLUME 4, ISSUE 1

PUBLICATION FREQUENCY: 6 ISSUES PER YEAR



Signal Processing: An International Journal (SPIJ)

Volume 4, Issue 1, 2010

Edited By
Computer Science Journals
www.cscjournals.org

Editor in Chief Dr. Saif alZahir

Signal Processing: An International Journal (SPIJ)

Book: 2010 Volume 4 Issue 1
Publishing Date: 31-03-2010
Proceedings
ISSN (Online): 1985-2339

SPIJ Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers

Editorial Preface

This is first issue of volume four of the Signal Processing: An International Journal (SPIJ). SPIJ is an International refereed journal for publication of current research in signal processing technologies. SPIJ publishes research papers dealing primarily with the technological aspects of signal processing (analogue and digital) in new and emerging technologies. Publications of SPIJ are beneficial for researchers, academics, scholars, advanced students, practitioners, and those seeking an update on current experience, state of the art research theories and future prospects in relation to computer science in general but specific to computer security studies. Some important topics covers by SPIJ are Signal Filtering, Signal Processing Systems, Signal Processing Technology and Signal Theory etc.

This journal publishes new dissertations and state of the art research to target its readership that not only includes researchers, industrialists and scientist but also advanced students and practitioners. The aim of SPIJ is to publish research which is not only technically proficient, but contains innovation or information for our international readers. In order to position SPIJ as one of the top International journal in signal processing, a group of highly valuable and senior International scholars are serving its Editorial Board who ensures that each issue must publish qualitative research articles from International research communities relevant to signal processing fields.

SPIJ editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for SPIJ. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. SPIJ provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
Signal Processing: An International Journal (SPIJ)

Editorial Board

Editor-in-Chief (EiC)

Associate Editors (AEiCs)

Editorial Board Members (EBMs)

Table of Contents

Volume 4, Issue 1, March 2010.

Pages

1 - 16	Frequency based criterion for distinguishing tonal and noisy spectral components **Maciej, Andrzej**
17 - 22	Improvement of minimum tracking in Minimum Statistics noise estimation method **Hassan Farsi**
23 - 37	A Novel Algorithm for Acoustic and Visual Classifiers Decision Fusion in Audio-Visual Speech Recognition System **Rajavel, P.S. Sathidevi**
38 - 53	A New Enhanced Method of Non Parametric power spectrum Estimation. **K.Suresh Reddy, S.Venkata Chalam, B.C.Jinaga**
54 - 61	A Combined Voice Activity Detector Based On Singular Value Decomposition and Fourier Transform **Jamal Ghasemi, Amard Afzalian, M.R. Karami Mollaei**
62 - 67	Reducting Power Dissipation in Fir Filter: an Analysis **Rakesh Kumar Bansal, Manoj Garg, Savina Bansal**

Frequency based criterion for distinguishing tonal and noisy spectral components

Maciej Kulesza m_kulesza@sound.eti.pg.pl
Multimedia Systems Department
Gdansk University of Technology
Gdansk, 80-233, Poland

Andrzej Czyzewski ac@sound.eti.pg.gda.pl
Multimedia Systems Department
Gdansk University of Technology
Gdansk, 80-233, Poland

Abstract

A frequency-based criterion for distinguishing tonal and noisy spectral components is proposed. For considered spectral local maximum two instantaneous frequency estimates are determined and the difference between them is used in order to verify whether component is noisy or tonal. Since one of the estimators was invented specially for this application its properties are deeply examined. The proposed criterion is applied to the stationary and nonstationary sinusoids in order to examine its efficiency.

Keywords: tonal components detection, psychoacoustic modeling, sinusoidal modeling, instantaneous frequency estimation.

1. INTRODUCTION

The algorithm responsible for distinguishing tonal from noisy spectral components is commonly used in many applications such a speech and perceptual audio coding, sound synthesis, extraction of audio metadata and others [1-9]. Since the tonal components present in a signal are usually of higher power than noise, the basic criterion for distinguishing tonal from noisy components is based on the comparison of the magnitudes of spectrum bins. Some heuristic rules may be applied to the local spectra maxima in order to determine whether they are noisy or tonal [1]. The other method relies on the calculation of terms expressing peakiness of these local maxima as it was proposed in [10] or level of similarity of a part of spectrum to the Fourier transform of stationary sinusoid, called sinusoidal likeness measure (SLM) [11]. In contrary to the magnitude-based criterions applied to the local spectra maxima, the ratio of geometric to arithmetic mean (spectral flatness measure – SFM) of magnitudes of spectrum bins may be used for tonality estimation of entire signal or for set of predefined bands [4, 5]. Instead of analysis of magnitude spectrum, it is also possible to extract the information related to the tonality of spectral components through comparison of the phase values coming from neighbouring bins as it was proposed in [12]. The method used in MPEG psychoacoustic model 2 employs linear prediction of phase and magnitude of spectrum bins. The tonality measure is then expressed as the difference between predicted values and the ones detected within particular time frame spectrum [1, 13-15]. Also various techniques for separation of periodic components within speech signal and signals composed of two pitched sounds were successfully investigated [3, 16-18].

Recently, it was proved that the tonality of spectral components within polyphonic recordings may be expressed as an absolute frequency difference between instantaneous frequencies of the local spectrum maxima calculated employing two different estimators [19-21]. While the first frequency estimator employs well known technique of polynomial fitting to the spectrum maximum and its two neighbouring bins, the second estimator is hybrid. It involves estimation results yielded by the first mentioned estimator and phase values coming down from three contiguous spectra. This algorithm was successfully combined with psychoacoustic model used in audio coding applications [13]. It was proved that this method allows detecting tonal spectra components even if they instantaneous frequency changes significantly over time. This property of the method is its main advantage over the tonality estimation algorithms commonly used in various applications. Although the efficiency of the mentioned algorithm has been already evaluated using artificial signals and polyphonic recordings, no investigation related to the hybrid frequency estimator and the tonality criterion being the basis for this method has been made. In this article we will focus on the experiments revealing properties of the hybrid frequency estimator and the properties of the tonality criterion employing it. The influence of the analysis parameters as well as influence of the analyzed signal characteristics on tonality estimation efficiency is investigated and deeply discussed. The properties of the hybrid estimator are compared to the properties of the estimator employing polynomial fitting to the spectral bins.

2. CONCEPT DESCRIPTION

For clarity of description, it is assumed here that the analyzed signal contains a single tonal component of constant or modulated instantaneous frequency and variable signal to noise ratio (SNR). A general diagram of the method used in order to examine the proprieties of proposed tonality criterion is shown in Fig.1.

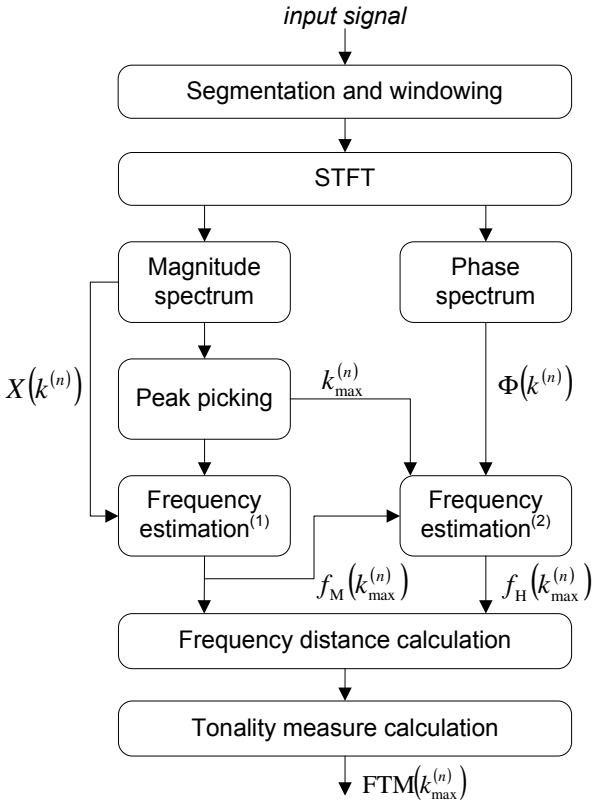


FIGURE 1: General diagram of investigated method for tonality measuring

The input signal is segmented into frames of equal length weighted by the von Hann window in conformity to the short time Fourier transform (STFT) concept [22]. Both the frame length and hop size are the parameters of the method. Moreover, the windowed frame of the signal is zero-padded before applying the FFT. Further, the magnitude and phase spectra denoted as $X\left(k^{(n)}\right)$ and $\Phi\left(k^{(n)}\right)$ are calculated, and spectral bin of highest magnitude is considered to be a candidate for the tonal component. The instantaneous frequency corresponding to the detected spectral component of highest energy (spectrum bin of $k_{\max}^{(n)}$ index) is then estimated using two methods. While the first one employs fitting of polynomial (binomial) to the detected component and its two adjacent bins within magnitude spectrum, the second one is based on the phase and magnitude–spectrum processing [23]. The results of frequency estimation obtained using two above-mentioned methods are denoted in Fig. 1 as $f_{\mathrm{M}}\left(k_{\max}^{(n)}\right)$ and $f_{\mathrm{H}}\left(k_{\max}^{(n)}\right)$. Finally, the absolute frequency difference is calculated and the level of tonality for selected component is assigned to it as a result of normalization of the yielded frequency distance (absolute frequency difference) by the assumed frequency distance threshold. The tonality measure calculated in accordance to the proposed scheme is called frequency-derived tonality measure (FTM).

2.1 Frequency estimator based on magnitude spectrum analysis

Assuming that the local maximum of magnitude spectrum being analyzed corresponds to the tonal component, the straightforward method for its instantaneous frequency estimation employs quadratic interpolation (known as QIFFT) which belongs to the approximate maximum likelihood (ML) estimators [24, 25]. In this approach the magnitude spectrum values of local maximum and two neighboring bins are involved in frequency estimator. The procedure is applied to the log spectrum values as it provides higher precision of frequency estimation in most cases [23, 26]. At the begining the fractional part of spectrum index is determined according to [27]

$$k_{\mathrm{frac}}^{(n)} = \frac{1}{2}\frac{X\left(k_{\max}^{(n)}-1\right)-X\left(k_{\max}^{(n)}+1\right)}{X\left(k_{\max}^{(n)}-1\right)-2X\left(k_{\max}^{(n)}\right)+X\left(k_{\max}^{(n)}+1\right)} \tag{1}$$

where $k_{\max}^{(n)}$ stands for the index of considered spectrum bin (the notation of spectrum bin indices is extended by the time index (number of frame) as superscript), $X\left(k_{\max}^{(n)}\right)$ represents the magnitude spectrum in log scale. The frequency of the spectrum peak detected in the *n*-th frame of signal is then estimated as follows

$$f_{\mathrm{M}}\left(k_{\max}^{(n)}\right) = \frac{k_{\max}^{(n)}+k_{\mathrm{frac}}^{(n)}}{N_{\mathrm{FFT}}} f_{\mathrm{s}} \tag{2}$$

where N_{FFT} is the length of FFT transform and f_{s} is the sampling rate in Sa/s (samples per second) and M in subscript indicates that the instantaneous frequency is estimated basing on magnitude spectrum processing. Since the signal frame is zero-padded before applying the FFT, the zero-padding factor is expressed as

$$Z_{\mathrm{p}} = \frac{N_{\mathrm{FFT}}}{N} \geq 1 \tag{3}$$

where *N* stands for the length of signal frame. The motivation for zero-padding of the signal frame before FFT calculation is the reduction of estimator bias resulting in an improved accuracy of frequency estimation. Basing on experimental results presented in [23], the maximum frequency bias of the QIFFT assuming the von Hann window is up-bounded in the following way

$$f_{\mathrm{Mbias}} \leq \frac{f_s}{N}\left(\frac{1}{4Z_p}\right)^3 \tag{4}$$

For zero-padding factor equal to 2 and frame length equivalent to 32 ms (for instance: f_s=32 kSa/s, N=1024) the bias of considered frequency estimator calculated according to (4) is less than 0.07 Hz. Using zero-padding factor higher than 2 seems to be impractical as it would result in significant increase of the computational complexity, assuring only slight increase of the frequency estimation accuracy. Thus, in investigated method for tonality measuring every frame of the input signal is zero-padded to its doubled length.

2.2 Hybrid frequency estimator

The second estimator suitable for combining with proposed method for tonal components detection and tonality estimation is required to:

- yield inadequate instantaneous frequency values when the spectrum bins involved into the estimator procedure do not correspond to the tonal components (the frequency distance between values obtained using quadratic interpolation and phase-based method should be abnormally high – i.e. higher than half of the frequency resolution of spectral analysis)

- allow of accurate instantaneous frequency estimation of frequency modulated tonal components

Various phase-based instantaneous frequency estimators have been proposed so far [28-32]. Assuming the STFT approach to the signal analysis, one of the straightforward methods for frequency estimation is based on an approach proposed in [28] where instantaneous frequency is computed basing on the phase difference between two successive frame short-term spectra. The hop size H equal to one sample is assumed in this method in order to allow for estimation of instantaneous frequency in full Nyquist band [32]. However, even if the analyzed spectrum maximum corresponds to the component totally noisy, the classic phase-difference estimator (assuming H=1) yields adequate instantaneous frequency estimates because the estimation error is lower than the frequency resolution of spectral analysis. Consequently, the first above-defined requirement for frequency estimator is not met. In order to overcome this problem, the higher hop size of STFT analysis should be used. When the higher hop size is chosen, the phase difference for particular frequency bin can be higher than 2π. In this case, the adequate phase increment cannot be calculated from the phase spectrum, as its values are bounded to ±π and then the phase difference never exceeds 2π. This causes the phase indetermination problem obstructing the instantaneous frequency estimation using classical phase-based method [22, 28, 32, 33]. Furthermore, when the higher hop size is selected the frequency of tonal component may be not longer constant in two successive steps of analysis or even the indices of spectral maxima corresponding to the same tonal component may be different $\left(k_{\max}^{(n)} \neq k_{\max}^{(n-1)}\right)$. Since the instantaneous frequency cannot be accurately determined in this case, the second requirement defined on the beginning of this subsection is not satisfied. Thus, the classical phase-difference estimator was not considered for employing it as an element in our method for tonal components detection. Although some phase-based methods for frequency estimation of nonstationary tonal components were already proposed in [30, 33], the proposed tonality estimation method is based on the dedicated estimator fulfilling the above-defined requirements and optimized for application considered here.

The instantaneous frequency determined by the hybrid estimator is defined as follows

$$f_{\mathrm{H}}\left(k_{\max}^{(n)}\right) = f_{\mathrm{M}}\left(k_{\max}^{(n-2)}\right) + \Delta f_{\Phi}^{(*)}\left(k_{\max}^{(n)}\right) \tag{5}$$

where: $f_{\mathrm{M}}\left(k_{\max}^{(n-2)}\right)$ is the instantaneous frequency of the spectrum maximum detected within *n*–2 analysis frame using estimator defined in Eq. (2), and $\Delta f_{\Phi}^{(*)}\left(k_{\max}^{(n)}\right)$ is the frequency jump between spectral maxima detected within *n*–2 and next *n* analysis frames estimated using phase-based method.

2.2.1 Phase-based frequency jump estimator

In the investigated method the frequency jump $\Delta f_{\Phi}^{(*)}\left(k_{\max}^{(n)}\right)$ is calculated basing on the phase values detected within three successive spectra. It is assumed that the phase values $\Phi\left(k_{\max}^{(n-2)}\right)$, $\Phi\left(k_{\max}^{(n-1)}\right)$ and $\Phi\left(k_{\max}^{(n)}\right)$ correspond to the same sinusoidal component detected within three contiguous spectra. The second order phase difference is then calculated according to [19]

$$\Delta^2\Phi\left(k_{\max}^{(n)}, k_{\max}^{(n-2)}\right) = \Phi\left(k_{\max}^{(n-2)}\right) - 2\Phi\left(k_{\max}^{(n-1)}\right) + \Phi\left(k_{\max}^{(n)}\right) \tag{6}$$

The phase offset which is non-zero in case of frequency modulated tonal components is given by

$$\Delta^2\phi\left(k_{\max}^{(n)}, k_{\max}^{(n-2)}\right) = \frac{\pi(N-1)}{Z_{\mathrm{p}}N}\left(k_{\max}^{(n-2)} - 2k_{\max}^{(n-1)} + k_{\max}^{(n)}\right) \tag{7}$$

Finally, the frequency jump can be estimated using following formula

$$\Delta f_{\Phi}\left(k_{\max}^{(n)}\right) = \frac{f_s}{\pi H}\left(\mathrm{princarg}\left(\Delta^2\Phi\left(k_{\max}^{(n)}, k_{\max}^{(n-2)}\right)\right) + \Delta^2\phi\left(k_{\max}^{(n)}, k_{\max}^{(n-2)}\right)\right) \tag{8}$$

where $\mathrm{princarg}(\varphi) = (\varphi + \pi)\bmod(-2\pi) + \pi$ is the function mapping the input phase φ into the ±π range [22]. Further the $\Delta f_{\Phi}\left(k_{\max}^{(n)}\right)$ is updated in order to overcome phase ambiguity problem [19]

$$\Delta f_{\Phi}^{(*)}\left(k_{\max}^{(n)}\right) = \Delta f_{\Phi}\left(k_{\max}^{(n)}\right) + m\frac{f_{\mathrm{s}}}{H} \tag{9}$$

where *m* is the integer value ensuring that the $\Delta f_{\Phi}^{(*)}\left(k_{\max}^{(n)}\right)$ falls within the maximal and minimal frequency jump range related to the $k_{\max}^{(n)} - k_{\max}^{(n-2)}$ difference [19].

2.3 Tonality measurements

The proposed criterion for distinguishing tonal from noisy components and their tonality measuring is based on the absolute difference between the frequency estimates obtained using the QIFFT method and the hybrid method described in previous subsection. Thus, the frequency distance for particular spectral maximum is given by

$$f_{\Delta}\left(k_{\max}^{(n)}\right) = f_{\mathrm{M}}\left(k_{\max}^{(n)}\right) - f_{\mathrm{H}}\left(k_{\max}^{(n)}\right) \tag{10}$$

When we combine the estimate (5) with the definition (10) the frequency distance may be expressed by

$$f_{\Delta}\left(k_{\max}^{(n)}\right) = f_{\mathrm{M}}\left(k_{\max}^{(n)}\right) - f_{\mathrm{M}}\left(k_{\max}^{(n-2)}\right) - \Delta f_{\Phi}\left(k_{\max}^{(n)}\right) \tag{11}$$

It is viewable that $f_{\Delta}\left(k_{\max}^{(n)}\right)$ is equal to the difference between frequency jumps derived from the magnitude spectrum analysis and from the phase spectrum analysis, respectively [19]. Let us define a measure based on the frequency distance $f_{\Delta}\left(k_{\max}^{(n)}\right)$ expressing the level of similarity of particular spectrum component to the pure sinusoid

$$\mathrm{FTM}\left(k_{\max}^{(n)}\right)=1-\frac{\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|}{\left|f_{\Delta}\right|_{\mathrm{thd}}} \tag{12}$$

where $\left|f_{\Delta}\right|_{\mathrm{thd}}$ is a frequency distance threshold which is assumed not to be exceeded when the $k_{\max}^{(n)}$ is a tonal spectral component. Tonality measure $\mathrm{FTM}\left(k_{\max}^{(n)}\right)$ is equal to 1 if spectral component considered corresponds to the sinusoid of high SNR and tends to gradually decrease when SNR falls. If $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|\geq\left|f_{\Delta}\right|_{\mathrm{thd}}$ for a particular spectral component, it is treated as a noisy one, and the tonality measure $\mathrm{FTM}\left(k_{\max}^{(n)}\right)$ equal to 0 is assigned to it. The experiments related to the properties of hybrid frequency estimator proposed here together with the criterion for tonal components detection as well as some remarks concerning selection of $\left|f_{\Delta}\right|_{\mathrm{thd}}$ threshold are presented in the following section.

3. EXPERIMENTS

3.1 The performance evaluation of instantaneous frequency estimators

In order to examine the properties of the proposed hybrid estimator, a set of real valued sinusoids with randomly chosen initial phases φ_0 and SNR ranging from 100 dB to -20 dB with 2 dB step were generated. It was assumed that the amplitude of sinusoid is equal to 1 and the power of noise is adjusted in order to achieve a desired SNR in dB according to the formula

$$\mathrm{SNR}[\mathrm{dB}]=10\log_{10}\frac{\sum_{s=1}^{L}x_{\mathrm{t}}^{2}[s]}{\sum_{s=1}^{L}x_{\mathrm{ns}}{}^{2}[s]} \tag{13}$$

where $x_{\mathrm{t}}[s]=a\cos(2\pi\omega s+\varphi_0)$, $\omega=f/f_{\mathrm{s}}$ is the normalized frequency in cycles per sample, $x_{\mathrm{ns}}[s]$ stands for a additive white Gaussian noise (AWGN) realization, *s* is sample number and *L* is the signal length.

For every selected SNR the sinusoids of constant normalized frequencies selected within range from 0.05 to 0.45 with 0.005 step (81 sinusoids) were generated and further analyzed resulting in vector of instantaneous frequency estimates related to the particular SNR. Then, the mean squared error (MSE) of estimates (2) and (5) was calculated basing on frequency estimation results and known apriori frequencies of generated sinusoids. Since this procedure was applied to sets of sinusoids of various SNR, the characteristic revealing frequency estimation errors versus SNR of analyzed sinusoids was obtained. The experiments were carried out for both considered estimators – the hybrid method and the QIFFT method, and the results were compared with lower Cramer-Rao bound (CRB) defining variance of unbiased frequency estimator of real sinusoid in a AWGN [25, 32]

$$\mathrm{var}(\hat{\omega})\geq\frac{12}{(2\pi)^2a^2N\left(N^2-1\right)}10^{-SNR/10} \tag{14}$$

where $\hat{\omega}$ is the normalized estimated frequency in cycles per sample, N is the same as in (3) and a=1 in our experiments.

The sampling rate of analyzed signals was adjusted to 8000 Sa/s, the frame length (von Hann window) was equal to 32 ms (N=256) and the hop size was switched between 32 ms (H=256) and 8 ms (N=64). The characteristics obtained for two above-defined hop sizes of analysis are presented in Fig. 2.

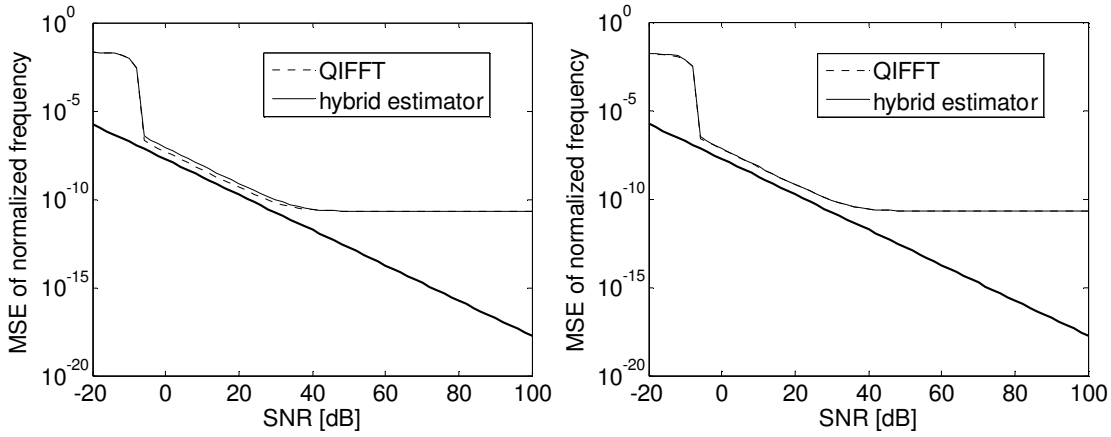


FIGURE 2: Performance of estimators for frequencies in (0.05, 0.45) normalized range for hop size equal to frame length (left), and quarter of frame length (right); Cramer-Rao bound – bold solid line

Since the spectrum bin of maximum energy is considered here to represent sinusoidal component, for lower SNRs the spurious noise peak may be detected instead of it. Thus, the frequency estimation MSEs presented in Fig. 2 are far beyond the CRB when the SNRs of sinusoids are lower than approximately -10 dB [23, 25]. Contrarily, in the SNR range from -10 dB to approximately 30 dB the MSE determined for examined estimators is close to the CRB. Although the curve representing results obtained using the QIFFT is approximately 4 dB above CRB regardless the hop size of analysis, the error of hybrid estimator tends to be slightly higher when the hop size is maximal possible. For SNRs higher than 40 dB the frequency estimation error reveals the bias of concerned estimators, which is related to the assumed zero-padding factor [23, 26].

The influence of the hop size on the estimation error in case of stationary sinusoid of SNR equal to 20 dB and 100 dB and normalized frequency equal to 0.13 is presented in Fig. 3 (sampling rate is the same as in previous experiment). It can be observed from Fig. 3 that the MSE of hybrid estimator is practically identical to the MSE obtained using the QIFFT regardless the hop size of analysis when the SNR is equal to 100 dB (compare results presented in Fig. 3 for the same SNR=100 dB). However, when the SNR is equal to 20 dB, the hybrid estimator performs slightly worse, by approximately 3 dB, than the QIFFT for hop sizes higher than a half of the frame length. For lower hop sizes the difference in performance of both estimators gradually decreases. It can be expected that for shorter hop sizes, the frequency change $\Delta f_{\Phi}^{(*)}\left(k_{\max}^{(n)}\right)$ derived from phase analysis according to (9) tends to have lower influence on the final estimation results. Thus, the shorter the hop size the properties of hybrid estimator are closer to the properties of the QIFFT method. This is not the case when the SNR is equal to -3 dB or lower, because the MSE of hybrid method tends to increase for the hop sizes below approximately a quarter of the frame length and higher than 220 samples. In this hop size range the hybrid method yields occasionally inadequate estimates when the SNR is low resulting in the MSE increase. Therefore, it can be deduced that the hybrid estimator operates most efficiently in the range of the hop size between approximately ¼ to ¾ of the frame length.

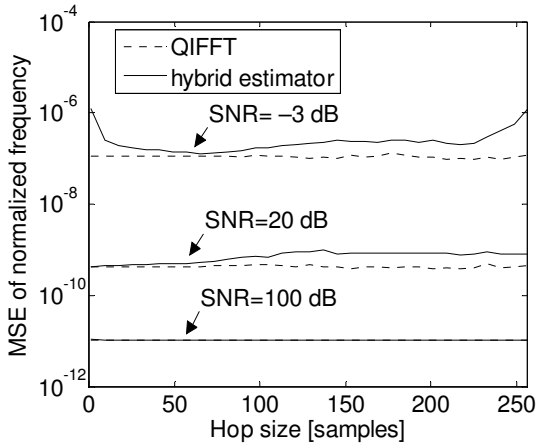


FIGURE 3: Impact of the hop size of analysis on the frequency estimation performance

Further, the MSE of frequency estimation results were determined for sinusoids of constant SNR equal to 100 dB and for normalized frequencies selected within 0.0025 and 0.4975 range with 0.001 step (496 sinusoids, f_s=8000 Sa/s, N=256, H=256). The results of our experiments presented in Fig. 4 indicate that the estimation errors for both considered methods are below 10^{-10} (see also Fig. 2) in almost entire bandwidth. However, when the normalized frequency of considered sinusoid is below approximately 0.005 or higher than 0.495 then the estimation error significantly increases.

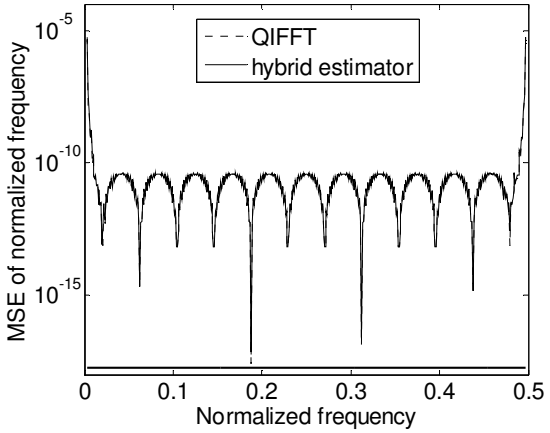


FIGURE 4: Performance of estimators for stationary sinusoids of SNR=100 dB and normalized frequencies selected within 0.0025 and 0.4975 range; Cramer-Rao bound – bold solid line.

Although the characteristics shown in Fig. 4 were determined assuming hop size equal to half of the frame length, they do not alter for other hop sizes. This is expected when considering the MSE obtained for stationary sinusoids of SNR equal to 100 dB presented in Fig. 3.

Since it is assumed that the proposed hybrid estimator should allow estimation of instantaneous frequency of non-stationary tonal components (see subsection 2.2), the set of linearly frequency modulated (LFM) chirps were analysed next [34]. The parameters of the STFT analysis as well as the sampling rate were identical to those used in the previous experiment described in this subsection. The initial normalized frequency of every LFM chirp signal was set to 0.05 and the frequency rates were altered from 0 to $f_s/2$ per second. The instantaneous frequencies estimated using the QIFFT and hybrid methods were compared with mean frequency values of LFM chirp calculated within a particular frame resulting in the MSE corresponding to the chirps of various

instantaneous frequency slopes. The experiments were carried out for LFM chirps of SNR equal to 100 dB and 20 dB. Although the limitations of the hybrid estimator when the hop size is equal to the frame length have been already revealed (see Fig. 3), in the experiments the hop size of analysis was chosen to be equal to frame length and a quarter of it for comparison purposes. In Fig. 5 the characteristics obtained are shown.

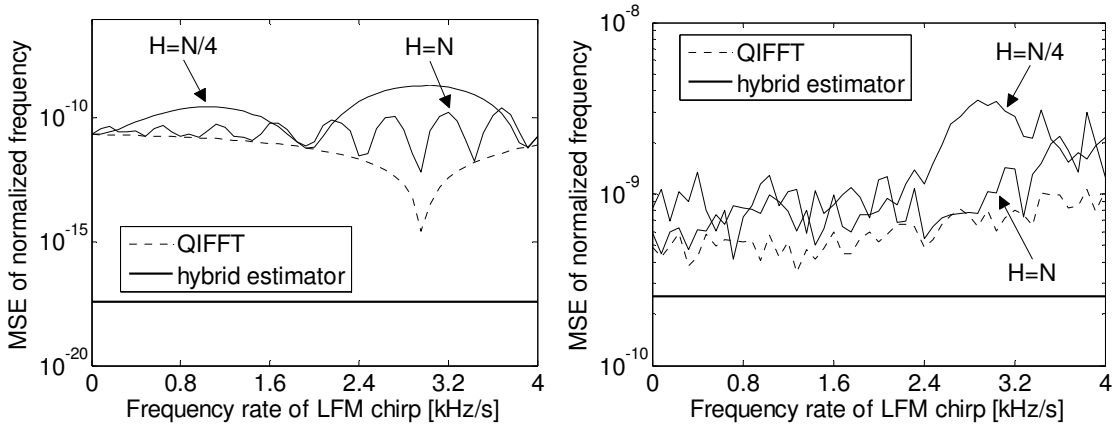


FIGURE 5: Performance of estimators for LFM chirps of various frequency rates and SNR equal to 100 dB (left) and 20 dB (right); Cramer-Rao bound – bold solid line.

When the hop size is equal to the quarter of the frame length the estimation error is higher for some chirps slopes (0.3-0.45) than the errors obtained with hop size equal to the frame length which is especially noticeable when considering characteristics obtained for signals of SNR equal to 100 dB. Furthermore, when SNR is equal to 20 dB (Fig. 5 - right), the errors corresponding to both estimation procedures are still close to the Cramer-Rao bound regardless the linear frequency modulation of analysed sinusoids.

Although the above experiments have confirmed that proposed hybrid estimator operates properly in case of sinusoids of linearly changing frequencies, its properties were also examined in case of non-linearly frequency modulated sinusoids. Thus, the frequency of carrier sinusoid equal to 0.13 (1040 Hz assuming f_s=8000 Sa/a) was modulated using sinusoid of normalized frequency equal to 2.5×10^{-4} (2 Hz). The modulation depth was altered so that the normalized frequency deviation of the carrier was changed between 0 and 0.025 (±200 Hz). Similarly to the experiments with LFM chirps the MSE of frequency estimates were determined for all generated sinusoids of SNR equal to 100 dB and 20 dB. The frame length was adjusted to 32 ms (N=256) and the hop size was switched between 32 ms and 8 ms (H=256, H=64). The results of those experiments are depicted by the curves shown in Fig. 6.

It can be noticed from Fig. 6 that the accuracy of frequency estimation is directly related to the depth of non-linear frequency modulation. The modulation depth seems to have less influence on the MSE for signals of lower SNRs, which is visible when comparing results obtained for sinusoids having the SNR of 100 dB and 20 dB. Additionally, when the framing hop size is short enough the performance of the QIFFT and hybrid estimators tends to be similar to each other.

It was suggested in subsection 2.2 that the desired property of estimator for application considered would be yielding inadequate frequency estimates when spectrum bins used in estimator do not correspond to sinusoidal component. In order to evaluate this property of proposed hybrid estimator, the white noise realization was analysed and in every spectrum the local maximum $k_{\max}^{(n)}$ laying closest to 800 Hz was selected.

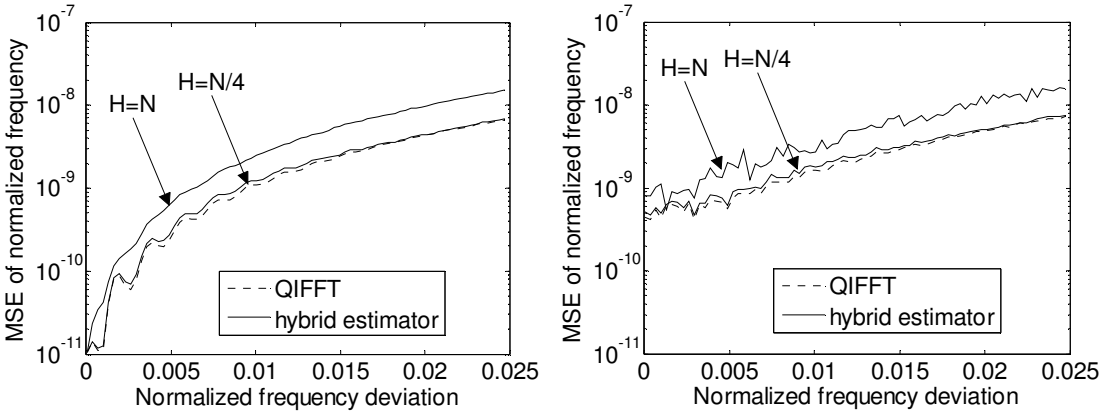


FIGURE 6: Performance of estimators for sinusoids of sinusoidally modulated frequencies of SNR equal to 100 dB (left) and 20 dB (right).

The QIFFT was applied to those peaks as well as the hybrid estimation method was used. Next, the frequencies estimated using these two methods were compared with frequency corresponding to detected local maximum

$$f_{\mathrm{b}}\left(k_{\max}^{(n)}\right)=\frac{k_{\max}^{(n)}}{N_{\mathrm{FFT}}}f_{\mathrm{s}} \tag{15}$$

The absolute frequency differences $\left|f_{\mathrm{b}}\left(k_{\max}^{(n)}\right)-f_{\mathrm{H}}\left(k_{\max}^{(n)}\right)\right|$ and $\left|f_{\mathrm{b}}\left(k_{\max}^{(n)}\right)-f_{\mathrm{M}}\left(k_{\max}^{(n1)}\right)\right|$ calculated for estimation results obtained in every frame of white noise realization (f_s=8000 Sa/s, frame length and hop size equal to 32 ms (N=256, H=256), signal length equal to 2 s) are presented in Fig. 7.

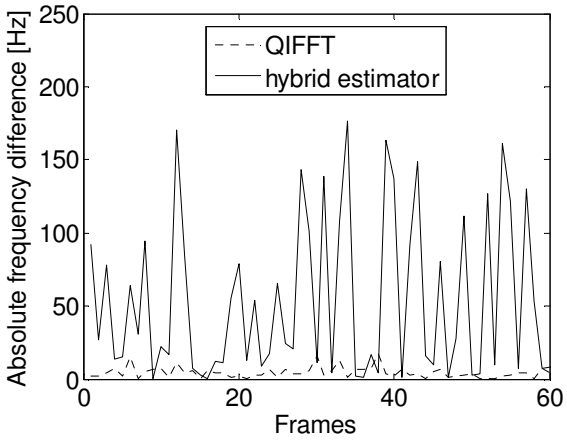


FIGURE 7: Absolute frequency differences between frequency of spectrum bin calculated according to Eq. (20) and estimates obtained using the QIFFT and hybrid method (noisy spectral peaks)

The maximum difference between frequency of spectrum local maximum defined by Eq. (15) and obtained using the QIFFT estimator is bounded to a half of the apparent frequency resolution of spectral analysis. Therefore, the curve depicting results yielded by the QIFFT estimator presented in Fig. 7 never exceeds $f_s/(2N_{\mathrm{FFT}})$=8000/512=15.625 Hz. Contrary to the QIFFT, the instantaneous frequency estimates yielded by the hybrid method are usually totally inadequate and are not bounded to the half of the apparent frequency resolution of spectral analysis. It can be concluded than that proposed hybrid estimator satisfies both requirements defined on the beginning of subsection 2.2, because it allows for frequency estimation of the modulated tonal

components and provides totally inadequate results when the selected spectral maxima do not correspond to the tonal components. Although additional experiments may be carried out in order to examine the properties of proposed hybrid estimator more deeply (i.e. estimation accuracy in case of complex sinusoids, influence of frame length and segmentation window type used, etc.), we have focused here only on the verification of those properties which are of primary importance for considered application.

3.2 Tonality measurements

In order to verify the concept of employing two different instantaneous frequency estimators for tonality measuring, a signal having a frequency modulated sinusoidal component of varying SNR was considered. As the spectrum bin of highest energy may not represent the tonal component when the SNR is very low (see Fig. 2), in our experiments the lowest SNR was adjusted to -3 dB. In the experiment the analysis was applied to the signal sampled at 8000 Sa/s rate, consisting of 24000 samples. The SNR of the sinusoidal components was constant in every segment of 8000 samples and equal to 30 dB, 3 dB and -3 dB, respectively. The instantaneous frequencies of tonal component were estimated within 32 ms frames of the signal (N=256) and the hop size was adjusted to 16 ms (H=128). The spectrogram of analyzed signal together with a curve representing the true pitch of sinusoid, the results of instantaneous frequency estimation employing two considered estimators and the frequency distance calculated according to (10) are presented in Fig. 8.

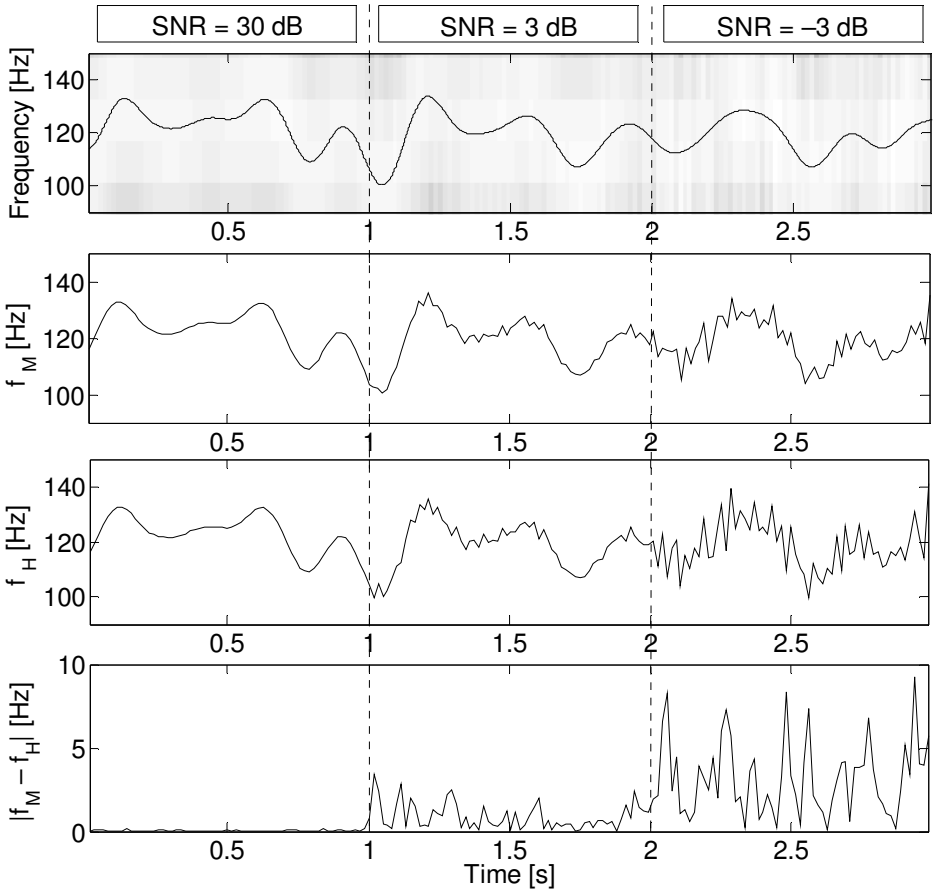


FIGURE 8: Looking from the top: a part of spectrogram together with a curve representing instantaneous frequencies of tonal component, estimated frequencies using the QIFFT and hybrid method, and absolute frequency difference calculated according to (10).

It can be noted that when the SNR is equal to 30 dB the instantaneous frequencies estimated using the QIFFT and hybrid estimates are close to each other resulting in negligible $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|$ values. However, when the SNR decreases, the $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|$ distance tends to have a higher mean value. This observation confirms that the absolute difference between frequencies estimated using the QIFFT and the hybrid method can be used as a measure of spectral components tonality [9].

Next, the influence of the hop size on the mean and maximum frequency distance $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|$ was examined. The single sinusoidal component of -3 dB SNR and constant frequency equal to 800 Hz (sampling rate 8000 Sa/s) was generated and further analysed with the hop size ranging from 0.125 ms (H=1) to 32 ms (H=256) with 1 ms (8 samples) step. For every selected hop size of the STFT analysis the arithmetic mean and maximum value of the vector containing all $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|$ values corresponding to the considered tonal component was calculated. The results are shown in Fig. 9.

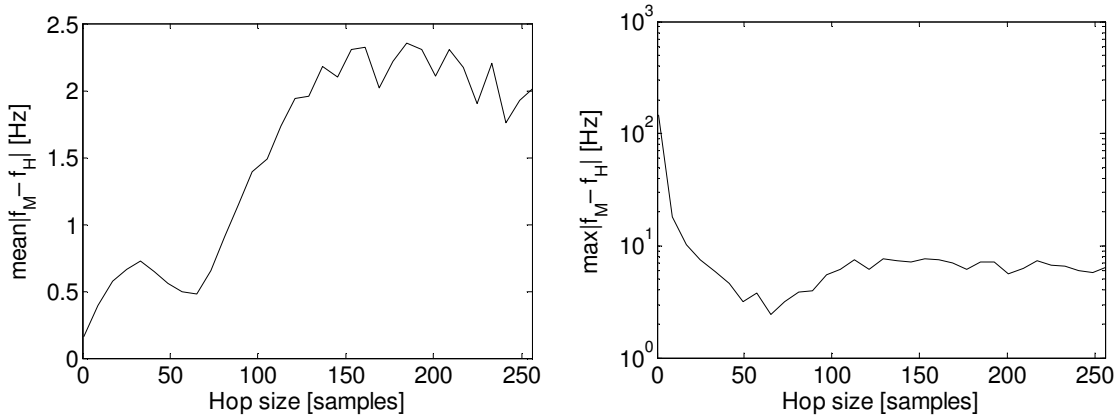


FIGURE 9: The mean (left) and maximum (right) frequency distances $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|$ obtained for sinusoids of constant frequency and SNR= -3 dB SNR analyzed with various hop sizes

The maximum value of frequency distance is the highest for hop size equal to one sample and decreases while the hop size increases to approximately N/4. This phenomenon is related to the properties of hybrid estimator which yields occasionally inadequate frequency estimates when the sinusoidal component of low SNR is analysed. Additionally, in the above-mentioned hop size range, the mean value of frequency distance is rather low. Thus, taking into account also computational complexity of the algorithm, the hop sizes below quarter of the frame length should be avoided.

Considering hop size range from 60 to about 150 samples it can be observed, that the mean value of $\left|f_{\Delta}\left(k_{\max}^{(n)}\right)\right|$ rises monotonically and then saturates beyond 2 Hz level. Adequately, the maximum value of frequency distance increases up to about 9 Hz, but saturates for hop size equal to approximately a half of the frame length. While the maximum values seem to be almost constant for higher hop sizes, the mean values tend to even slight decrease for the hop sizes longer than 200 samples. Therefore, the proposed criterion for tonal components detection and their tonality estimation would operate most efficiently when the hop size would be selected within range between ¼ to approximately ¾ of the frame length in the analysis. This observation is coherent with conclusions related to the results presented in Fig. 3. Although the curves presented in Fig. 9 would slightly vary depending on the frequency of analysed sinusoid, their

major character would be retained. Therefore, the presented considerations tend to be valid regardless the frequency of tonal component.

In order to determine the tonality measure of a particular spectral component according to (12) the appropriate value of $|f_\Delta|_{\text{thd}}$ threshold must be selected. This threshold must be at least as high as the maximum value of frequency distance yielded by the algorithm providing that the tonal component is analysed. Since the maximum value of the frequency distance depends on the chosen hop size *H* (see Fig. 9) threshold $|f_\Delta|_{\text{thd}}$ may be selected in accordance to it. However, in the proposed approach it is assumed to be constant regardless the selected *H* value of the STFT analysis. Actually it was selected to be a half of the frequency width corresponding to a bin of zero-padded spectrum

$$|f_\Delta|_{\text{thd}} = \frac{f_s}{2N_{\text{FFT}}} \tag{19}$$

Further, a set of stationary and frequency modulated sinusoids of nominal frequency equal to 120 Hz and SNR values ranging from 100 dB to -20 dB with 2 dB step were generated and analyzed. The frequency deviation of modulated sinusoid was set to 20 Hz and the carrier frequency was modulated using sinusoid of 3 Hz frequency. The sampling rate was equal to 8000 Sa/s, the frame length was selected to be equal to 32 ms (*N*=256) and hop size was adjusted to 16 ms (*H*=128). Since the length of every analysed signal was equal to 3 s, resulting in a vector of FTM values corresponding to the sinusoid of a particular SNR, the arithmetic mean values of these vectors were determined. The results of experiment are presented in Fig. 10.

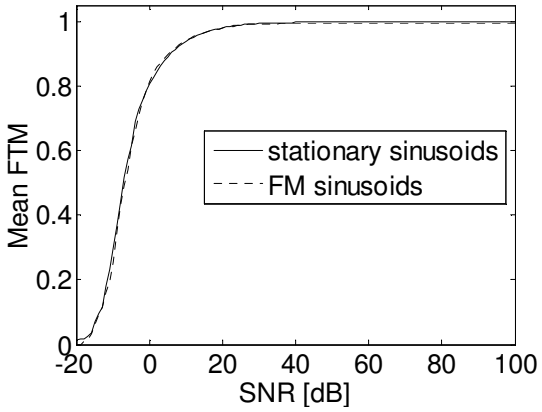


FIGURE 10: Mean values of FTM determined for pure and frequency modulated sinusoids of various SNR

The mean FTM for tonal component of the SNR higher than approximately 40 dB is equal or close to the value of 1, because the instantaneous frequencies estimated using estimators (2) and (5) are almost identical to each other. In the SNR range from 40 dB to -20 dB the mean FTM values gradually decrease indicating lower tonality of the considered spectral component. It can be observed that when the tonal component is totally masked with noise which is the case when SNR is equal to -20 dB, the FTM is close to the value of 0. This confirms that the proposed tonality criterion is efficient in terms of distinguishing tonal from noisy spectral components. Additionally, the curves representing the mean FTM for a pure sinusoid and a frequency modulated one are practically identical to each other indicating that frequency modulation does not affect significantly the tonality measurements provided by the proposed method.

4. CONCLUSIONS

A criterion for distinguishing tonal from noisy spectral components based on a comparison of their instantaneous frequencies estimated using two different methods was proposed and evaluated. Since one of the estimators was specially developed for application considered, the experiments revealing its properties were carried out. It was shown that the proposed hybrid estimator provides satisfactory accuracy of frequency estimation in case of the analysis of pure and modulated sinusoidal components. Regardless the way the tonal components changes its frequency (linearly or periodically) the MSE of the frequency estimation remains below reasonable threshold for the hybrid method. However, it yields inadequate estimation results when the spectral component corresponds to a noise. These two above-mentioned properties of the estimator engineered here were found to be essential for application of the developed tonality criterion (FTM). The experiments revealed that the absolute difference between frequencies estimated using the QIFFT method and the hybrid one is directly related to the SNR of the sinusoids analysed. It was shown that the investigated algorithm operates most efficiently when the hop size of analysis is chosen between ¼ to ¾ of the frame length. The experimental results proved that characteristics of FTM values versus SNR of sinusoidal component are almost identical to each other whenever the sinusoid of constant or modulated instantaneous frequency is analysed. The presented tonality measure may substitute the tonality estimators employed so far in the psychoacoustic models and may be used also in various applications requiring tonal components detection.

5. ACKNOWLEDGMENTS

Research subsidized by the Polish Ministry of Science and Higher Education under grant PBZ MNiSW-02/II/2007 and N N517 378736.

6. REFERENCES

[1] ISO/IEC MPEG, "*IS11172-3 Coding of moving pictures and associated audio for digital storage media up to 1.5 Mbit/s*", Part 3: Audio, Annex D. ISO/IEC JTCI, 1992.

[2] ISO/IEC "*13818-7 Information technology — Generic coding of moving pictures and associated audio information*", Part 7: Advanced Audio Coding (AAC), 4th edition, 2006.

[3] M.G. Christensen, A. Jakobsson, "*Multi-Pitch Estimation*". Synthesis Lectures on Speech and Audio Processing, 5(1):1-160, 2009.

[4] J. D. Johnston, "*Transform coding of audio signals using perceptual noise criteria*". IEEE J. on Selected Areas in Comm., 6:314-323, 1988.

[5] O. Hellmuth, E. Allamanche, J. Herre, T. Kastner, M. Cermer, W. Hirsch, "*Advanced audio identification using MPEG-7 content description*". In proceedings of 111th Audio Eng. Soc. Int. Conf., New York, USA, 2001.

[6] R. J. McAulay, T. F. Quatieri, "*Speech analysis/synthesis based on a sinusoidal representation*". IEEE Transactions on Acoustics, Speech, and Signal Processing, 34:744-754, 1986.

[7] S. Levine, J. O. Smith III, "*Improvements to the switched parametric & transform audio coder*". In proceedings of IEEE Workshop on Application of Signal Processing to Audio and Acoustics, New York, USA, 1999.

[8] T. Painter, A. Spanias, "*Perceptual coding of digital audio*". In proceedings of. of IEEE, 88:451-513, 2002.

[9] S.-U. Ryu, K. Rose, "*Enhanced accuracy of the tonality measure and control parameter extraction modules in MPEG-4 HE-AAC*". In proceedings of 119th Audio Eng. Soc. Int. Conf., New York, USA, 2005.

[10] K. Lee, K. Yeon, Y. Park, D. Youn, *"Effective tonality detection algorithm based on spectrum energy in perceptual audio coder*". In proceedings of 117th Audio Eng. Soc. Int. Conf., San Francisco, USA, 2004.

[11] X. Rodet, "*Musical sound signal analysis/synthesis: sinusoidal+residual and elementary waveform models*". In proceedings of IEEE Time-Frequency and Time-Scale Workshop, Coventry, Grande Bretagne, 1997.

[12] A. J. S. Ferreira, "*Tonality detection in perceptual coding of audio*". In proceedings of 98th Audio Eng. Soc. Int. Conf., Paris, France, 1995.

[13] M. Kulesza, A. Czyzewski, „*Audio codec employing frequency-derived tonality measure*". In proceedings of 127th Audio Eng. Soc. Int. Conf., New York, USA, 2009.

[14] M. Kulesza, A. Czyzewski, "*Novel approaches to wideband speech coding*". GESTS Int. Trans. On Computer Science and Engineering, 44(1):154-165, 2008.

[15] D. Schulz, "*Improving audio codecs by noise substitution*". J. Audio Eng. Soc., 44:593-598, 1996.

[16] P. J. B. Jackson, C. H. Shadle, "*Pitch-scaled estimation of simultaneously voiced and turbulence-noise components in speech*". IEEE Trans. On Speech and Audio Processing, 9:713-726, 2001.

[17] Y. Wang, R. Kumaresan, "*Real time decomposition of speech into modulated components*". J. Acoust. Soc. Am., 119(6):68-73, 2006.

[18] B. Yegnanarayana, C. Alessandro, V. Darisons, "*An iterative algorithm for decomposition of speech signals into periodic and aperiodic components*". IEEE Trans. on Speech and Audio Proc., 6: 1-11, 1998.

[19] M. Kulesza, A. Czyzewski, "*Tonality Estimation and Frequency Tracking of Modulated Tonal Components*". J. Audio Eng. Soc., 57(4):221–236, 2009.

[20] G. Peeters, X. Rodet, "*Signal characterization in terms of sinusoidal and non-sinusoidal components*". In proceedings of Digital Audio Effects (DAFx) Conf., Barcelona, Spain, 1998.

[21] G. Peeters, X. Rodet, "*SINOLA: a new analysis/synthesis method using spectrum peak shape distortion, phase and reassigned spectrum*". In proceedings of the Int. Computer Music Conf., Beijing, China, 1999.

[22] U. Zolzer, "*DAFX Digital Audio Effects*". John Wiley & Sons, United Kingdom, 2002.

[23] M. Abe, J. O. Smith III, "*Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT magnitude peaks*". In proceedings of 117th Audio Eng. Soc. Int. Conf., San Francisco, USA, 2004.

[24] F. Keiler, S. Marchand, "*Survey on extraction of sinusoids in stationary sound*". In proceedings of the 5th Int. Conf. on Digital Audio Effects (DAFx-02), Hamburg, Germany, 2002.

[25] D. C. Rife, R. R. Boorstyn, "*Single-tone parameter estimation from discrete-time observations*". IEEE Trans. Info. Theory, 20(5):591-598, 1974.

[26] M. Betser, P. Collen, G. Richard, B. David, "*Preview and discussion on classical STFT-based frequency estimators*. In proceedings of 120th Audio Eng. Soc. Int. Conf., Paris, France, 2006.

[27] J.C. Brown, M.S. Puckette, "*A high resolution fundamental frequency determination based on phase changes of the Fourier transform*". J. Acoust. Soc. Am., 94(2):662-667, 1998.

[28] J. Flanagan, R. Golden, "*Phase vocoder*". Bell Syst. Tech. J., 45:1493–1509, 1966.

[29] F.J. Charpentier, "*Pitch detection using the short-term Fourier transform*". In proceedings of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 11:113-116, Tokyo, 1986.

[30] S.W. Lang, B.R Musicus, "*Frequency estimation from phase difference*". In proceedings of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 4:2140-2143, United Kingdom, 1989.

[31] M.S. Puckette, J.C. Brown, "*Accuracy of frequency estimates using the phase vocoder*". IEEE Trans. On Speech and Audio Processing, 6(2):166-176, 1998.

[32] M. Lagrange, S. Marchand, "*Estimating the instantaneous frequency of sinusoidal components using phase-based methods*". J. Audio Eng. Soc., 55:385-399, 2007.

[33] M. Betser, P. Collen, G. Richard, B. David, "*Estimation of frequency for AM/FM models using the phase vocoder framework*". IEEE Trans. On Signal Processing, 56(2):505-517, 2008.

[34] M. Lagrange, S. Marchand, J.B. Rault, "*Sinusoidal parameter extraction and component selection in non stationary model*". In proceedings of Digital Audio Effects (DAFx) Conf., Hamburg, Germany, 2002.

Improvement of Minimum Tracking in Minimum Statistics Noise Estimation Method

Hassan Farsi hfarsi@birjand.ac.ir
Department of Electronics and Communications Engineering,
University of Birjand,
Birjand, IRAN.

Abstract

Noise spectrum estimation is a fundamental component of speech enhancement and speech recognition systems. In this paper we propose a new method for minimum tracking in Minimum Statistics (MS) noise estimation method. This noise estimation algorithm is proposed for highly non-stationary noise environments. This was confirmed with formal listening tests which indicated that the proposed noise estimation algorithm when integrated in speech enhancement was preferred over other noise estimation algorithms.

Keywords: Speech enhancement, Statistics noise, noise cancellation, Short time Fourier transform

1. INTRODUCTION

Noise spectrum estimation is a fundamental component of speech enhancement and speech recognition systems. The robustness of such systems, particularly under low signal-to-noise ratio (SNR) conditions and non-stationary noise environments, is greatly affected by the capability to reliably track fast variations in the statistics of the noise. Traditional noise estimation methods, which are based on voice activity detectors (VAD's), restrict the update of the estimate to periods of speech absence.

Additionally, VAD's are generally difficult to tune and their reliability severely deteriorates for weak speech components and low input SNR [1], [2], [3]. Alternative techniques, based on histograms in the power spectral domain [4], [5], [6], are computationally expensive, require much memory resources, and do not perform well in low SNR conditions. Furthermore, the signal segments used for building the histograms are typically of several hundred milliseconds, and thus the update rate of the noise estimate is essentially moderate.

Martin (2001)[7] proposed a method for estimating the noise spectrum based on tracking the minimum of the noisy speech over a finite window. As the minimum is typically smaller than the mean, unbiased estimates of noise spectrum were computed by introducing a bias factor based on the statistics of the minimum estimates. The main drawback of this method is that it takes slightly more than the duration of the minimum-search window to update the noise spectrum when the noise floor increases abruptly. Moreover, this method may occasionally attenuate low energy phonemes, particularly if the minimum search window is too short [8]. These limitations can be overcome, at the price of significantly higher complexity, by adapting the smoothing parameter and the bias compensation factor in time and frequency [9]. A computationally more efficient minimum tracking scheme is presented in [10]. Its main drawbacks are the very slow update rate of the noise estimate in case of a sudden rise in the noise energy level, and its tendency to cancel the signal [1].In this paper we propose a new approach for minimum tracking , resulted improving the performance of MS method.

The paper is organized as follows. In Section II, we present the MS noise estimator. In Section III, we introduce an method for minimum tracking, and in section IV, evaluate the proposed method, and discuss experimental results, which validate its effectiveness.

2. MINIMUM STATISTICS NOISE ESTIMATOR

Let x(n) and d(n) denote speech and uncorrelated additive noise signals, respectively, where n is a discrete-time index. The observed signal y(n), given by y(n)=x(n)+d(n), is divided into overlapping frames by the application of a window function and analyzed using the short-time Fourier transform (STFT). Specifically,

$$Y(k,l) = \sum_{n=0}^{N-1} y(n+lM)h(n)e^{-j\left(\frac{2\pi}{N}\right)nk} \quad (1)$$

Where k is the frequency bin index, l is the time frame index, h is an analysis window of size N (e.g., Hamming window), and M is the framing step (number of samples separating two successive frames). Let $X(k,l)$ and $D(k,l)$ denote the STFT of the clean speech and noise, respectively.
For noise estimation in MS method, first compute the short time subband signal power $\lambda_y(k,l)$ using recursively smoothed periodograms. The update recursion is given by eq.(2). The smoothing constant is typically set to values between $\alpha = 0.9, \ldots, 0.95$.

$$\lambda_y(k,l) = \alpha\lambda_y(k,l-1) + (1-\alpha)|Y(k,l)|^2 \tag{2}$$

The noise power estimate $\lambda_d(k,l)$ is obtained as a weighted minimum of the short time power estimate $\lambda_y(k,l)$ within window of D subband power samples [11], i.e.

$$\bar{\lambda}_d(k,l) = B_{min}\lambda_{min}(k,l) \tag{3}$$

$\lambda_{min}(k,l)$ is the estimated minimum power and B_{min} is a factor to compensate the bias of the minimum estimate. The bias compensation factor depends only on known algorithmic parameters [7].
For reasons of computational complexity and delay the data window of length D is decomposed into U sub-windows of length V such that For a sampling rate of fs=8 kHz and a framing step M=64 typical window parameters are V=25 and U=4,thus D=100 corresponding to a time window of ((D-1).M+N)/f_s=0.824s. Whenever V samples are read, the minimum of the current sub-window is determined and stored for later use. The overall minimum is obtained as the minimum of past samples within the current sub-window and the U previous sub-window minima.
In [7] shown that the bias of the minimum subband power estimate is proportional to the noise power $\sigma^2(k)$ and that the bias can be compensated by multiplying the minimum estimate with the inverse of the mean computed for $\sigma^2(k) = 1$.

$$B_{min} = \frac{1}{E\{\lambda_{min}(k,l)\}|_{\sigma^2(k)=1}} \tag{4}$$

Therefore to obtain B_{min} We must generate data of variance $\sigma^2(k) = 1$, compute the smoothed periodogram (eq. (2)), and evaluate the mean and the variance of the minimum estimate.
As discussed earlier, minimum of the smoothed periodograms, obtained within window of D subband power samples. In next section we propose a method to improve this minimum tracking.

3. PROPOSED METHOD FOR MINIMUM TRACKING

The local minimum in MS method was found by tracking the minimum of noisy speech over a search window spanning D frames. Therefore, the noise update was dependent on the length of the minimum-search window. The update of minimum can take at most 2D frames for increasing noise levels. A different non-linear rule is used in our method for tracking the minimum of the noisy speech by continuously averaging past spectral values [12]

$$\begin{aligned}
&if \quad S_{min}(k,l-1) < S(k,l)\\
&S_{min}(k,l) = \alpha\ S_{min}(k,l-1)\\
&\qquad + \frac{1-\alpha}{1-\beta}(S(k,l) - \beta S(k,l-1))\\
&else\\
&S_{min}(k,l) = \alpha S_{min}(k,l-1)\\
&\qquad - \gamma(S_{min}(k,l-1) - \lambda S(k,l))\\
&end
\end{aligned} \tag{5}$$

where $S_{min}(k,l)$ is the local minimum of the noisy speech power spectrum and α, β, γ and λ are constants which are determined experimentally. The lookahead factor β controls the adaptation time of the local minimum. Typically, we use $\alpha = 0.998$, $\beta = 0.8$, $\gamma = 0.01$ and $\lambda = 0.9$. Because Improve the minimum tracking in this method, the bias compensation factor decreases, as in MS method it is obtained $B_{min} = 1.5$ and in this method it is obtained $B_{min} = 1.2$.

4. PERFORMANCE EVALUATION

The performance evaluation of the proposed method (PM), and a comparison to the MS method, consists of three parts. First, we test the tracking capability of the noise estimators for non-stationary noise. Second, we measure the segmental relative estimation error for various noise types and levels. Third, we integrate the noise estimators into a speech enhancement system, and determine the improvement in the segmental SNR. The results are conformed by a subjective study of speech spectrograms and informal listening tests.
The noise signals used in our evaluation are taken from the Noisex92 database [13]. They include white Gaussian noise (WGN), F16 cockpit noise, and babble noise. The speech signal is sampled at 8 kHz and degraded by the various noise types with segmental SNR's in the range [-5, 10] dB. The segmental SNR is defined by [14]

$$SegSNR = \frac{10}{|\mathcal{L}|}\sum_{l\in\mathcal{L}} log\frac{\sum_k |X(k,l)|^2}{\sum_k |D(k,l)|^2} \quad (6)$$

where $\mathcal{L}$ represents the set of frames that contain speech,
and $|\mathcal{L}|$ its cardinality. The spectral analysis is implemented with Hamming windows of 256 samples length (32ms) and 64 samples frame update step.
Fig. 1 plots the ideal (True), PM, and MS noise estimates for a babble noise at 0 dB segmental SNR, and a single frequency bin k = 5 (the ideal noise estimate is taken as the recursively smoothed periodogram of the noise $|D(k,l)|^2$, with a smoothing parameter set to 0.95). Clearly, the PM noise estimate follows the noise power more closely than the MS noise estimate. The update rate of the MS noise estimate is inherently restricted by the size of the minimum search window (D). By contrast, the PM noise estimate is continuously updated even during speech activity.
Fig. 2 shows another example of the improved tracking capability of the PM estimator. In this case, the speech signal is degraded by babble noise at 5 dB segmental SNR. The ideal, PM, and MS noise estimates, averaged out over the frequency, are depicted in this figure.
A quantitative comparison between the PM and MS estimation methods is obtained by evaluating the segmental relative estimation error in various environmental conditions. The segmental relative estimation error is defined by [15]

$$SegEr = \frac{1}{L}\sum_{l=0}^{L-1}\frac{\sum_k [\hat{\lambda}_d(k,l) - \lambda_d(k,l)]^2}{\sum_k \lambda_d^2(k,l)} \quad (7)$$

where $\lambda_d(k,l)$ is the ideal noise estimate, $\hat{\lambda}_d(k,l)$ is the noise estimated by the tested method, and L is the number of frames in the analyzed signal. Table 1 presents the results of the segmental relative estimation error achieved by the PM and MS estimators for various noise types and levels. It shows that the PM method obtains significantly lower estimation error than the MS method.
The segmental relative estimation error is a measure that weighs all frames in a uniform manner, without a distinction between speech presence and absence. In practice, the estimation error is more consequential in frames that contain speech, particularly weak speech components, than in frames that contain only noise. We therefore examine the performance of our estimation method when integrated into a speech enhancement system. Specifically, the PM and MS noise estimators are combined with the Optimally-Modified Log-Spectral Amplitude (OM-LSA) estimator, and evaluated both objectively using an improvement in segmental SNR measure, and subjectively by informal listening tests. The OM-LSA estimator [16], [17] is a modified version of the conventional LSA estimator [18-19], based on a binary hypothesis model. The modification includes a lower bound for the gain, which is determined by a subjective criterion for the noise naturalness, and exponential weights, which are given by the conditional speech presence probability [20, 21].

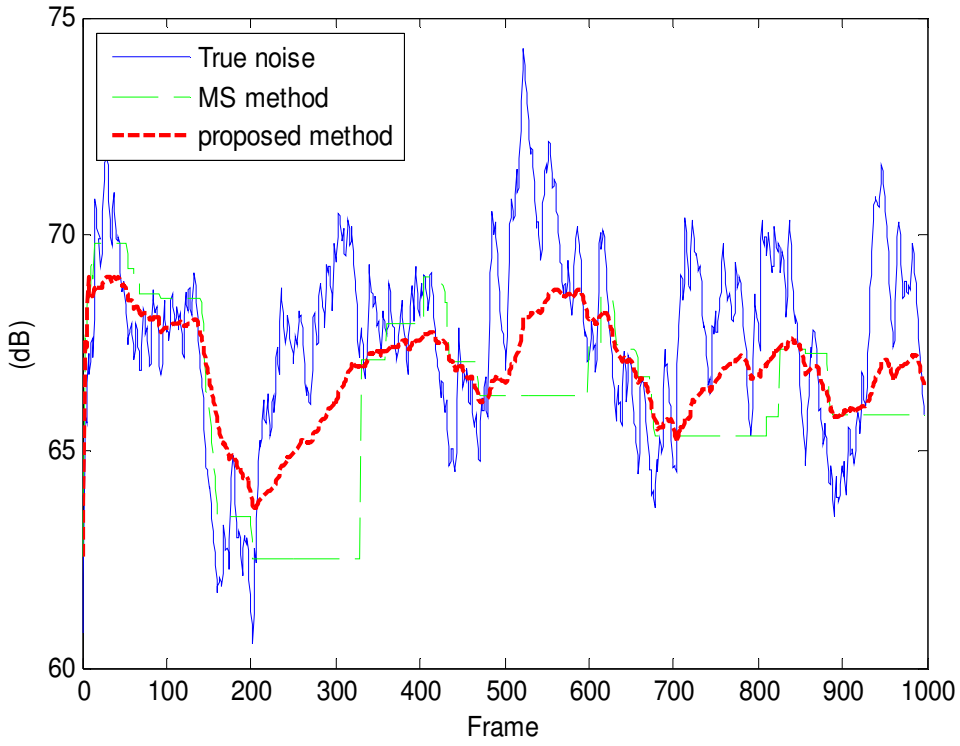


FIGURE 1. Plot of true noise spectrum and estimated noise spectrum using proposed method and MS method for a noisy speech signal degraded by babble noise at 0 dB segmental SNR, and a single frequency bin k = 5.

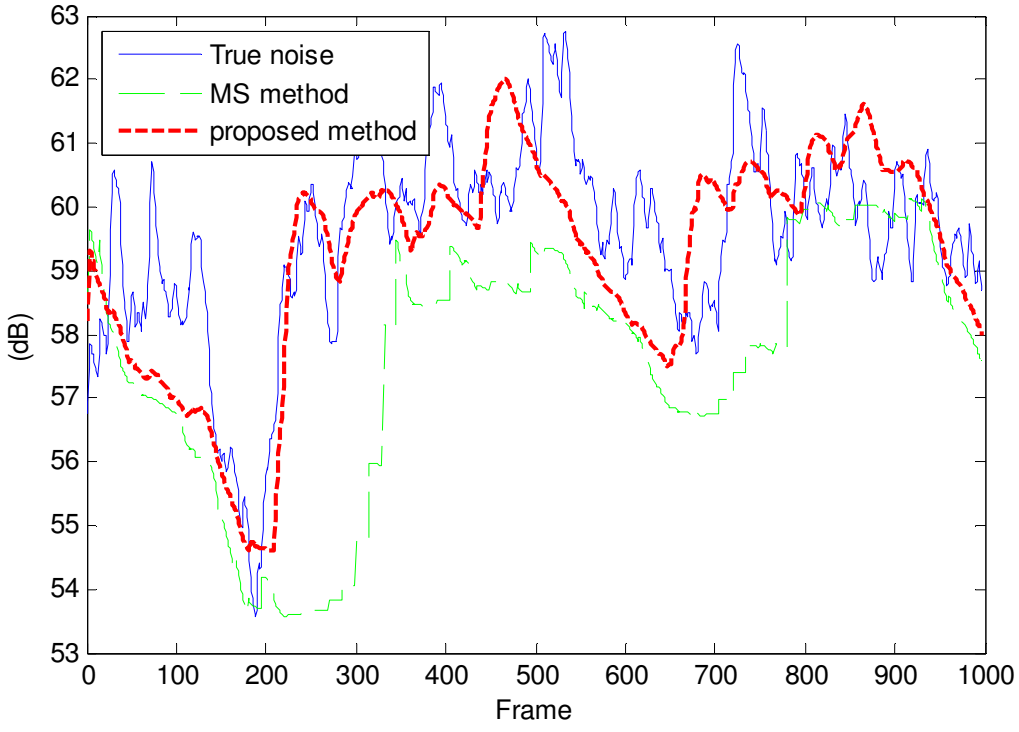


FIGURE 2. Ideal, proposed and MS average noise estimates for babble noise at 5 dB segmental SNR.

Input SegSNR (dB)	WGN Noise		F16 Noise		Babble Noise	
	MS	PM	MS	PM	MS	PM
-5	0.147	0.139	0.192	0.189	0.401	0.397
0	0.170	0.163	0.197	0.193	0.398	0.395
5	0.181	0.173	0.231	0.228	0.427	0.422
10	0.241	0.231	0.519	0.512	0.743	0.736

TABLE 1. Segmental Relative Estimation Error for Various Noise Types and Levels, Obtained Using the MS and proposed method (PM) Estimators.

Input SegSNR (dB)	WGN Noise		F16 Noise		Babble Noise	
	MS	PM	MS	PM	MS	PM
-5	8.213	8.285	6.879	6.924	3.254	3.310
0	7.231	7.312	6.025	6.165	2.581	2.612
5	6.215	6.279	5.214	5.298	2.648	2.697
10	5.114	5.216	3.964	4.034	1.943	1.998

TABLE 2. Segmental SNR Improvement for Various Noise Types and Levels, Obtained Using the MS and proposed method (PM) Estimators.

Table 2 summarizes the results of the segmental SNR improvement for various noise types and levels. The PM estimator consistently yields a higher improvement in the segmental SNR, than the MS estimator, under all tested environmental conditions.

5. SUMMARY AND CONCLUSION

In this paper we have addressed the issue of noise estimation for enhancement of noisy speech. The noise estimate was updated continuously in every frame using minimum of the smoothed noisy speech spectrum. Unlike the MS method, the update of local minimum was continuous over time and did not depend on some fixed window length. Hence the update of noise estimate was faster for very rapidly varying non-stationary noise environments. This was confirmed by formal listening tests that indicated significantly higher preference for our proposed algorithm compared to the MS noise estimation algorithm.

6. REFERENCES

1. J. Meyer, K. U. Simmer and K. D. Kammeyer "*Comparison of one- and two-channel noise-estimation techniques*," Proc. 5th International Workshop on Acoustic Echo and Noise Control, IWAENC-97, London, UK, 11-12 September 1997, pp. 137-145.

2. J. Sohn, N. S Kim and W. Sung, "*A statistical model-based voice activity detector,*" IEEE Signal Processing Letters, 6(1): 1-3, January 1999.

3. B. L. McKinley and G. H. Whipple, "*Model based speech pause detection,*" Proc. 22th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-97, Munich, Germany, 20-24 April 1997, pp. 1179-1182.

4. R. J. McAulay and M. L. Malpass "*Speech enhancement using a soft-decision noise suppression filter,*" IEEE Trans. Acoustics, Speech and Signal Processing, 28(2): 137-145, April 1980.

5. H. G. Hirsch and C. Ehrlicher, *"Noise estimation techniques for robust speech recognition,"* Proc. 20th IEEE Inter. Conf. Acoust. Speech Signal Process., ICASSP-95, Detroit, Michigan, 8-12 May 1995, pp. 153-156.

6. C. Ris and S. Dupont, "*Assessing local noise level estimation methods: application to noise robust ASR,*" Speech Communication, 34(1): 141-158, April 2001.

7. R. Martin, "*Spectral subtraction based on minimum statistics,*" Proc. 7th European Signal Processing Conf., EUSIPCO-94, Edinburgh, Scotland, 13-16 September 1994, pp. 1182-1185.

8. I. Cohen and B. Berdugo, "*Speech Enhancement for Non-Stationary Noise Environments,*" Signal Processing, 81(11): 2403-2418, November 2001.

9. R. Martin, "*Noise power spectral density estimation based on optimal smoothing and minimum statistics,*" IEEE Trans. Speech and Audio Processing, 9(5): 504-512, July 2001.

10. G. Doblinger, "*Computationally efficient speech enhancement by spectral minima tracking in subbands,*" Proc. 4th EUROSPEECH'95, Madrid, Spain, 18-21 September 1995, pp. 1513-1516.

11. R. Martin: "*An Efficient Algorithm to Estimate the instantaneous SNR of Speech Signals*," Proc. EUROSPEECH '93, pp. 1093-1096, Berlin, September 21-23, 1993.

12. Doblinger, G., 1995. "*Computationally efficient speech enhancement by spectral minima tracking in subbands,*" in Proc. Eurospeech' 2002, 1513–1516.

13. A. Varga and H. J. M. Steeneken, "*Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effiect of additive noise on speech recognition systems,*" Speech Communication, 12(3): 247-251, July 1993.

14. S. Quackenbush, T. Barnwell and M. Clements, "*Objective Measures of Speech Quality,"* Englewood Cliffs, NJ: Prentice-Hall, 1988.

15. I. Cohen, "*Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging*," IEEE Trans. Speech Audio Process. 11 (5): 466–475, 2003.

16. I. Cohen, "*On speech enhancement under signal presence uncertainty,*" Proc. 26th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-2001, 7-11 May 2001, pp. 167-170.

17 I. Cohen and B. Berdugo, *"Speech Enhancement for Non-Stationary Noise Environments,"* Signal Processing, 81(11): 2403-2418, November 2001.

18 J. Ghasemi, K. Mollaei, *"A new approach for speech enhancement based on eigenvalue spectral subtraction,"* in Signal Processing: An International Journal (SPIJ), 3(4): 34-41, Sep. 2009.

19 Y. Ephraim and D. Malah, *"Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,"* IEEE Trans. Acoustics, Speech and Signal Processing, 33(2): 443-455, April 1985.

20. M. Satya Sai Ram, P. Siddaiah, M. M. Latha, *" Usefullness of speech coding in voice banking,"* in Signal Processing: An International Journal (SPIJ), 3(4): 42-54, Sep. 2009.

21 M.S. Salam, D. Mohammad, S-H Salleh, *" Segmentation of Malay Syllables in connected digit speech using statistical approach,"* in Signal Processing: An International Journal (SPIJ), 2(1): 23-33, February 2008.

A Novel Algorithm for Acoustic and Visual Classifiers Decision Fusion in Audio-Visual Speech Recognition System

R. Rajavel rettyraja@gmail.com
Research Scholar, ECE Department,
National Institute of Technology Calicut,
Kerala-673601, India

P.S. Sathidevi sathi@nitc.ac.in
Professor, ECE Department,
National Institute of Technology Calicut,
Kerala-673601, India

Abstract

Audio-visual speech recognition (AVSR) using acoustic and visual signals of speech has received attention recently because of its robustness in noisy environments. Perceptual studies also support this approach by emphasizing the importance of visual information for speech recognition in humans. An important issue in decision fusion based AVSR system is the determination of the appropriate integration weight for the speech modalities to integrate and ensure the combined AVSR system's performances better than that of the audio-only and visual-only systems under various noise conditions. To solve this issue, we present a genetic algorithm (GA) based optimization scheme to obtain the appropriate integration weight from the relative reliability of each modality. The performance of the proposed GA optimized reliability-ratio based weight estimation scheme is demonstrated via single speaker, mobile functions isolated word recognition experiments. The results show that the proposed scheme improves robust recognition accuracy over the conventional uni-modal systems and the baseline reliability ratio-based AVSR system under various signals to noise ratio conditions.

Key words: Audio-visual speech recognition, side face visual feature extraction, audio visual decision fusion, Reliability-ratio based weight optimization, late integration

1. INTRODUCTION

Many researchers were trying to design automatic speech recognition (ASR) systems which can understand human speech and respond accordingly [16]. However, the performances of the past and current ASR systems are still far behind as compared to human's cognitive ability in perceiving and understanding speech [18]. The weaknesses of most modern ASR systems are their inability to cope robustly with audio corruption which can arise from various sources, for example environment noises such as engine noise or other people speaking, reverberation effects or transmission channel distortion etc. Thus one of the main challenges being faced by the ASR research community is how to develop ASR systems which are more robust to these kinds of corruptions that are typically encountered in real-world situations. One approach to this problem is to introduce another modality to complement the acoustic speech information which will be invariant to these sources of corruptions [18]. Visual speech is one such source, obviously not perturbed by the acoustic environment and noise. Such systems that combine the audio and visual modalities to identify the utterances are known as audio-visual speech recognition systems [18]. The first AVSR system was reported in 1984 by Petajan [19]. During the last decade more than hundred articles have appeared on AVSR [5, 6, 13, 18]. AVSR systems can enhance the

performance of the conventional ASR not only under noisy conditions but also in clean condition when the talking face is visible [20]. The major advantage of utilizing the acoustic and the visual modalities for speech understanding comes from "Complementarity" of the two modalities: The two pronunciations /b/ and /p/ are easily distinguishable with the acoustic signal, but not with the visual signal; on the other hand, the pronunciations /b/ and /g/ can be easily distinguished visually, but not acoustically [21] and, "synergy" : Performance of audio-visual speech perception can outperform those of acoustic-only and visual-only perception for diverse noise conditions [22]. Generally, the AVSR systems work by the following procedures. First, the acoustic and the visual signals of speech are recorded by a microphone and a camera, respectively. Then, each signal is converted into an appropriate form of compact features. Finally, the two modalities are integrated for recognition of the given speech. The integration can take place either before the two information sources are processed by a recognizer (early integration/feature fusion) or after they are classified independently (late integration/decision fusion). Some studies are in favor of early integration [1, 5, 6, 7, 13, 23], and other prefers late integration [2, 3, 4, 5, 7, 23, 24]. Despite of all these studies, which underline the fact that speech reading is part of speech recognition in humans, still it is not well understood when and how the acoustic and visual information are integrated. This paper takes the advantages of late integration on practical implementation issue to construct a robust AVSR system. The integration weight which determines the amount of contribution from each modality in decision fusion AVSR is calculated from the relative reliability measure of the two modalities [32]. In this work, the integration weight calculated from the reliabilities of each modality is optimized against the recognition accuracy using genetic algorithm. The performance of the proposed GA optimized reliability ratio-based weight estimation scheme is demonstrated via single speaker, mobile functions isolated word recognition experiments. An outline of the remainder of the paper is as follows. The following section explains the integration schemes in AVSR and the reason for decision fusion in this work. Section 3 describes our own recorded experimental database, audio and visual feature extraction schemes. How Genetic Algorithm can be used to obtain the appropriate integration weight from the relative reliability of two modalities for decision fusion is explained in section 4. Section 5 discusses the HMM training and recognition results. The discussion, conclusion and future direction of this work are outlined in the last section.

2. APPROACHES FOR INFORMATION FUSION IN AVSR

The primary focus of AVSR is to obtain the recognition performance which is equal to or better than the performance of any individual modality for various SNR conditions. Secondly, the use of audio-visual information for speech recognition is to improve the recognition performance with as high synergy of the modalities as possible [2]. Generally, while combining two modalities, the integrated system should show high synergy effect for a wide range of SNR conditions. On the contrary, when the fusion is not performed appropriately, we cannot expect complementarity and synergy of the two information sources and moreover, the integrated recognition performance may be even inferior to that of any of the uni-modal systems, which is called "attenuating fusion" [25].

In general, the audio-visual information fusion can be categorized into feature fusion (or early integration) and decision fusion (or late integration), which are shown in figure 1. In feature fusion approach the features of two modalities are concatenated before given to the classifier for recognition, where as in decision fusion approach, the features of each modality are used for recognition separately and, then the outputs of the two classifiers are combined for the final recognition result [2]. Each approach has its own advantages and disadvantages. Most of the audio-visual speech recognition systems [1, 5, 6, 7, 13] are based on feature fusion. The main attraction of this approach is its computational tractability, since only a single classifier is used, and that existing procedures for training and testing of HMMs can be applied without significant modification [4, 26]. There are many advantages in implementing a noise-robust AVSR system using decision fusion. First, in the decision fusion approach it is relatively easy to employ an adaptive weighting scheme for controlling the amounts of the contributions of the two modalities to the final recognition [2, 5]. Second, the decision fusion allows flexible modeling of the temporal coherence of the two information streams, whereas the feature fusion assumes a perfect synchrony between the acoustic and the visual feature sequences [2]. It is proved [27] that there exists an asynchronous characteristic between the acoustic and the visual speech: The lips and the tongue sometimes start to

move up to several hundred milliseconds before the acoustic speech. Finally and most importantly, in the feature fusion approach the combination of the acoustic and the visual features results in high dimensional data sets, which makes training HMMs difficult. Since we have very limited training samples, practical implementation of feature fusion is impossible. Hence this work focuses on the decision fusion for AVSR system

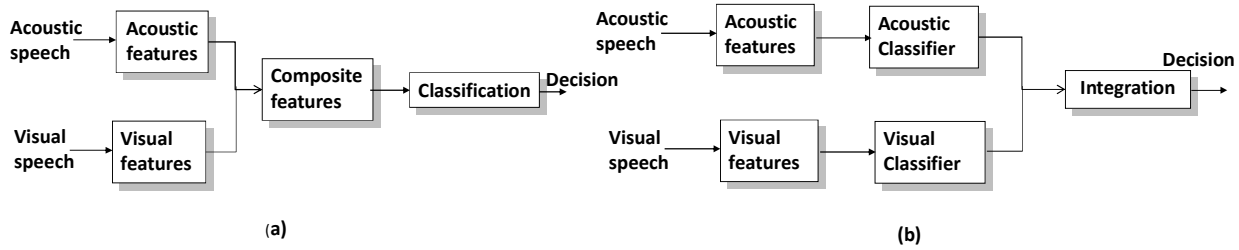


FIGURE 1: AVSR integration schemes. (a) Feature fusion. (b) Decision fusion

3. AUDIO-VISUAL FEATURES EXTRACTION SCHEMES

3.1. Experimental database

This paper focuses on a slightly different type of AVSR system which is mainly useful for mobile applications. Most of the past and current AVSR systems [2, 3, 4, 5, 6, 7] use the mouth information extracted from frontal images of the face, but these systems cannot be used directly for mobile applications because the user needs to hold a handset with a camera in front of their mouth at some distance, which may be unnatural and inconvenient for conversation. As the distance between the mouth and the mobile phone increases, SNR decreases which may worsen the recognition accuracy. If the mouth information can be taken by using a handset held in the usual way for telephone conversation this would greatly improve the usefulness of the system [1]. This paper focuses on this point of view and proposes an audio-visual speech recognition system using side-face images, assuming that a small camera can be installed near the microphone of the mobile device in the future.

Potamianos et al. has demonstrated that using mouth videos captured from cameras attached to wearable headsets produced better results as compared to full face videos [28]. With reference to the above, as well as to make the system more practical in real mobile application, around 70 commonly used mobile functions isolated words were recorded 25 times each by a microphone and web camera located approximately 10-15 cm away from single speaker's right cheek mouth region. Samples of the recorded side-face videos are shown in figure 2. Advantage of this kind of arrangement is that face detection, mouth location estimation and identification of the region of interest etc. are no longer required and thereby reducing the computational complexity [9]. Most of the audiovisual speech databases available are recorded in ideal studio environment with controlled lighting or kept some of the factors like background, illumination, distance between camera and speaker's mouth, view angle of the camera etc. as constant. But in this work, the recording was done purposefully in the office environment on different days with different values for the above factors and also to include natural environment noises such as fan noise, bird's sounds, sometimes other people speaking and shouting sounds.

FIGURE 2: Samples of recorded side-face images

3.2. Acoustic features extraction

Most of the speech recognition systems [1, 2, 3, 4, 5] use the so-called Mel Frequency Cepstral Coefficients (MFCC) and its derivatives as acoustic features for recognition since it shows good recognition performance. This work also adapts, MFCC and its derivatives as acoustic features for recognition. This section briefly reviews the MFCC feature extraction process.

Assume that $s(k)$ represents a speech signal that is multiplied by a hamming window $w(k)$ to obtain a short segment $V_m(k)$ of speech signal defined as:

$$V_m(k) = \begin{cases} s(k).w(k-m) & if \quad k = m,....m+1-N \\ 0 & else \end{cases} \qquad \text{----------} \qquad (1)$$

Where N is the window length and m is the overlapping segment length. [In this work N=256 samples (or 32ms) and m=100 samples (or 12.5ms) with the sampling frequency of fs=8000Hz]. The short speech segment $V_m(k)$ is transformed from time domain to frequency domain by applying an N-point Fast Fourier Transform (FFT). The resulting amplitude spectrum is | V (n) |. For further processing, only power spectrum of the signal is interested, which is computed by taking squares of | V (n) |. Since $V(n)$ is periodic and symmetry, only the values | V (n) |^2 . . . | V (N/2) |^2 are used, giving a total number of $N/2 + 1$ value. Next, the coefficients of the power spectrum | V (n) |^2 are transformed to reflect the frequency resolution of the human ear. A common way to do this is to use K triangle-shaped windows in the spectral domain to build a weighted sum over those power spectrum coefficients | V (n) |^2 which lie within the window. We denote the windowing coefficients as

$$\eta_{kn} \quad ; \quad k = 0,1,....k-1 \quad ; \quad n = 0,1,....N/2 \qquad \text{----------} \qquad (2)$$

In this work, the window coefficients are computed with fs=8000Hz, N=256, and K=22. This gives a new set of coefficients $G(k)$; k = 0, 1, ...K – 1 the so-called mel spectral coefficients

$$G(k) = \sum_{n=0}^{N/2} \eta_{kn} . \left| V(n) \right|^2 \quad ; \quad k = 0,1,....K-1 \qquad \text{----------} \qquad (3)$$

After this, a discrete cosine transform (DCT) is applied to log of mel spectral coefficients. Thus, the Mel frequency cepstral coefficients for frame m can be expressed as

$$c_m(q) = \sum_{k=0}^{K-1} \log(G(k)).\cos\left[\frac{\pi q(2k+1)}{2K}\right] \quad ; \quad q = 0,1,.....Q-1 \qquad \text{----------} \qquad (4)$$

Where $0 \le q \le Q-1$ and Q=12 is the desired number of cepstral features.

The segmented speech signal's energy is also considered as one of the features in this work, which is computed as

$$e = \sum_{n=0}^{N-1} s^2(n) \qquad \text{----------} \qquad (5)$$

In order to better reflect the dynamic changes of the MFCC in time, usually the first and second derivatives in time are also computed, i.e. by computing the difference of two coefficients lying τ times indices in the past and in the future of the time index. The first derivative is computed as:

$$\Delta c_m(q) = c_{m+\tau}(q) - c_{m-\tau}(q) \quad ; \quad q = 0,1,....Q-1 \qquad \text{----------} \qquad (6)$$

The second derivative is computed from the difference of the first derivatives:

$$\Delta\Delta c_m(q) = \Delta c_{m+\tau}(q) - \Delta c_{m-\tau}(q) \quad ; \quad q = 0,1,....Q-1 \qquad \text{----------} \qquad (7)$$

The time interval τ is taken as 4.

3.3. Visual features extraction

Visual features proposed in the literature of AVSR can be categorized into shape-based, pixel-based and motion-based features [29]. Pixel-based and shape based features are extracted from static frames and hence viewed as static features. Motion-based features are features that directly utilize the dynamics of speech [11, 12]. Dynamic features are better in representing distinct facial movements and static features are better in representing oral cavity that cannot be captured either by lip contour or motion-based features. This work focuses on the relative benefits of both static and dynamic features for improved AVSR recognition.

3.3.1. DCT based static feature extraction

G. Potamianos et al. [13] reported that intensity based features using discrete cosine transform (DCT) outperform model-based features. Hence DCT is employed in this work to represent static features. Each side-face mouth region video is recorded with a frame rate of 30 frames/sec and [240 x 320] pixel resolutions. Prior to the image transform the recorded video frames $\{V_t(a, b, c); 1 \le t \le 60; 1 \le a \le 240; 1 \le b \le 320; 1 \le c \le 3\}$ are converted to equivalent RGB image. This RGB image is converted to the YUV color space and only the luminance part (Y) of the image is kept as such since it retains the image data least affected by the video compression [14]. The resultant Y- image was sub sampled to [16 x 16] and then passed as the input $\{A_t(m, n); 1 \le t \le 60; 1 \le m \le 16; 1 \le n \le 16\}$ to the DCT. The images of [16 x 16] pixels provided slightly better performance than [32 x 32] pixel images [14], and hence in this work [16 x 16] pixel images are taken as input to the DCT.

The DCT has the property that, most of the visually significant information about the image is concentrated in just a few coefficients of the DCT. The two dimensional DCT of an m-by-n image sequence $\{A_t(m, n); 1 \le t \le 60; 1 \le m \le 16; 1 \le n \le 16\}$ is defined as:

$$B_t(p,q) = \left\{ \frac{1}{\sqrt{2N}} C(p)\, C(q) \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_t(m,n) \cos\left[\frac{(2m+1)p\pi}{2M} \right] \cos\left[\frac{(2n+1)q\pi}{2N} \right] \right\} \quad \text{---------} \quad (8)$$

Where, $M = N = 16$; $0 \le p \le M-1$; $0 \le q \le N-1$; and

$$C(x) = \begin{cases} 1/\sqrt{2} & if \quad x = 0 \\ 1 & if \quad x > 0 \end{cases}$$

The DCT returns a 2D matrix $B_t(p, q)$ of coefficients and moreover, the triangle region feature selection outperforms the square region feature selection, as those include more of the coefficients corresponding to low frequencies [14]. Hence in this work, [6 x 6] triangle region DCT coefficients without the DC component are considered as 20 static features of a frame.

3.3.2. Motion segmentation based dynamic feature extraction

In this work, dynamic visual speech features which show the side-face mouth region movements of the speaker are segmented from the video using an approach called motion history images (MHI) [11]. MHI is a gray scale image that shows where and when movements of speech articulators occur in the image sequence.

Let $\{A_t(m, n); 1 \le t \le 60; 1 \le m \le 16; 1 \le n \le 16\}$ be a luminance part (Y) image sequence, the difference of frames is defined as

$$DIF_t(m,n) = \left| A_t(m,n) - A_{t-1}(m,n) \right| \quad \text{----------} \quad (9)$$

Where $A_t(m, n)$ is the intensity of each pixel at location (*m, n*) in the t^{th} frame and $DIF_t(m, n)$ is the difference of consecutive frames representing region of motion. Binarization of the difference image $DIF_t(m, n)$ over a threshold τ is

$$DOF_t(m,n) = \begin{cases} 1 & if \quad DIF_t(m,n) \ge \tau \\ 0 & otherwise \end{cases} \quad \text{----------} \quad (10)$$

The value of the threshold τ is optimized through experimentation. Finally MHI (m, n) is defined as

$$MHI(m,n) = Max \bigcup_{t=1}^{N-1} DOF_t(m,n) \times t \quad \text{----------} \quad (11)$$

Where N represents the number of frames used to capture the side-face mouth region motion. In equation (11), to show the recent movements with brighter value, the binarized version of the *DOF* is multiplied with a ramp of time and integrated temporally [11]. Next, DCT was applied to MHI (m, n) and the transformed coefficients are obtained. Similar to static feature extraction, only [6 x 6] triangle region DCT coefficients without DC component are considered as the dynamic features. Finally, the static and dynamic features are concatenated to represent visual speech.

4. DECISION FUSION WITH GA OPTIMIZED RELIABILITY RATIO-BASED INTEGRATION

The main focus of this work is the estimation of optimal integration weight for the modalities in the decision fusion. After the acoustic and visual subsystems perform recognition separately, their outputs are combined by a weighted sum rule to produce the final decision. For a given audio-visual speech test datum of O_A and O_V the recognized utterance C^* is given by [5],

$$C^* = \arg\max_i \left\{ \gamma \log P(O_A / \lambda_A^i) + (1-\gamma) \log P(O_V / \lambda_V^i) \right\} \quad \text{----------} \quad (12)$$

Where λ_A^i and λ_V^i are the acoustic and the visual HMMs for the i^{th} utterance class, respectively, and $\log P(O_A / \lambda_A^i)$ & $\log P(O_V / \lambda_V^i)$ are their outputs. The weighting factor γ ($0 \le \gamma \le 1$) determines how much each modality contributes to the final decision. If it is not estimated appropriately we cannot expect complementarity and synergy of the two information sources and moreover, the combined recognition performance may be even inferior to that of any uni-modal systems [25].

One simple solution to this problem is assigning a constant weight value over various SNR conditions or manual determination of the weight [30]. In some other work, the weight is determined from SNR by assuming that SNR of the acoustic signal is known which is not always a feasible assumption [4]. Indeed, some researchers determine the weight by using an additional adaptation data [31]. Finally, the most popular approach among such schemes is the reliability ratio (RR)-based method in which the integration weight is calculated from the relative reliability measures of the two modalities [32]. This work proposes a Genetic Algorithm based optimization scheme to determine appropriate integration weight from the relative reliability measures of the two modalities, which ensures complementarity and synergy of AVSR without a priori knowledge of the SNR or additional adaptation data. The following subsections briefly explain the baseline reliability ratio - based integration method [32] and the proposed GA optimized reliability ratio - based integration procedure to determine the appropriate integration weight from the reliability measures of acoustic and visual classifiers.

4.1. Baseline reliability ratio - based integration

The reliability of each modality can be measured from the outputs of the corresponding HMMs. When the acoustic speech is not corrupted by any noise, there are large differences between the acoustic HMMs output or else the differences become small. Considering this observation, the reliability of a modality is defined by the most appropriate and best in performance [2]

$$S_m = \frac{1}{N_c - 1} \sum_{i=1}^{N} (\max_j \log P(O / \lambda^j) - \log P(O / \lambda^i)) \quad \text{----------} \quad (13)$$

Which means the average differences between the maximum log-likelihood and the other ones and N_c is the number of classes being considered to measure the reliability of each modality m $\in$ {A, V}. In this case, N_c is 70 i.e. all class recognition hypotheses are considered to measure the reliability. Then, the integration weight γ can be calculated by [32]

$$\gamma = \frac{S_A}{S_A + S_V} \quad \text{----------} \quad (14)$$

Where S_A and S_V are the reliability measure of the outputs of the acoustic and visual subsystems, respectively.

4.2. Proposed GA optimized reliability ratio - based integration

The audio-visual integration method proposed in sections 4.1 can improve the recognition accuracy as compared to the audio-only over certain SNR conditions. This may not be the optimal method of integration weight estimation since that did not show performance improvement at all SNRs. This was experimentally proved in this work for the noisy speech data. To overcome this problem and ensure performance improvement at all SNR conditions, this work proposes a method which optimizes the integration weight estimated in section 4.1 by using genetic algorithm.

The genetic algorithm is a method for solving both constrained and unconstrained optimization problems. It is built on the principles of evaluation via natural selection: an initial population of individual is created and by iterative application of the genetic operators (selection, crossover, mutation) an optimal solution is reached according to the defined fitness function. The procedure of the proposed algorithm is as follows:

Step 1: Initialization: Generate a random initial population of size 20.

Step 2: Fitness evaluation: Fitness of all the solutions in the populations is evaluated. The steps for evaluating the fitness of a solution are given below:

Step 2a: Assume the matrix P of size [$N_c \times N_c$] with all zero values. Where N_c is the Number of utterance class.

Step 2b: class = 1 : No of class (N_c = 70).

Step 2c: test datum = 1 : No of test datum (*Nts* = 5).

Step 2d: Get the acoustic and visual subsystems log likelihood $\log P(O_A / \lambda_A^i)$ and $\log P(O_V / \lambda_V^i)$; respectively, for the class and test datum given in steps 2*b* & 2*c*.

Step 2e: Find the maximum value of acoustic log likelihood
i.e., *amax* = *max* (*sort* (log P (O_A/λ_A^i))*, decend*) for the class and test datum given in steps 2*b* & 2*c*.

Step 2f: Find the maximum value of visual log likelihood
i.e., *vmax* = *max* (*sort* (log *P* (O_V/λ_V^i))*, decend*) for the class and test datum given in steps 2*b* & 2*c*.

Step 2g: Compute the acoustic reliability S_A as:

$$S_A = \frac{1}{N_c - 1} \sum_{i=1}^{N_c} (a\max - \log P(O_A / \lambda_A{}^i))$$

Where N_c is the number of classes being considered.

Step 2h: Compute the visual reliability S_V as:

$$S_V = \frac{1}{N_c - 1} \sum_{i=1}^{N_c} (v\max - \log P(O_V / \lambda_V{}^i))$$

Step 2i: Estimate the integration weight γ as:

$$\gamma = x \times \left[\frac{S_A}{S_A + S_V} \right]$$

According to the solution *x*.

Step 2j: Integrate the log likelihoods as follows

$$C = \arg\max_{i} \left\{ \gamma \log P(O_A / \lambda_A^i) + (1-\gamma) \log P(O_V / \lambda_V^i \right\}$$

Using the estimated integration weight value γ.

Step 2k: Find the maximum value and its corresponding index in *C*.

Step 2l: Increment the value of matrix *P* according to the class and index of *C* as follows

P (class, index) = P (class, index) + 1

Step 2m: Go to step 2*c* until all the test datum are over.

Step 2n: Go to step 2*b* until all the classes are over.

Step 2o: The recognition accuracy or fitness value is defined as

$$\text{Recognition Accuracy} = \frac{\sum diag(P)}{\sum\sum(P)} \times 100$$

Step 3: Updating Population: Two best solutions in the current population are forwarded to the next generation parents without any changes, the remaining solutions in the new population are generated using crossover and mutation.

Step 4: Termination: Repeat steps 2 to 3 until the algorithm reaches the maximum number of iterations.

The final best fitness value gives the maximum recognition accuracy and its corresponding solution gives best integration weight multiplier to obtain the appropriate integration weight for decision fusion.

5. HMM TRAINING AND RECOGNITION RESULTS

The bimodal decision fusion speech recognition system using side-face mouth region image is shown in figure 3. Both speech and side-face mouth region images are simultaneously recorded using low cost microphone and web camera. Audio signals are sampled at 8 kHz with 16-bit resolution. A single frame contains 32 milliseconds speech samples and the frame window proceeds by 12.5 milliseconds. The 12th-order MFCCs, the normalized energy and their delta terms are used for the acoustic features. Further, the cepstral mean subtraction (CMS) technique was applied to remove the channel distortion contained in the speech samples. Visual signals focusing side-face mouth region images are recorded with a frame rate of 30 frames/sec and [240 x 320] pixel resolutions. This work involves decision fusion and hence there is no frame rate synchronization problem between the acoustic and visual speech features. The static visual speech features are computed via DCT image transform approach and dynamic visual speech features are computed via MHI approach. The dynamic features are computed for the whole word not for individual frames. Finally, the static and dynamic features are concatenated to represent visual speech.

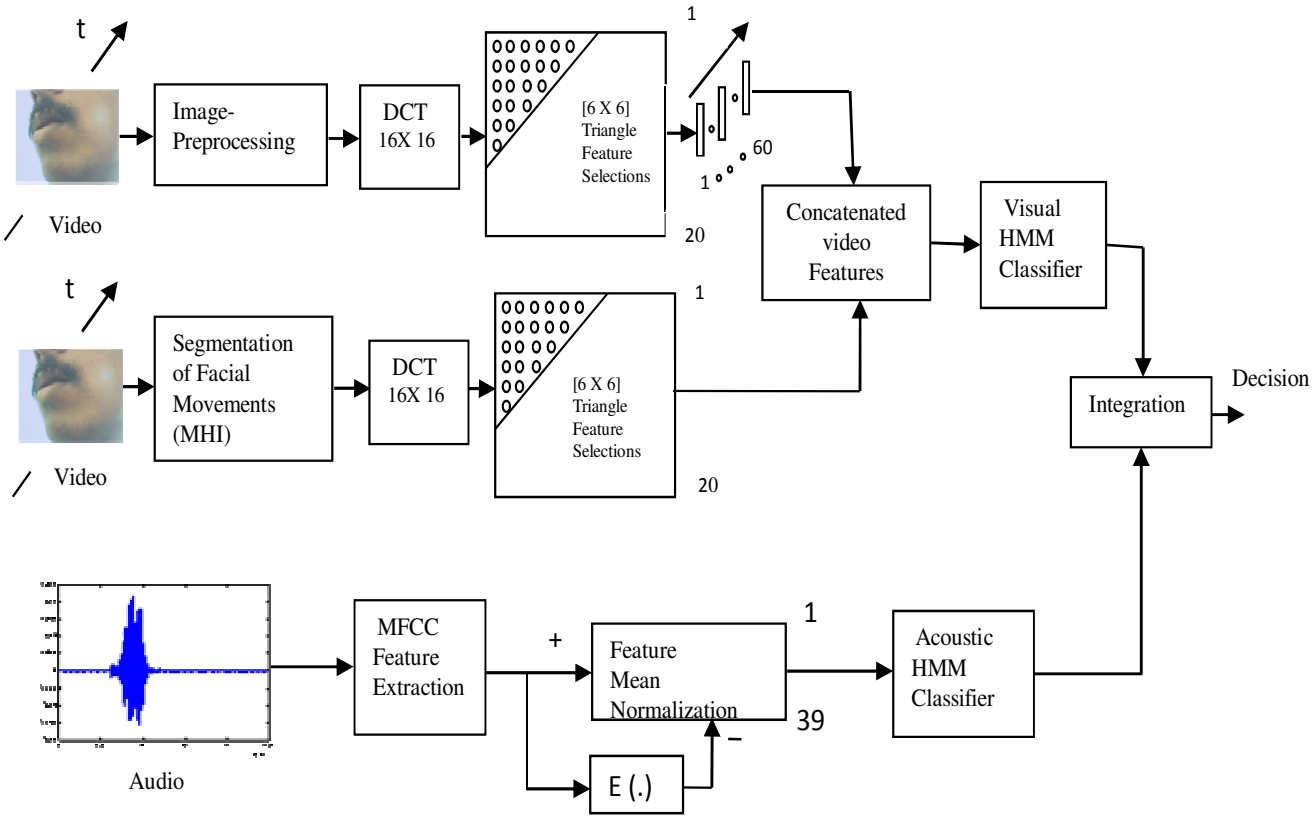


FIGURE 3: Audio-Visual decision fusion speech recognition system using mouth region side-face images

5.1. HMM Recognizer

HMM is a finite state network based on stochastic process. The left-right HMM is a commonly used classifier in speech recognition, since it has the desirable property that it can readily model the time-varying speech signal [16]. This work also adopts left-right continuous HMMs having Gaussian mixture models (GMMs) in each state. The whole- word model which is a standard approach for small vocabulary speech recognition task was used. The number of states in each HMM and number of Gaussian functions in each GMM are set to 10 and 6 respectively, which are determined experimentally. The initial parameters of the HMMs are obtained by uniform segmentation of the training data onto the states of the HMMs and iterative application of the segmental k-means algorithm and the Viterbi alignment. For training the HMMs, the standard Baum-Welch algorithm was used [16]. The training was terminated when the relative change of the log-likelihood value is less than 0.001 or maximum number of iteration is reached, which is set to 25.

SNR	Audio only (%)	Visual only (%)	AV Baseline-RR (%)	AV GA Optimized-RR (%)	Optimum Weight γ
20 dB	94.86	54	98	98.29	0.91
10 dB	50	54	68.57	78	0.25
5 dB	13.71	54	32	66.57	0.07
0 dB	2.86	54	14.57	58	0.04
-5 dB	1.43	54	10	56	0.02
-10 dB	1.43	54	8	54.86	0.01
Average (%)					
(-10dB ~20 dB)	27.38	54	38.52	68.82	
(-10dB ~ 5 dB)	4.86	54	16.14	58.86	

TABLE 1: Audio - only, visual-only, audio-visual speech recognition accuracies

5.2. Results

The proposed GA optimized reliability ratio-based integration algorithm has been tested on single speaker, seventy mobile functions isolated word. The dataset was recorded in an office environment with background noise. Each word was recorded 25 times, 80% of which have been used for training and 20% for testing. The recorded noisy acoustic signal is again artificially degraded with additive white Gaussian noise at SNRs of 20, 10, 5, 0, -5, and -10dB. As mentioned earlier, the main focus of this work is estimating the optimal integration weight for the modalities and in turn maximizing the synergy effect.

Table 1 shows recognition accuracies obtained by the audio-only, visual-only, audio-visual baseline reliability ratio, and the proposed bimodal system at various SNR conditions. Similarly figure 4 compares the recognition performance of all the systems. From the results, the following observations were made,

1. The audio-only recognition system shows nearly 95% for the recorded real time noisy speech at 20dB SNR but, as the speech becomes more noisy, its performance is degraded sharply; the recognition accuracy is even less than 2% at -5 and -10dB SNR conditions.

2. The visual-only system shows 54%, recognition accuracy at all SNRs.

3. The baseline reliability ratio-based method shows synergy effect only at 20 and 10dB SNR conditions but, in the remaining SNR conditions (i.e., -10dB 5dB) their performances are inferior to that of visual-only system i.e. they show attenuation fusion at these SNR conditions.

4. But, the proposed GA optimized reliability ratio-based bimodal system shows synergy effect at all SNR conditions. The amount of synergy at all SNRs is plotted in figure 5. The maximum synergy of 24% occurs at 10dB SNR.

5. Compared to the acoustic-only system, relative recognition performance by the proposed bimodal system is 41.44% on average at all SNR conditions. Under high-noise conditions (i.e., -10dB 5dB), relative recognition performance is 54%.

6. Similarly, compared to the baseline reliability ratio-based system, relative recognition performance by the proposed bimodal system is 30.3% on average at all SNR conditions. Under high-noise conditions (i.e., -10dB 5dB), relative recognition performance is 42.72%, which demonstrates that the noise robustness of recognition is achieved by the proposed system.

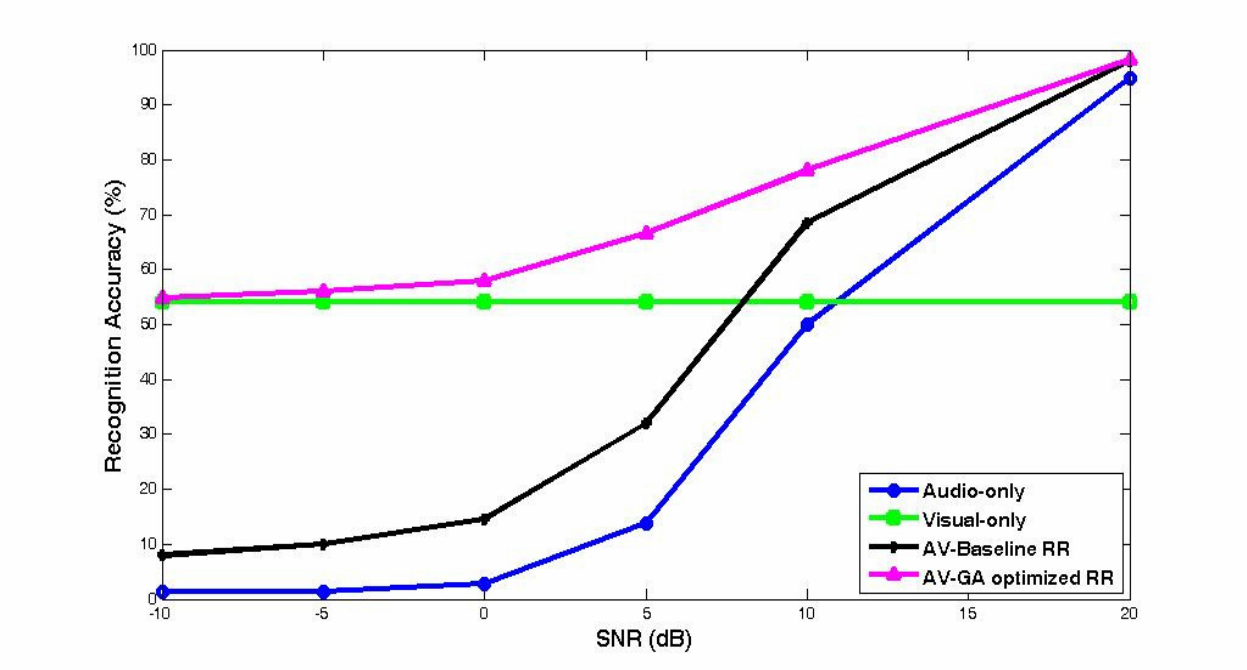


FIGURE 4: Recognition performance of the uni-modal and bimodal systems

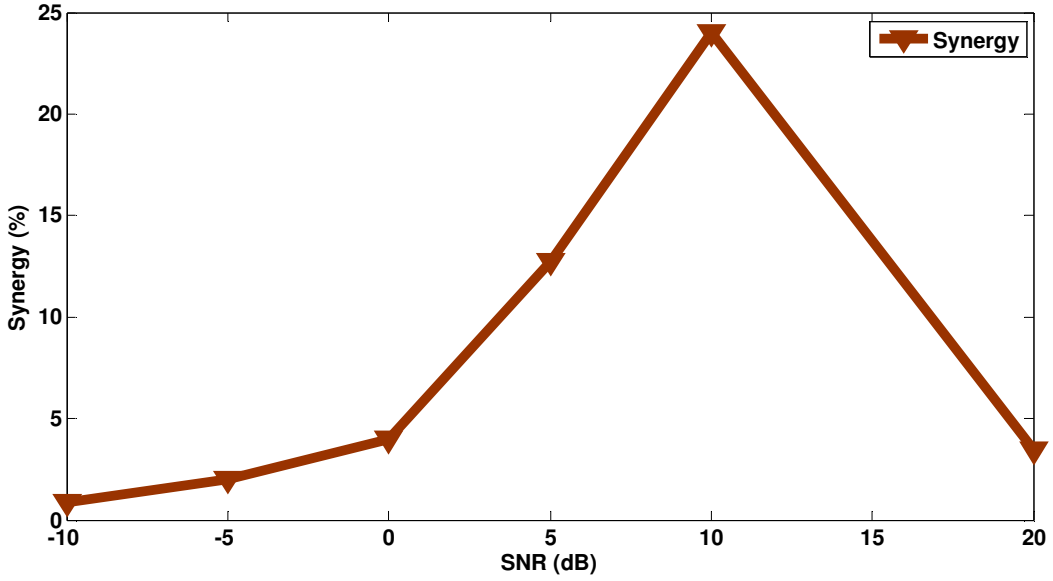


FIGURE 5: Synergy effect of proposed GA optimized RR-based system on various SNR

6. DISCUSSION AND CONCLUSION

In this paper, a GA optimized reliability ratio-based integration weight estimation scheme for decision fusion AVSR system is proposed. The proposed system uses an audio-visual speech data base developed by us, which extracts visual features from the side-face mouth region images rather than frontal face images to focus on mobile applications. Generally, the dynamic visual speech features are obtained by derivatives of static features [14], but in this work the dynamic features are obtained via MHI approach and concatenated with static features to represent the visual speech. For evaluating the proposed method, the recognition accuracy is compared with the related method called baseline reliability ratio-based method in section 5.2. Results show that the proposed method significantly improves the recognition accuracy at all SNR conditions as compared to the baseline reliability ratio-based method. At low SNR, baseline reliability ratio-based method shows very poor recognition accuracy. But the proposed method solves this issue and improves the recognition accuracy considerably. Our future work needs to address the following issues:

1. The baseline reliability ratio-based system show "attenuating fusion" on high-noise conditions (i.e., -10dB 5dB). Therefore an effective denoising algorithm is to be developed to improve the performance further.

2. Moreover, this work was done on a single speaker, small vocabulary mobile function isolated words recognition task. In practice to cover all the recent mobile applications this work needs to be extended to multi speaker, medium size vocabulary, and continuous word recognition task.

7. REFERENCES

1. K. Iwano, T. Yoshinaga, S. Tamura, S. Furui. "*Audio-visual speech recognition using lip information extracted from side-face images*". EURASIP Journal on Audio, Speech, and Music Processing, (2007): 9 pages, Article ID 64506, 2007

2. J.S. Lee, C. H. Park. "*Adaptive Decision Fusion for Audio-Visual Speech Recognition*"'. In: F. Mihelic, J. Zibert (Eds.), Speech Recognition, Technologies and Applications, pp. 550 (2008)

3. J.S. Lee, C. H. Park. "*Robust audio-visual speech recognition based on late integration*"'. IEEE Transaction on Multimedia, 10: 767-779, 2008

4. G. F. Meyer, J. B.Mulligan, S. M.Wuerger. "*Continuous audiovisual digit recognition using N-best decision fusion*". Information Fusion. 5: 91-101, 2004

5. A. Rogozan, P. Delglise. "*Adaptive fusion of acoustic and visual sources for automatic speech recognition*". Speech Communication. 26: 149-161, 1998

6. G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. "*Recent advances in the automatic recognition of audio-visual speech*". In Proceedings of IEEE, 91(9), 2003

7. S. Dupont, J. Luettin. "*Audio-visual speech modeling for continuous speech recognition*". IEEE Transanction on Multimedia, 2: 141-151, 2000

8. G. Potamianos, H. P. Graf, and E. Cosatto. "*An image transform approach for HMM based automatic lipreading*". In Proceedings of International Conference on Image Processing. Chicago, 1998

9. R. Rajavel, P. S. Sathidevi. "*Static and dynamic features for improved HMM based visual speech recognition*". In Proceedings of 1st International Conference on Intelligent Human Computer Interaction, Allahabad, India, 2009

10. G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu. "*A cascade image transform for speaker independent automatic speechreading*". In Proceedings of IEEE International Conference on Multimedia and Expo. New York, 2000

11. W. C. Yau, D. K. Kumar, S. P. Arjunan. "*Voiceless speech recognition using dynamic visual speech features*". In Proceedings of HCSNet Workshop on the Use of Vision in HCI. Canberra, Australia, 2006

12. W. C. Yau, D. K. Kumar, H. Weghorn. "*Visual speech recognition using motion features and Hidden Markov models*". In: M. Kampel, A. Hanbury (Eds.), LNCS, Springer, Heidelberg, pp. 832-839 (2007)

13. G. Potamianos, C. Neti, J. Luettin, and I. Matthews. "*Audio-visual automatic speech recognition: An overview*". In: G. Baily, E. Vatikiotis-Bateson, P. Perrier (Eds.), Issues in visual and audio-visual speech processing, MIT Press, (2004)

14. R. Seymour, D. Stewart, J. Ming. "*Comparison of image transformbased features for visual speech recognition in clean and corrupted videos*". EURASIP Journal on Image and Video Processing. (2008), doi:10.1155/2008/810362, 2008

15. B. Plannerer. "*An introduction to speech recognition: A tutorial* ". Germany, 2003

16. L. Rabiner, B.H. Juang. "*Fundamentals of Speech Recognition*"'. Prentice Hall, Englewood Cliffs (1993)

17. B. Nasersharif, A. Akbari. "*SNR-dependent compression of enhanced Mel sub-band energies for compensation of noise effects on MFCC features*". Pattern Recognition Letters, 28:1320-1326, 2007

18. T. Chen. "*Audiovisual speech processing. Lip reading and lip synchronization*". IEEE Signal Processing Magazine, 18: 9-21, 2001

19. E. D. Petajan. "*Automatic lipreading to enhance speech recognition*". In Proceedings of Global Telecommunications Conference. Atlanta, 1984

20. P. Arnold, F. Hill. "*Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact*". Brit. J. Psychol., 92: 339-355, 2001

21. A. Q. Summerfield. "*Some preliminaries to a comprehensive account of audio-visual speech perception*". In: B. Dodd, R. Campbell (Eds.), Hearing by Eye: The Psychology of Lip-reading. Lawrence Erlbarum, London, pp. 3-51 (1987)

22. C. Benoit, T. Mohamadi, S. D. Kandel. "*Effects of phonetic context on audio-visual intelligibility of French*". Journal of Speech and Hearing Research. 37: 1195-1203, 1994

23. C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. "*Audio visual speech recognition, Final Workshop 2000 Report"*. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 2000

24. P. Teissier, J. Robert-Ribes, J. L. Schwartz. "*Comparing models for audiovisual fusion in a noisy-vowel recognition task*". IEEE Transaction on Speech Audio Processing, 7: 629-642, 1999

25. C. C. Chibelushi, F. Deravi, J. S. D. Mason. "*A review of speech-based bimodal recognition*". IEEE Transactions on Multimedia, 4(1): 23-37, 2002

26. P.L. Silsbee. "*Sensory integration in audiovisual automatic speech recognition*". In Proceedings of the 28th Annual Asilomar Conference on Signals, Systems, and Computers, 1: 561-565, 1994

27. C. Benot. "*The intrinsic bimodality of speech communication and the synthesis of talking faces*". In: M. M. Taylor, F. Nel, D. Bouwhuis (Eds.), The Structure of Multimodal Dialogue II. Amsterdam, Netherlands, pp. 485-502 (2000)

28. G. Potamianos, C. Neti, J. Huang, J.H. Connell, S. Chu, V. Libal, E.Marcheret, N. Hass, J. Jiang. "*Towards practical development of audiovisual speech recognition*". In Proceedings of IEEE International Conf. on Acoustic, Speech, and Signal Processing. Canada, 2004

29. S.W.Foo, L. Dong. "*Recognition of Visual Speech Elements Using Hidden Markov Models*". In: Y. C. Chen, L.W. Chang, C.T. Hsu (Eds.), Advances in Multimedia Information Processing-PCM02, LNCS2532. Springer-Verlag Berlin Heidelberg, pp.607-614 (2002)

30. A. Verma, T. Faruquie, C. Neti, S. Basu. "*Late integration in audiovisual continuous speech recognition*". In Proceedings of Workshop on Automatic Speech Recognition and Understanding. Keystone, 1999

31. S. Tamura, K. Iwano, S. Furui. "*A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization*". In Proceedings of ICASSP. Philadelphia, 2005

32. A. Adjoudani, C. Benot. "*On the integration of auditory and visual parameters in an HMM-based ASR*". In: D. G. Stork and M. E. Hennecke (Eds.), Speech reading by Humans and Machines: Models, Systems, and Speech Recognition, Technologies and Applications, Springer, Berlin, Germany, pp. 461-472 (1996)

A New Enhanced Method of Non Parametric power spectrum Estimation.

K.Suresh Reddy reddysureshk375@rediffmail.com
Associate Professor, ECE Department,
G.Pulla Reddy Engineering College,
Kurnool,518 002 AP, India.

Dr.S.Venkata Chalam sv_chalam2005@yahoo.com
Professor, ECE Department,
Ace college of Engineering,
Hyderabad, 500 003, AP, India..

Dr.B.C.Jinaga bcjinaga@jntu.ac.in
OSD, JNTU,
Hyderabad, AP, India.

Abstract

The spectral analysis of non uniform sampled data sequences using Fourier Periodogram method is the classical approach.In view of data fitting and computational standpoints why the Least squares periodogram (LSP) method is preferable than the “classical” Fourier periodogram and as well as to the frequently-used form of LSP due to Lomb and Scargle is explained. Then a new method of spectral analysis of nonuniform data sequences can be interpreted as an iteratively weighted LSP that makes use of a data-dependent weighting matrix built from the most recent spectral estimate. It is iterative and it makes use of an adaptive (i.e., data-dependent) weighting, we refer to it as the iterative adaptive approach (IAA).LSP and IAA are nonparametric methods that can be used for the spectral analysis of general data sequences with both continuous and discrete spectra. However, they are most suitable for data sequences with discrete spectra (i.e., sinusoidal data), which is the case we emphasize in this paper. Of the existing methods for nonuniform sinusoidal data, Welch, MUSIC and ESPRIT methods appear to be the closest in spirit to the IAA proposed here. Indeed, all these methods make use of the estimated covariance matrix that is computed in the first iteration of IAA from LSP. Comparative study of LSP with MUSIC and ESPRIT methods are discussed.

Keywords: A Nonuniform sampled data, periodogram, least-squares method, iterative adaptive approach, Welch, Music and Esprit spectral analysis.

1. INTRODUCTION

Let the data sequence $\{y(t_n)\}_{n=1}^{N}$ consists of N number of samples whose spectral analysis is our goal. We assume that the observations $\{t_n\}_{n=1}^{N}$ are given, $y(t_n)\varepsilon R(n=1,...N)$ and that a possible

nonzero mean has been removed from $\{y(t_n)\}_{n=1}^N$, so that $\sum_{n=1}^N y(t_n) = 0$. We will also assume throughout this paper that the data sequence consists of a finite number of sinusoidal components and of noise, which is a case of interest in many applications. Note that, while this assumption is not strictly necessary for the nonparametric spectral analysis methods discussed in this paper, these methods perform most satisfactorily when it is satisfied.

2. MOTIVATION FOR THE NEW ESTIMATOR

There are two different non parametric approaches to find the spectral analysis of nonuniform data sequences. First is the classical periodogram approach and the second is Least Squares periodogram approach. The proposed enhanced method of Iterative adaptive approach is explained.

2.1 Classical Periodogram Approach*:* The classical periodogram estimate for the power spectrum of non uniformly sampled data sequence $\{y(t_n)\}_{n=1}^N$ of length N can be interpreted by

$$P_{FP}(\omega) = \frac{1}{N}\left|y(t_n)e^{-jwt_n}\right|^2 \tag{1}$$

Where ω is the frequency variable and where, depending on the application, the normalization factor might be different from 1/N (such as $1/N^2$, see, e.g., [1] and [2]). It can be readily verified that can be obtained from the solution to the following least-squares (LS) data fitting problem:

$$p_F(\omega) = \left|\hat{\beta}(\omega)\right|^2 ,$$

$$\hat{\beta}(\omega) = \min_{\beta(\omega)} \sum_{n=1}^{N} \left| y(t_n - \beta(\omega)e^{-j\omega t_n} \right|^2 \tag{2}$$

In the above (2), if we keep $\beta(\omega) = \left|\beta(\omega)\right|e^{j\phi(\omega)}$, the LS criterion can be written as

$$\sum_{n=1}^{N}[\,y(t_n) - \left|\beta(\omega)\right|\cos(\omega t_n + \phi(\omega))]^2$$

$$+ \left|\beta(\omega)\right|^2 \sum_{n=1}^{N} \sin^2(\omega t_n + \phi(\omega)) \tag{3}$$

Minimization of the first term in (3) makes sense, given the sinusoidal data assumption made previously. However, the same cannot be said about the second term in (3), which has no data fitting interpretation and hence only acts as an additive data independent perturbation on the first term.

2.2 The LS Periodogram: It follows from the discussion in the previous subsection that in the case of real-valued (sinusoidal) data, considered in this paper, the use of Fourier Periodogram is not completely suitable, and that a more satisfactory spectral estimate should be obtained by solving the following LS fitting problem:

$$\min_{\alpha} \sum_{n=1}^{N} [\,y(t_n) - \alpha\cos(\omega t_n + \phi)]^2 \tag{4}$$

The dependence of α and ω can be eliminated using a = α cos (φ) ; b= -α sin(φ) (5)
so that LS criterion can be written as

$$\min_{a,b} \sum_{n=1}^{N} [\,y(t_n) - a\cos(\omega t_n) - b\sin(\omega t_n)]^2 \tag{6}$$

The solution to the minimization problem in (6) is well known to be $\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = R^{-1} r$ (7)

Where $$R = \sum_{n=1}^{N} \begin{bmatrix} \cos(\omega t_n) \\ \sin(\omega t_n) \end{bmatrix} \begin{bmatrix} \cos(\omega t_n) & \sin(\omega t_n) \end{bmatrix} \quad (8)$$

and $$r = \sum_{n=1}^{N} \begin{bmatrix} \cos(\omega t_n) \\ \sin(\omega t_n) \end{bmatrix} y(t_n) \quad (9)$$

The power of the sinusoidal frequency component ω Can be given as

$$\begin{aligned} &\frac{1}{N} \sum_{n=1}^{N} \left(\begin{bmatrix} \hat{a} & \hat{b} \end{bmatrix} \begin{bmatrix} \cos(\omega t_n) \\ \sin(\omega t_n) \end{bmatrix} \right)^2 \\ &= \frac{1}{N} \begin{bmatrix} \hat{a} & \hat{b} \end{bmatrix} R \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} \\ &= \frac{1}{N} r^T R^{-1} r \end{aligned} \quad (10)$$

Hence the periodogram for Least Squares Criterion can be given as

$$p_{LSP}(\omega) = \frac{1}{N} r^T(\omega) R(\omega) r(\omega) \quad (11)$$

The LSP has been discussed, for example, in [3]–[8], under different forms and including various generalized versions. In particular, the papers [6] and [8] introduced a special case of LSP that has received significant attention in the subsequent literature.

2.3 Iterative Adaptive Approach: The algorithm for the proposed estimate is discussed as with the notations. Let $\Delta\omega$ denote the step size of the grid considered for the frequency variable, and let $K = \frac{\omega_{\max}}{\Delta_\omega}$ denote the number of the grid points needed to cover the frequency interval $[0, \omega_{\max}]$, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x ; also, let $\omega_k = k\Delta_\omega$ for k=1,…,K.

$$Y = \begin{bmatrix} y(t_1) \\ y(t_2) \\ . \\ . \\ y(tn) \end{bmatrix}, \ \theta_k = \begin{bmatrix} a(\omega_k) \\ b(\omega_k) \end{bmatrix}, \ A_K = \begin{bmatrix} c_k & s_k \end{bmatrix},$$

$$c_k = \begin{bmatrix} \cos(\omega_k t_1) \\ . \\ \cos(\omega_n t_n) \end{bmatrix}, s_k = \begin{bmatrix} \sin(\omega_k t_1) \\ . \\ \sin(\omega_n t_n) \end{bmatrix} \quad (12)$$

Using this notation we can write the Least squares criterion in (6) as follows in the vector form at, $\omega = \omega_k$

$$\left\| Y - A_K \theta_K \right\|^2 \quad (13)$$

Where $\|\cdot\|$ denotes the Euclidean norm. The LS estimate of θ_k in (7) can be rewritten as

$$\hat{\theta}_k = \left(A_k^T A_k \right)^{-1} A_k^T Y. \quad (14)$$

In addition to the sinusoidal component with frequency ω_k ,the data of Y also consists of other sinusoidal components with frequencies different from ω_k. as well as noise. Regarding the latter, we do not consider a noise component of explicitly, but rather implicitly via its contributions to the data spectrum at $\{\omega_k\}_{k=1}^{K}$; for typical values of the signal-to-noise ratio, these noise contributions to the spectrum are comparatively small. Let us define

$$Q_k = \sum_{P=1, P\neq K}^{K} \left(A_P D_P A_P{}^T \right); \tag{15}$$

$$D_P = \frac{a^2(\omega_p) + a^2(\omega_p)}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{16}$$

which can be thought of as the covariance matrix of the other possible components in Y, besides the sinusoidal component with frequency ω_k considered in (13).

In some applications, the covariance matrix of the noise component of Y is known (or, rather, can be assumed with a good approximation) to be

$$\sum = \begin{bmatrix} \sigma_1^2 & . & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & . & . & . & \sigma_N^2 \end{bmatrix}, \text{ with given } \{\sigma_n^2\}_{n=1}^{N} \tag{17}$$

In such cases, we can simply add $\sum$ to the matrix Q_k in (16).Assuming Q_k that is available, and that it is invertible, it would make sense to consider the following weighted LS (WLS) criterion, instead of (13),

$$[Y - A_k \theta_k]^T Q_k^{-1} [Y - A_k \theta_k] \tag{18}$$

It is well known that the estimate of θ_k obtained by minimizing (18) is more accurate, under quite general conditions, than the LS estimate obtained from (13).Note that a necessary condition for Q_k^{-1} to exist is that (2K-1)>N, which is easily satisfied in general.

The vector that minimizes (18) can be given by

$$\hat{Q}_k = \left(A_k^T Q_k^{-1} A_k \right)^{-1} \left(A_k^T Q_k^{-1} Y \right). \tag{19}$$

Similar to that of (11) the IAA estimate which makes us of Weighted Least Squares an be given by

$$P_{IAA} = \frac{1}{N} \hat{\theta}_k^T (A_k^T A_k) \hat{\theta}_k \tag{20}$$

The IAA estimate in (20) requires the inversion of NXN matrix Q_k for k=1, 2,…, K and also N≥1 which is computationally an intensive task.

To show how we can simply reduce the computational complexity of (19), let us introduce the matrix

$$\psi = \sum_{p=1}^{K} \left(A_p D_p A_P^T \right) = Q_k + A_k D_k A_k^T \tag{21}$$

A simple calculation shows that

$$Q_k^{-1} A_k = \psi^{-1} A_k \left(I + D_k A_k^T Q_k^{-1} A_k \right). \tag{22}$$

To verify this equation, premultiply it with

The ψ in (21) and observe that $\psi Q_k^{-1} A_k = A_k + A_k D_k A_k^T Q_k^{-1} A_k$

$$= A_k \left(I + A_k D_k A_k^T Q_k^{-1} A_k \right) \tag{23}$$

Inserting (22) in (19) yields the another expression for the IAA estimate

$$\hat{\theta}_k = \left(A_k^T \psi^{-1} A_k\right)^{-1} \left(A_k^T \psi^{-1} Y\right) \quad (24)$$

This is more efficient than in (19) computationally.

2.4 Demerits of Fourier Periodogram and LSP:

The spectral estimates obtained with either FP or LSP suffer from both local and global (or distant) leakage problems. Local leakage is due to the width of the main beam of the spectral window, and it is what limits the resolution capability of the periodogram. Global leakage is due to the side lobes of the spectral window, and is what causes spurious peaks to occur (which leads to "false alarms") and small peaks to drown in the leakage from large peaks (which leads to "misses"). Additionally, there is no satisfactory procedure for testing the significance of the periodogram peaks. In the uniformly sampled data case, there is a relatively well-established test for the significance of the most dominant peak of the periodogram; see [1], [2], and [13] and the references therein. In the nonuniform sampled data case, [8] (see also [14] for a more recent account) has proposed a test that mimics the uniform data case test mentioned above. However, it appears that the said test is not readily applicable to the nonuniform data case; see [13] and the references therein. As a matter of fact, even if the test were applicable, it would only be able to decide whether $\{y(t_n)\}$ are white noise samples, and not whether the data sequence contains one or several sinusoidal components (we remark in passing on the fact that, even in the uniform data case, testing the existence of multiple sinusoidal components, i.e., the significance of the second largest peak of the periodogram, and so forth, is rather intricate [1], [2]). The only way of correcting the test, to make it applicable to nonuniform data, appears to be via Monte Carlo simulations, which may be a rather computationally intensive task (see [13]) The main contribution of the present paper is the introduction of a new method for spectral estimation and detection in the nonuniform sampled data case, that does not suffer from the above drawbacks of the periodogram (i.e., poor resolution due to local leakage through the main lobe of the spectral window, significant global leakage through the side lobes, and lack of satisfactory tests for the significance of the dominant peaks). A pre- view of what the paper contains is as follows.

Both LSP and IAA provide nonparametric spectral estimates in the form of an estimated amplitude spectrum (or periodogram $P(\omega)$). We use the frequencies and amplitudes corresponding to the dominant peaks of $P(\omega)$ (first the largest one, then the second largest, and so on) in a Bayesian information criterion see, e.g., [19] and the references therein, to decide which peaks we should retain and which ones we can discard. The combined methods, viz. LSP BIC and IAA BIC, provide parametric spectral estimates in the form of a number of estimated sinusoidal components that are deemed to fit the data well. Therefore, the use of BIC in the outlined manner not only bypasses the need for testing the significance of the periodogram peaks in the manner of [8] (which would be an intractable problem for RIAA, and almost an intractable one for LSP as well—see [13]), but it also provides additional information in the form of an estimated number of sinusoidal components, which no periodogram test of the type discussed in the cited references can really provide.

Finally, we present a method for designing an optimal sampling pattern that minimizes an objective function based on the spectral window. In doing so, we assume that a sufficient number of observations are already available, from which we can get a reasonably accurate spectral estimate. We make use of this spectral estimate to design the sampling times when future measurements should be per- formed. The literature is relatively scarce in papers that ap- proach the sampling pattern design problem (see, e.g., [8] and [20]). One reason for this may be that, as explained later on, spectral window-based criteria are relatively in- sensitive to the sampling pattern, unless prior information (such as a spectral estimate) is assumed to be available—as in this paper. Another reason may be the fact that measure- ment plans might be difficult to realize in some applications, due to factors that are beyond the control of the experimenter. However, this is not a serious problem for the sampling pattern design strategy proposed here which is flexible enough to tackle cases with missed measurements by revising the measurement plan on the fly.

The amplitude and phase estimation (APES) method, proposed in [15] for uniformly sampled data, has significantly less leakage (both local and global) than the periodogram. We follow here the ideas in [16]–[18] to extend APES to the nonuniformly sampled data case. The so-obtained generalized method is referred to as RIAA for reasons explained in the Abstract.

2.5 The Iterative Adaptive Algorithm: The proposed algorithm for power spectrum estimation can be explained as follows

- Initialization: Using the Least Squares method in (13) obtain the initial estimates of $\{\theta_k\}$ which are denoted by $\left\{ \hat{\theta}_k^{\,0} \right\}$.
- Iteration: Let $\left\{ \hat{\theta}_k^{\,i} \right\}$ denote the estimates of $\{\theta_k\}$ at the i[th] iteration (i=0, 1, 2…), and let $\left\{ \hat{\psi}^{\,i} \right\}$ denote the estimate of ψ obtained from $\left\{ \hat{\theta}_k^{\,i} \right\}$.
- For i=0, 1, 2.., Compute: $\hat{\theta}_k^{\,i+1} = \left[A_k^T (\hat{\psi}^{\,i})^{-1} A_k \right]^{-1} \left[A_k^T (\hat{\psi}^{\,i})^{-1} y \right]$ for k=1,…K.

  Until a given number of iterations are performed.
- Periodogram calculations:

Let $\left\{ \hat{\theta}_k^{\,I} \right\}$ denotes the estimation of $\{\theta_k\}$ Obtained by the iterative process (I denote iteration number at which iteration is stopped). Using $\left\{ \hat{\theta}_k^{\,I} \right\}$ compute the IAA periodogram as

$$P_{IAA}(\omega_k) = \frac{1}{N} \left(\hat{\theta}_k^{\,I} \right)^T \left(A_k^T A_k \right) \left(\hat{\theta}_k^{\,I} \right), \quad \text{for k=1,… K.}$$

3. PROPOSED SYSTEM AND SIMULATED DATA:

The system model for the proposed algorithm is shown in Figure 1.

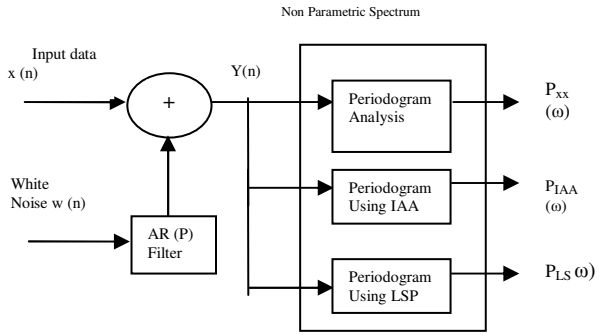


FIGURE 1: Proposed system model for the simulated data.

The system model for the proposed algorithm is shown in Figure 1. We consider a data sequence consisting of M=3 sinusoidal components with frequencies 0.1, 0.4 and 0.41 Hz, and amplitudes 2,4 and 5, respectively. The phases of the three sinusoids are independently and uniformly distributed over $[0,2\pi]$ and the additive noise is white normally distributed with mean of 0 and variance of σ^2 =0.01. We define the signal-to-noise ratio (SNR) of each sinusoid as

$$SNR_m = 10\log_{10}\left(\frac{\alpha_m^2/2}{\sigma^2}\right) \text{ dB} \quad \text{m=1,2,3.} \qquad (25)$$

Where α_m is the amplitude of the m[th] sinusoidal component hence SNR1=23 dB, SNR2=29 dB and SNR3= 31 dB in this simulation example. The input data sequence for the system model is as follows

$$x(t) = 2\cos(2\pi 0.1t) + 3\cos(2\pi 0.4t) + 4\cos(2\pi 0.4t) + w(t) \qquad (26)$$

Where $w(t)$ zero mean Gaussian is distributed white noise with variance of 0.01 and the sampling pattern follows a Poisson process with parameter $\lambda = 0.1s^{-1}$, that is, the sampling intervals are exponentially distributed with mean $\mu = \frac{1}{\lambda} = 10$ s. We choose N=64 and show the sampling pattern in Fig. 3(a). Note the highly irregular sampling intervals, which range from 0.2 to 51.2 s with mean value 9.3 s. Fig. 3(b) shows the spectral window corresponding to Fig. 3(a). The smallest frequency at which the spectral $f_0 \rangle 0$ at which the spectral window has a peak close to N^2 is approximately 10 Hz. Hence $f_{\max}=f_0/2$=5Hz. The step Δf of the frequency grid is chosen as 0.005 Hz. However, they are most suitable for data sequences with discrete spectra (i.e., sinusoidal data), which is the case we emphasize in this paper. Of the existing methods for nonuniform sinusoidal data, Welch, MUSIC and ESPRIT methods appear to be the closest in spirit to the IAA.

4. RESULT ANALYSIS:

The results in Fig. 2 presents the spectral estimates averaged over 100 independent realizations of Monte-Carlo trials of periodogram and Welch estimates. Fig. 4 presents the spectral estimates averaged over 100 independent realizations of LSP and IAA estimates. Fig. 5 presents the spectral estimates averaged over 100 independent realizations of Monte- Carlo trials of Music and Esprit estimates. LSP nearly misses the smallest sinusoid while IAA successfully resolves all three sinusoids. Note that IAA suffers from much less variability than LSP from one trial to another. The plots were taken with the help MATLAB programming by the authors. LSP and IAA are nonparametric methods that can be used for the spectral analysis of general data sequences with both continuous and discrete spectra. However, they are most suitable for data sequences with discrete spectra (i.e., sinusoidal data), which is the case we emphasize in this paper. Of the existing methods for nonuniform sinusoidal data, Welch, MUSIC and ESPRIT methods appear to be the closest in spirit to the IAA proposed here. Indeed, all these methods make use of the estimated covariance matrix that is computed in the first iteration of IAA from LSP. MUSIC and ESPRIT, on the other hand, are parametric methods that require a guess of the number of sinusoidal components present in the data, otherwise they cannot be used furthermore.

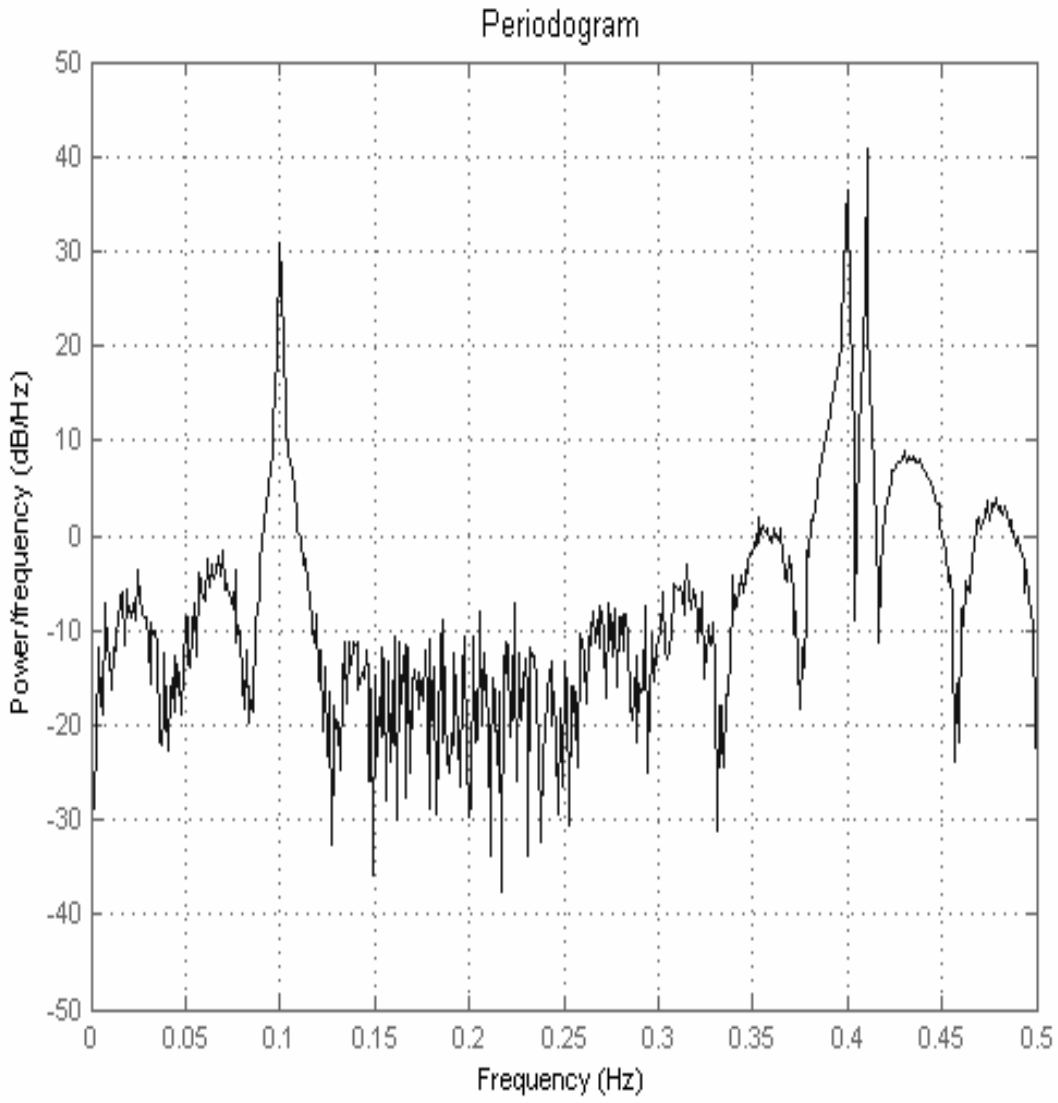


FIGURE 2: Average spectral estimates from 100 Monte Carlo trials of Fourier periodogram

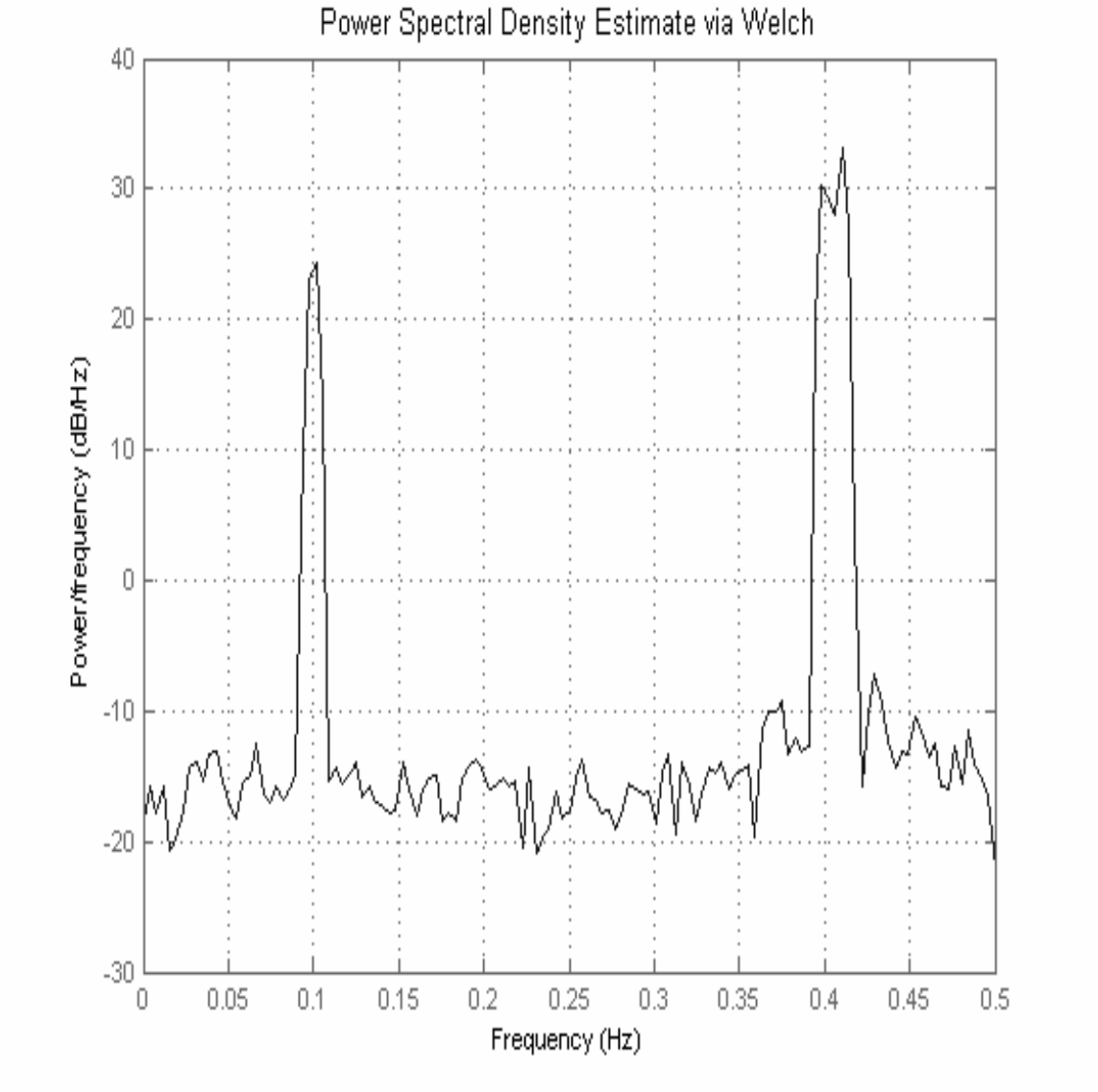


FIGURE 3: Average spectral estimates from 100 Monte Carlo trials of Welch estimates.

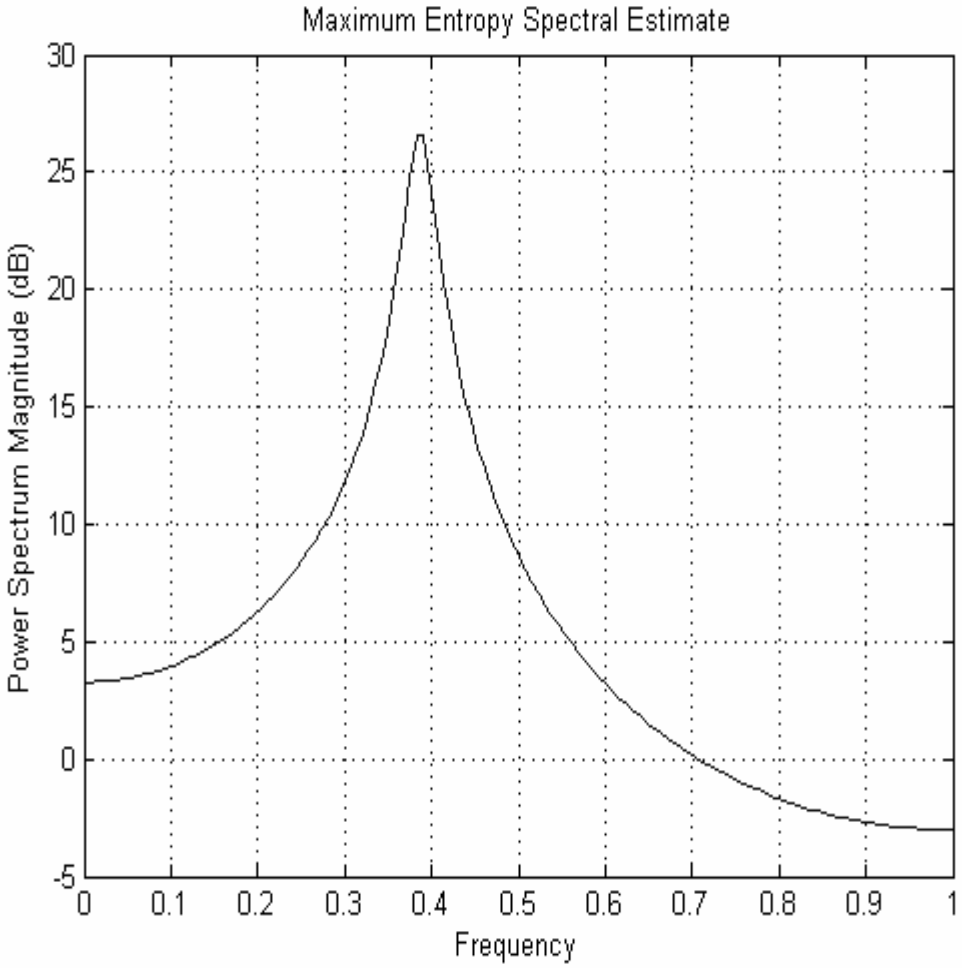


FIGURE 4: Average spectral estimates from 100 Monte Carlo trials of MEM estimates.

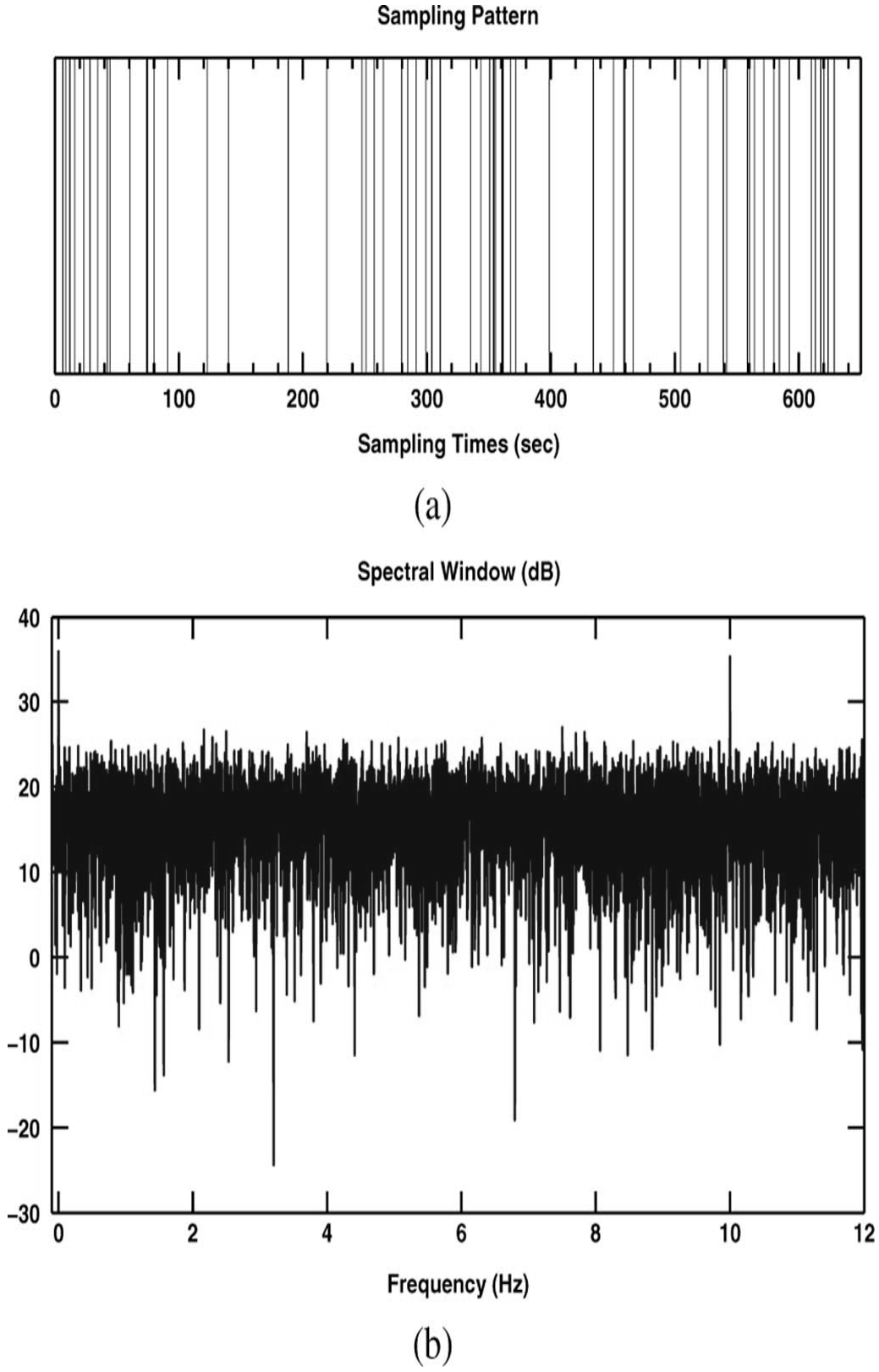


FIGURE6: Sampling pattern and spectral window for the simulated data case. (a) The sampling pattern used for all Monte Carlo trials in Figs. 2–4. The distance between two consecutive bars represents the sampling interval. (b) The corresponding spectral window

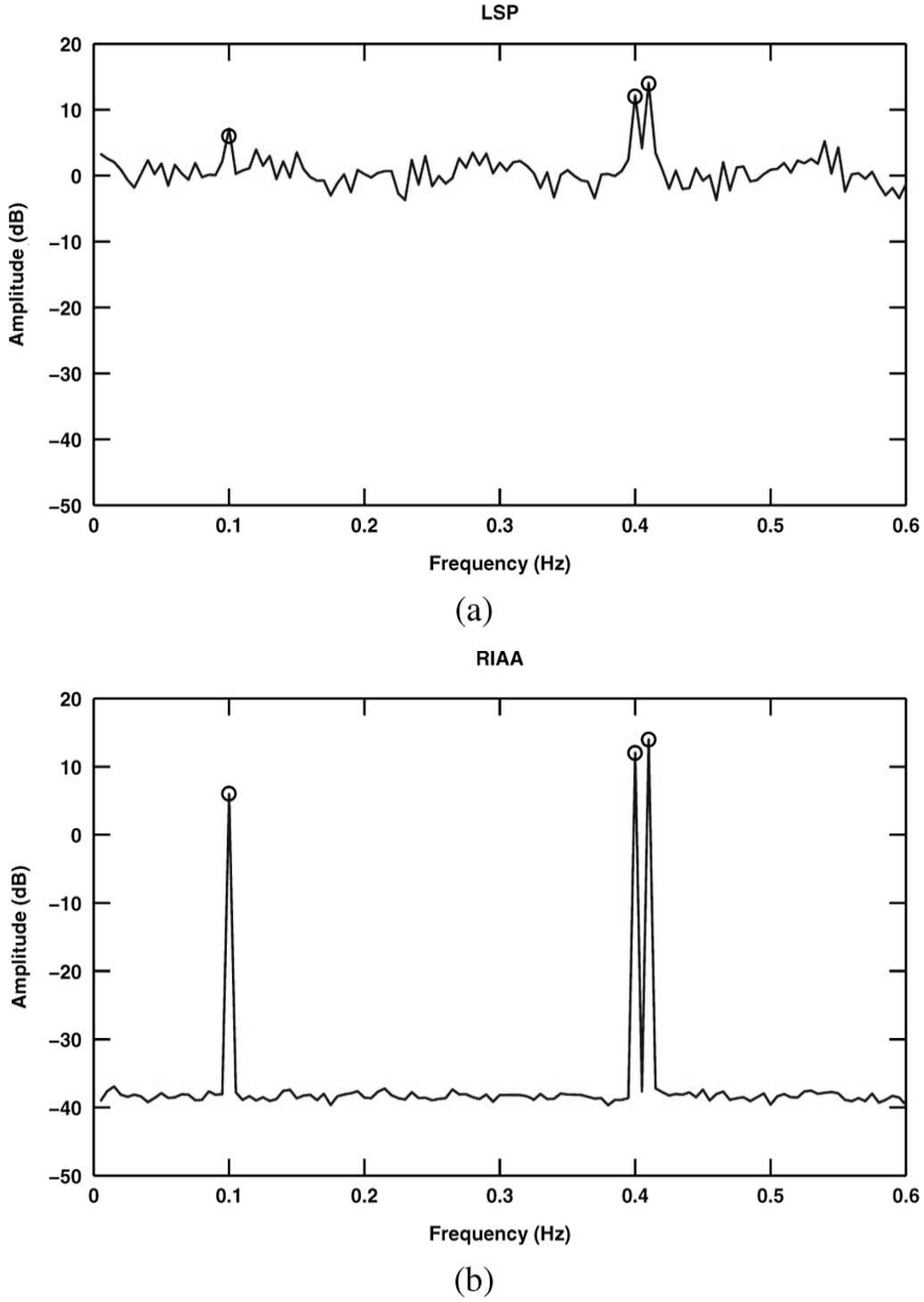
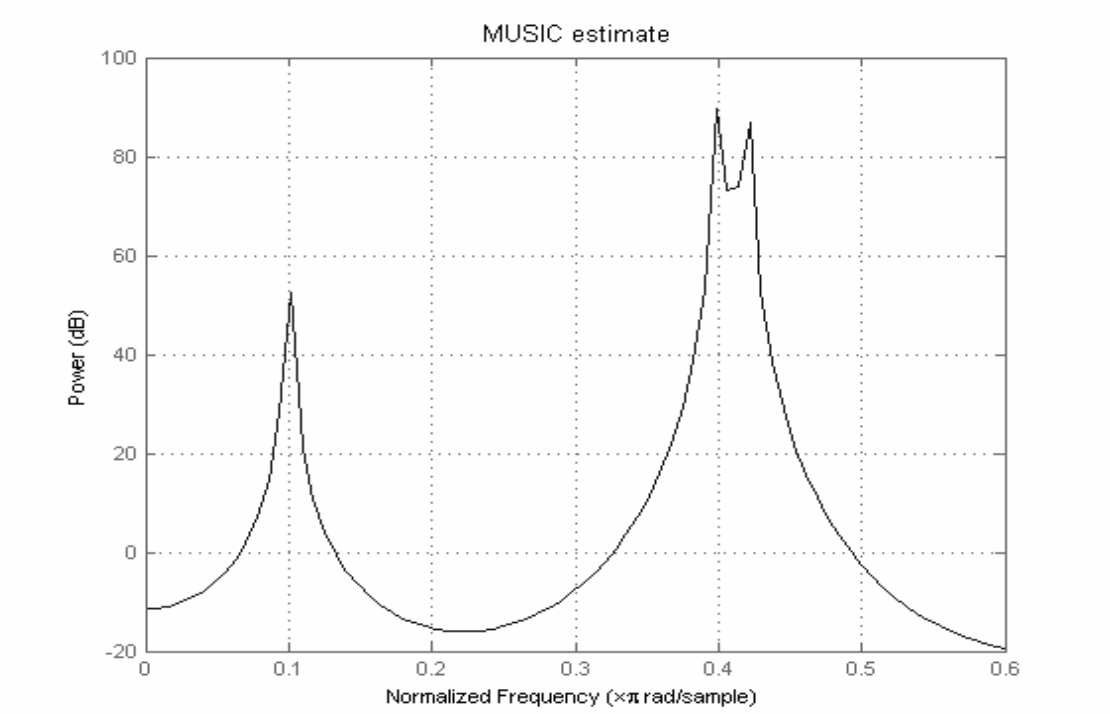
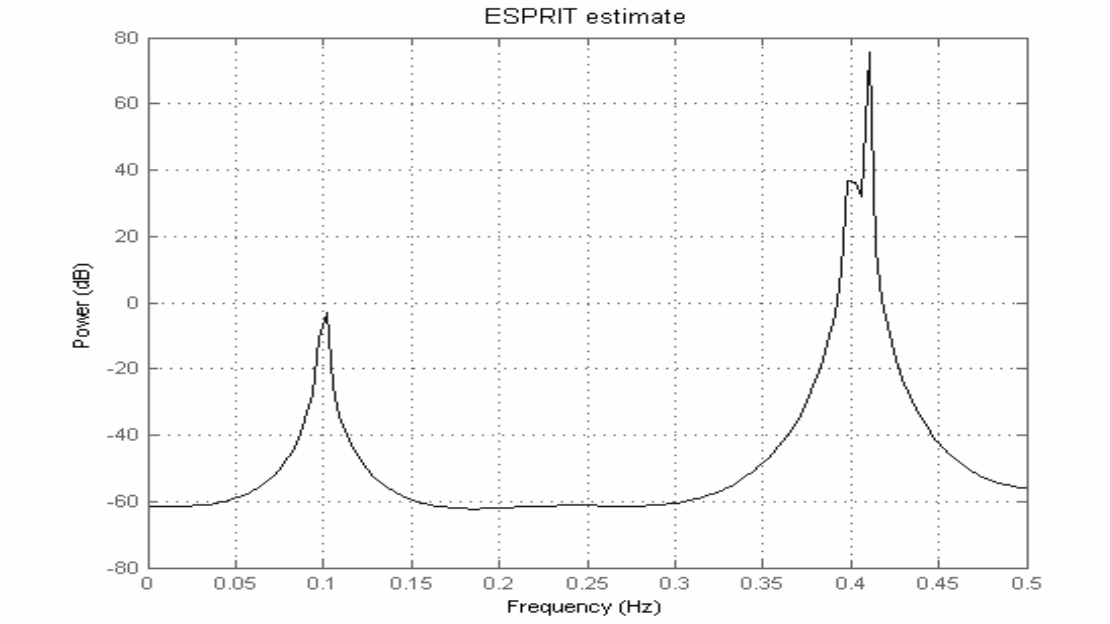


FIGURE7: Average spectral estimates from 100 Monte Carlo trials. The solid line is the estimated spectrum and the circles represent the true frequencies and amplitudes of the three sinusoids. (a) LSP (b) IAA.





(a)

(b)

FIGURE8: Average spectral estimates from 100 Monte Carlo trials. (a) Music estimate and (b) Esprit estimate.

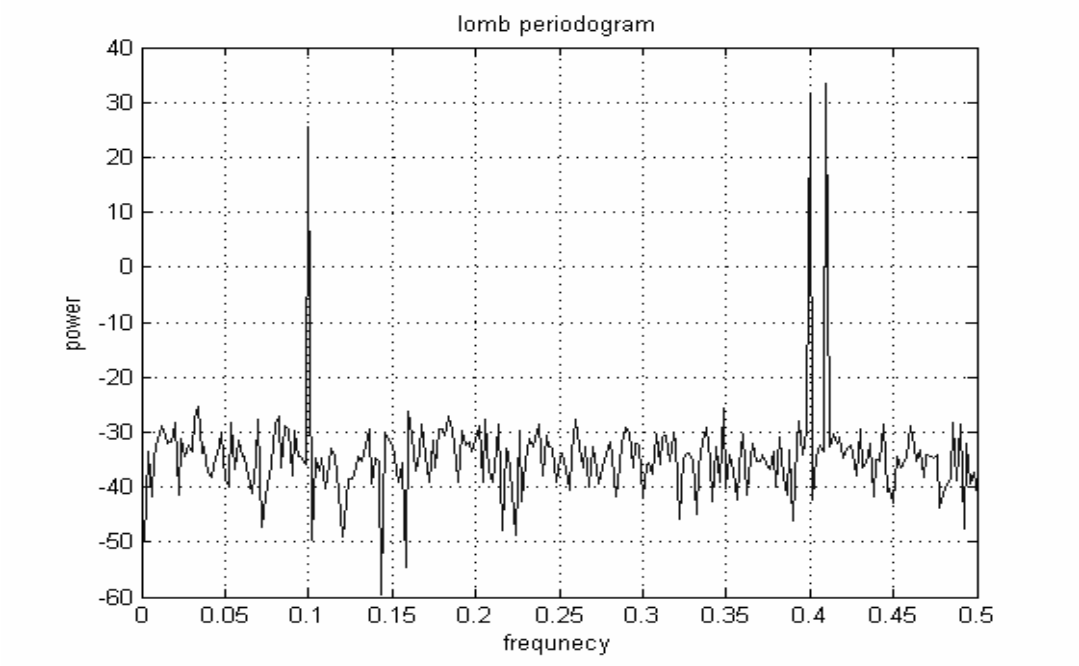


FIGURE9: Average spectral estimates from 100 Monte Carlo trials of Lomb periodogram.

. 4. CONSLUSIONS:

Of the existing methods for nonuniform sinusoidal data, the MUSIC and ESPRIT methods appear to be the closest in spirit to the IAA proposed here (see the cited paper for explanations of the acronyms used to designate these methods). Indeed, all these methods make use of the estimated covariance matrix that is computed in the first iteration IAA from LSP. In fact Welch (when used with the same covariance matrix dimension as IAA) is essentially identical to the first iteration of IAA. MUSIC and ESPRIT.In the case of a single sinusoidal signal in white Gaussian noise, the LSP is equivalent to the method of maximum likelihood and therefore it is asymptotically statistically efficient. Consequently, in this case LSP can be expected to outperform IAA. In numerical computations we have observed that LSP tends to be somewhat better than IAA for relatively large values of N or SNR; however, we have also observed that, even under these conditions that are ideal for LSP, the performance of IAA in terms of MSE (mean squared error) is slightly better (by a fraction of a dB) than that of LSP when or SNR becomes smaller than a certain threshold.

5. REFERENCES

[1] Stoica,P and R.L. Moses, Introduction to Spectral Analysis, Prentice-Hall, 1997, pp. 24-26.

[2] Welch, P.D, "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodogram," IEEE Trans. Audio electro acoustics, Vol. AU-15 (June 1967), pp. 70-73.

[3] Oppenheim, A.V., and R.W. Schafer, Discrete-Time Signal Processing, Prentice-Hall, 1989, pp. 730-
742.

[4] B. Priestley, *Spectral Analysis and Time Series, Volume 1: Univariate Series,* New York: Academic, 1981.

[5] P. Stoica and R. L. Moses, *Spectral Analysis of Signals* Upper Saddle River, NJ: Prentice-Hall, 2005.

[6] F.J.M.Barning,"The numerical analysis of the light-curve of12 lacerate," *Bull.*

Astronomy. Inst Netherlands, vol. 17, no. 1, pp. 22–28, Aug. 1963.

[7] P. Vanicek, "Approximate spectral analysis by least-squares fit," *Astro- phys. Space Sci.*, vol. 4, no. 4, pp. 387–391, Aug. 1969.

[8] P. Vanicek, "Further development and properties of the spectral anal- ysis by least- squares," *Astrophys. Space Sci.*vol. 12, no. 1, pp. 10–33, Jul. 1971.

[9]N. R. Lomb, "Least-squares frequency analysis of unequally spaced data," *Astrophysics . Space Sci.*, vol. 39, no. 2, pp. 447–462, Feb. 1976.

[10] S. Ferraz-Mello, "Estimation of periods from unequally spaced obser- vations," *Astronom. J.*, vol.86, no.4, pp. 619–624, Apr. 1981.

[11] J. D. Scargle, "Studies in astronomical time series analysis. II—Statis- tical aspects of spectral analysis of unevenly spaced data," *Astrophys. J.*, vol. 263, pp 835–853, Dec. 1982.

[12] W. H. Press and G. B. Rybicki, "Fast algorithm for spectral analysis of unevenly sampled data," *Astrophys. J.*, vol. 338, pp. 277–280, Mar.1989.

[13] J. A. Fessler and B. P. Sutton, "Nonuniform fast Fourier transforms using min-max interpolation," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 560–574, Feb. 2003.

[14] N. Nguyen and Q. Liu, "The regular Fourier matrices and nonuniform fast Fourier transforms," *SIAM J. Sci. Comput.*, vol. 21, no. 1,pp. 283–293, 2000.

[15] L. Eyer and P. Bartholdi, "Variable stars: Which Nyquist frequency?,"*Astron. Astrophys. Supp Series*, vol. 135, pp. 1–3, Feb. 1999.

[16] F. A. M. Frescura, C. A. Engelbrecht, and B. S. Frank, "Significance tests for periodogram peaks," NASA Astrophysics Data System, Jun. 2007

[17] P. Reegen, "SigSpec—I. Frequency- and phase resolved significance in Fourier space," *Astron Astrophys.*, vol. 467, pp. 1353–1371, Mar.2007.

[18] J. Li and P. Stoica, "An adaptive filtering approach to spectral estimation and SAR imaging,"T*rans.*, vol. 44, no. 6,pp. 1469–1484, Jun. 1996.

[19] T. Yardibi, M. Xue, J. Li, P. Stoica, and A. B.Baggeroer, "Iterative adaptive approach for sparse signal representation with sensing applications," *IEEE Trans. Aerosp. Electron . Syst.*, 2007.

[20] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp.36–47, Jul. 2004.

[21] E. S. Saunders, T. Naylor, and A. Allan, "Optimal placement of a limited number of observations for period searches," *Astron. Astrophys.*,vol. 455, pp. 757–763, May 2006.

A Combined Voice Activity Detector Based On Singular Value Decomposition and Fourier Transform

Amard Afzalian a.afzalian@gmail.com
Islamic Azad University, Science and Research,Tehran, IRAN.

Mohammad Reze Karami Mollaei mkarami@nit.ac.ir
Faculty of Electrical and Computer Engineering, Babol Noshirvani University of Technology Babol, P.O. Box 47135-484, IRAN

Jamal Ghasemi jghasemi@stu.nit.ac.ir
Faculty of Electrical and Computer Engineering, Signal Processing Laboratory Babol Noshirvani University of Technology, Babol, P.O. Box 47135-484, IRAN

Abstract

Voice activity detector (VAD) is used to separate the speech data included parts from silence parts of the signal. In this paper a new VAD algorithm is represented on the basis of singular value decomposition. There are two sections to perform the feature vector extraction. In first section voiced frames are separated from unvoiced and silence frames. In second section unvoiced frames are silence frames. To perform the above sections, first, windowing the noisy signal then Hankel's matrix is formed for each frame. The basis of statistical feature extraction of purposed system is slope of singular value curve related to each frame by using linear regression. It is shown that the slope of singular values curve per different SNRs in voiced frames is more than the other types and this property can be to achieve the goal the first part can be used. High similarity between feature vector of unvoiced and silence frame caused to approach for separation of the two categories above cannot be used. So in the second part, the frequency characteristics for identification of unvoiced frames from silent frames have been used. Simulation results show that high speed and accuracy are the advantages of the proposed system.

Keywords: Speech, Voice Activity Detector, Singular Value.

1. INTRODUCTION

Voice activity detection is an important step in some speech processing systems, such as speech recognition, speech enhancement, noise estimation, speech compression ... etc. In speech recognition when a word or utterance begins or ends (the end points) must be specified [1]. Also VAD is used to disable speech recognition for silence segments. Some speech transition systems transmits active segments in high rate of bits and transmits silence in low rate of bits, by this method they improve the band-width [2]. In some speech enhancement algorithm for example

spectral subtraction method, a VAD method is required to know when to update the noise reference [3,20,21]. Conversational speech is a sequence of consecutive segments of silence and speech. In noisy signal silence regions are noisy. Voice sound contain more energy than unvoiced sound, while unvoiced sounds are more noise-like, so in noisy condition activity detection is harder In unvoiced regions. Feature extraction is the most important section in VAD system that elicits required parameters from desired frame. To achieve an accurate algorithm, the system parameters must be selected until by them can be able to separate from each other areas. After proper feature election, threshold is applied to the extracted parameters and the decisions are made. To achieving good detection level threshold can be adapted to the change of the noise conditions. Many of the algorithms assume that the first frames are silence [5, 6, and 7], so we can initialize noise reference from these frames. Common features used in VAD's are short term Fourier transform and zero crossing rate [4, 6, 11, and 12]. Another important and widely used parameter in this regard is Cepstral Coefficient [7, 9]. In this method the Cepstral coefficients are calculated within frames and then by calculating the difference between this vector and the value assigned to the noise signal and then comparing the result with the basic threshold value, the frame identity could be determined. LPC method is also another major applicable method for VAD implementation [13]. Generally in LPC based algorithms a series of mean coefficients are experimentally considered for voice, unvoiced and silent modes. In the next step the LPC coefficients of suspicious frame and their relative difference with mean indices are calculated and the frame identity is recognized based on these values. The other parameters for implementing VAD in combined algorithms are LTSE (long term Spectral Estimation)[5], wavelet coefficient [8,10], the ratio of signal to noise in sub-band [14], LSPE(Least Square Periodicity Estimator)[11] and AMDF(Average Magnitude Difference Function)[15]. One of most important cases in VAD system is speed of system performance beside proper accuracy. In this paper is to present a new algorithm of VAD based on single value decomposition and frequency features, specifications accuracy and speed simultaneously be fulfilled. Based on this, paper organization as follows that in Section 2 single values decomposition (SVD) will be explained. In Section 3 the proposed method with the system block diagram is given. In Section 4 simulation results in terms of quantitative and qualitative criteria is evaluated. Finally, the article concludes with Section 5 ends.

2. Singular Value Decomposition

The singular value composition is one of the main tools in digital signal processing and statistical data. By doing SVD on a matrix with dimensions of M×N, we have:

$$X = U \Sigma V^T \tag{1}$$

On above relation U and V matrixes are singular vectors matrix with dimensions of M×M and N×N, respectively. Also, $\sum$ with r order of a diagonal matrix M×N is included singular values so that components on the main dial gauge are not zero and other components are zero. The elements on main dial gauge areas $\sigma_{11} > \sigma_{22} > ... > \sigma_{rr} > 0$ And are the values of X matrix. For exercising SVD to one dimensional matrix, the vector of signal samples must map to subspace with more dimensions, on the other hand must be changed to a matrix in certain way. Different ways have indicated for one dimensional signal transformation to a matrix that in this article (here) have used Hankel's way [16,19].

3. Purposed algorithm

The main question on sound discovery is the classification of listening signal characteristic to diagnosis sound parts. Thus, listening signal are classified to sound classes: silence, voiced, and unvoiced. For classifying, suitable characteristics must elicit from the speech signal parts (frame). Before studying the details of purposed system, general block diagram are showed in figure 1.

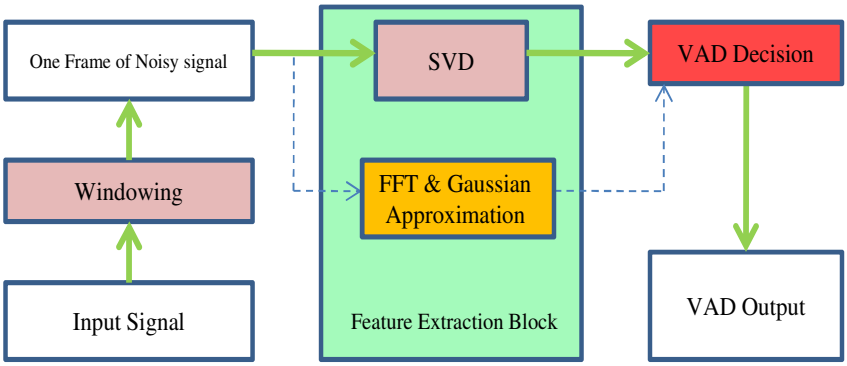


Figure (1) block diagram of propositional method for indicator system of voice activity.

As shown on figure 1, feature vector extraction done in two parts. In first part, voiced frames are separated from unvoiced and silence frame as for statistical characteristics of singular values matrix. In second part, unvoiced frames are separated from silence frames that this separation is based on its frequency spectrum and Gaussian rate in each frame. At the end, one value accrues to voiced and unvoiced frames that including voice information, and zeros one to silence frames.

3.1. Voiced frames separation from the other parts

In suggested system to separate the voiced parts from two parts of unvoiced and silence ones, it is used on slop of singular value curve in related part. For doing atop stages, first noisy signal divide to 16ms frames. According to relation (2), Hankel matrix makes for every frame.

$$X_k=[x_0,x_1,...,x_{n-1}]\rightarrow H_k=\begin{pmatrix} x_0 & x_1 & ... & x_{M-1} \\ x_1 & x_2 & ... & x_M \\ \vdots & \vdots & \vdots & \vdots \\ x_{L-1} & x_L & ... & x_{N-1} \end{pmatrix} \quad (2)$$

Where X_k is the vector of exist samples in K frame in input signal and H_k is the isomorphic Hankel matrix of L×M dimensions with X_k . The percent of sample overlapping in matrix and the conditions of H_k dimensions are brought on (3) and (4) relations.

$$\%overlapping=\frac{L-1}{L}\times 100 \quad (3)$$

$$M+L=N+1, L\geq M \quad (4)$$

For gaining full primitive pattern of X_k frame, the number of added zeros must lessen in M column of H_k matrix that its results have brought on (5) relation.

$$ZeroPadding=L-\text{mod}(N,L)\Rightarrow L=[\frac{N}{2}]+1, M=N-[\frac{N}{2}] \quad (5)$$

Gained values of L, M in (5) relation get H_k semi rectangular matrix with maximum sample overlapping. The singular values of each frame are got by Hankel matrix of existed frame and using SVD map on related singular values.

$$H_k=U_k\sum_k V_k^T \quad (6)$$

In (6) relation U_k and V_k the singular vectors of diagonal matrix and also $\sum_k$ are isomorphic singular values matrix with H_k (part 2). Figure (2), singular values vectors show the every voiced frames in 16 millisecond length and SNR=10db with white Gaussian noise.

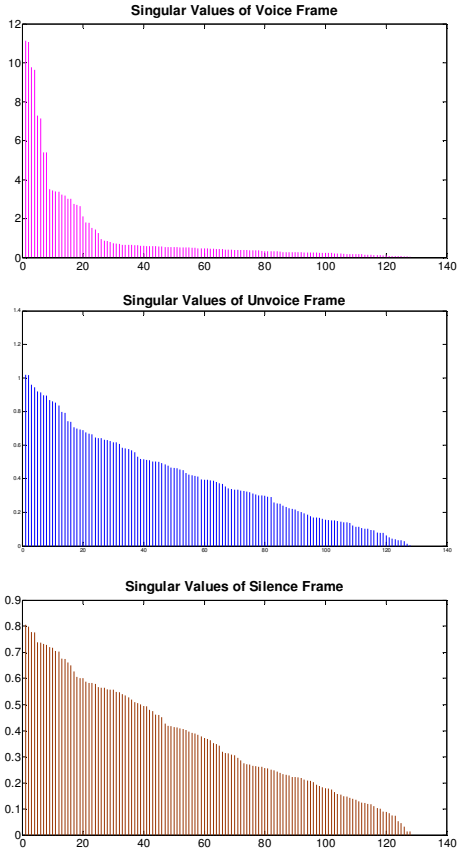


Figure (2) singular values vectors of voiced, unvoiced and silence frame in SNR=10db (voice signal from housewives_16.wav on TIMIT database)

According as seeing on figure (2) the slop of curve in singular values between voiced frames are different from the other parts. In fact, singular values of voiced frames have more slop than singular values of unvoiced and silence frames. The base of statistical feature extraction for separating voiced frames is the slop of singular values curve related to each frame by using linear regression. Table 1 shows the values of this slop in different SNRs on three certain frames that have chosen from voiced, unvoiced and silence parts.

Table (1):
slope of singular values curve in linear regression related to species of frame on different SNRs (voice signal from housewives_16.wav on TIMIT database)

SNR	Mean amount of singular values curve slope on 10 times repeat for each SNR		
	Voiced frame	Unvoiced frame	Silence frame
0db	0.1498	0.0987	0.0949
5db	0.1232	0.0566	0.0528
10db	0.1170	0.0338	0.0305
15db	0.1143	0.0232	0.0176
20db	0.1139	0.0175	0.0098

Results of table (1) are support on this thesis that the slop of singular values curve on different SNRs in voiced frames are more than the others and by using this trait we can achieve the goal of first section that was the feature vector extraction in voiced frames from related singular values matrix.

2.3. Separating the voiced and silence frames

By studying figure (2) and table (1) are deduced that according to approximation through slop of singular values curves related to voiced and silence frames, it can't divide these two type frames from each other. Because of this in purposed system have used the other trait to separate these two parts. In this part, frequency trait has used for recognition unvoiced frames from silence ones. The base of comparison is the curve of Gaussian function, (7) relation.

$$f(x;\gamma,c) = e^{\frac{-(x-c)^2}{2\gamma^2}} \tag{7}$$

Atop relation C is mean and γ is variance of curve. By doing study, we see the discrete Fourier transform of unvoiced frames in meddle frequency are similar to the curve of Gaussian functions to some extent. Figure (3) and (4) show the smooth frequency spectrum related to unvoiced and silence frames and their comparison with the curve of Gaussian function. (Voice signal from TIMIT database with SNR=15db)

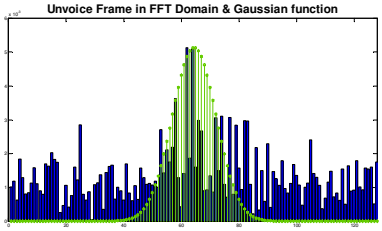


Figure (3) frequency spectrum of Unvoice frame in SNR=15db and the curve of Gaussian function ($\gamma = 0.6$)

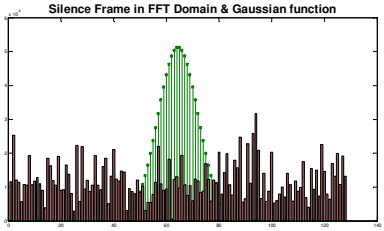
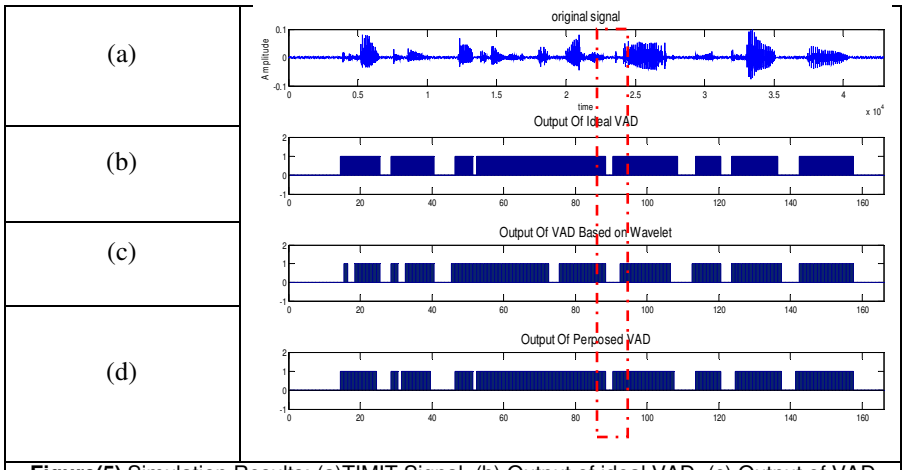


Figure (4) frequency spectrum of Silence frame in SNR=15db and the curve of Gaussian function ($\gamma = 0.6$)

According as atop figures are seen, frequency spectrum of voiced frame is more similar to Gaussian curve than silence frame noise in meddle frequency. Also by different examinations, the optimal values of γ parameter in varying SNRs are as 0.2 with a view of minimizing the difference between Gaussian pattern and frequency spectrum of silence frame pattern.

4. SIMULATION RESULTS

In this section, the operation of purposed system study and compare considered as accuracy and speed. Sx114.wav voice signal from TIMIT database including 166 of 16 millisecond frames with sampling rate is 16 kHz that each frame has 256 samples. Figure (5) shows the noisy signal with SNR=10db and the indicator systems output of voice activity.



Figure(5). Simulation Results: (a)TIMIT Signal, (b) Output of ideal VAD, (c) Output of VAD based on Wavelet, (d) Output of Purposed VAD

According as seeing, suggested method has more efficiency to keep the parts of signals that the activity signal is weak. In figure (5) and dashed line-drawn area, one unvoiced letter has omitted in VAD system as wavelet (figure 5-c) that it has kept on VAD output propositional system (figure 5-d). In table (2) the accuracy rate of VAD propositional system is shown as wavelet transform [17] about voice signal sx114.wav for different SNRs.

Table (2): percent of VAD system error comparison by using purposed algorithm and wavelet based algorithm in different SNRs

SNR	Purposed algorithm	Wavelet based algorithm
0db	25%	36%
5db	18%	20%
10db	15%	14%
15db	13%	12%
20db	8%	10%

One of the strength points of the purposed algorithm is its speed. In table (3), speed of two algorithms has compared with each other in processing the sx114.wav sound file (CPU Intel Core2Duo 2.5 GHz, 4 M Cache 733 MB RAM)

Table (3):
Speed comparison of VAD system by using purposed algorithm and wavelet based algorithm

Consumed time in purposed algorithm	Consumed time in wavelet based algorithm
4 second	12 second

In this part by using two VAD systems as preprocessing block has brought the results of hearing test for a speech signal rich-making system that accuracy of VAD operation system be proved in keeping unvoiced areas. In table 4 the standards has come that is used in evaluation of speech with hearing factor.

Table (4) Five-Point adjectival scales for quality and impairment, and associated scores	
Score	Impairment
5 (Excellent)	Imperceptible
4 (Good)	(Just) Perceptible but not Annoying
3 (Fair)	(Perceptible and) Slightly Annoying
2 (Poor)	Annoying (but not Objectionable)
1 (Bad)	Very Annoying (Objectionable)

In table 5, results of using two said VAD algorithms have shown as preprocessing block for enhancement method of multi band spectral subtraction. Specifications of this test are in [3].

Table (5) Results of MOS test; 17 clean speech signal from TIMIT database; Noise Type: White Gaussian Noise.			
Used Algorithm	Input SNR		
	0db	5db	10db
Wavelet Based	1.6	2.3	2.8
Purposed	1.8	2.7	3.3

Studying the results of table (5) are shown the reform efficiency of speech enhancement by using of propositional VAD algorithm to wavelet transform way.

5. CONCLUSION

In this paper a new method for Voice activity detector based on Singular Value Decomposition and discrete Fourier transform was proposed. The proposed method is evaluated by using various criteria. By using the mentioned criteria, it is presented that this method can compete with other methods. Also, the aim of indicator systems of voice activity is control of destruction unvoiced signal sites in rich-making operation that observantly to results of hearing test, the propositional algorithm have proper power in compare with wavelet transform way. The propositional system has manifold speed than usual method that is one of the obvious characters in practical usage and hardware implementation.

6. REFERENCES

[1] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, A. J. Rubio, “A new Kullback-Leibler VAD for speech recognition in noise”, IEEE Signal Processing Letters, 11(2) :266– 269. 2004.

[2] A. Kondoz, “Digital speech: coding for low bit rate communication systems”, J. Wiley, New York, 1994.

[3] Y. Ghanbari, M. R. Karami Mollaei, “A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets”, Speech Communication 48 (2006) 927–940.

[4] B. V. Harsha, “A Noise Robust Activity Detection Algorithm”, proc. Of int. symposium of intelligent multimedia, video and speech processing, pp. 322-325, Oct. 2004, Hong Kon.

[5] J. Ramırez, J. C. Segura, C. Benıtez, A. de la Torre, A. Rubio, “A New Adaptive Long-Term Spectral Estimation Voice Activity Detector,” EUROSPEECH, pp. 3041-3044, 2003, Geneva.

[6] J. Faneuff, “ Spatial, Spectral, and Perceptual Nonlinear Noise Reduction for Hands-free Microphones in a Car,” Master Thesis Electrical and Computer Engineering July 2002.

[7] S. Skorik, F. Berthommier, “On a Cepstrum-Based Speech Detector Robust To White Noise,” Accepted for Specom 2000, St. Petersbourg.
[8] J. Stegmann, G. Schroeder, “Robust Voice Activity Detection Based on the Wavelet Transform", Proc. IEEE Workshop on Speech Coding, Sep. 1997, pp. 99-100, Pocono Manor, Pennsylvania, USA.
[9] J.A. Haigh and J.S. Mason, “Robust Voice Activity Detection Using Cepstral Features,” In Proc. of IEEE TENCON'93, vol. 3, pp. 321-324, 1993, Beijng.
[10] J. Shaojun, G. Hitato, Y. Fuliang, “A New Algorithm For Voice Activity Detection Based On Wavelet Transform,” proc. Of int. symposium of intelligent multimedia, video and speech processing, pp. 222-225, Oct. 2004, Hong Kong.
[11] Tanyer S G, Ozer H, “ Voice activity detection in nonstationary guassian noise,” proceeding of ICSP’98 pp.1620-1623.
[12] Sangwan, A., Chiranth, M. C., Jamadagni, H. S., Sah, R.,Prasad, R. V., Gaurav, V., “VAD Techniques for Real-Time Speech Transmission on the Internet”, 5th IEEE International Conference on High-Speed Networks and Multimedia Communications, pp. 46–50, 2002.
[13] L. R. Rabiner, M. R. Sambur, “ Application of an LPC Distance Measure to the Voiced-Unvoiced-Silence Detection Problem,” IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-25, No. 4, pp. 338-343, August 1977.
[14] A. Vahatalo, I. Johansson, “Voice Activity Detection For GSM Adaptive Multi-Rate Codec,” IEEE 1999, pp. 55-57.
[15] M.Orlandi, a. santarelli, D. Falavigna, “Maximum Likelihood endpoint Detection with time-domain features,”eurospeech 2003, Geneva, pp.1757-1760.
[16] S. H. Jensen, P. C. Hansen, S. D. Hansen, “Reduction of Broad-Band Noise in Speech by Truncated QSVD,” IEEE, Trans on speech & Audio Processing, Vol.3, No.6, November 1995.
[17] Y. Ghanbari, M. Karami, “Spectral subtraction in the wavelet domain for speech enhancement,” International Conference on Information Knowledge Technology (IkT2004), CD ROM, 2004.
[18] H. Sameti, H. Sheikhzadeh, Li Deng, R. L.Brennan, “HMM-Based Strategies for Enhancement of Speech Signals Embedded in Nonstationary Noise”, IEEE Transactions on Speech and Audio Processing, Vol. 6, No. 5, September 1998.
[19] S. Sayeed, N. S. Kamel, R. Besar, “A Sensor-Based Approach for Dynamic Signature Verification using Data Glove”. Signal Processing: An International Journal (SPIJ), 2(1)1:10, 2008.
[20] J.Ghasemi, M. R. Karami Mollaei, "A New Approach for Speech Enhancement Based On Eigenvalue Spectral Subtraction" Signal Processing: An International Journal, (SPIJ) 3(4), 34:41, 2009.
[21] A. Afzalian, M.R. Karami Mollaei, J. Ghasemi, " A New Approach for Speech Enhancement Based On Singular Value Decomposition and Wavelet Transform", AJBAS In Press(2010).

Reducing Power Dissipation in FIR Filter: An Analysis

Manoj Garg manojgarg78@yahoo.co.in
M Tech Research Scholar
ECE Deptt, GZSCET,
Bathinda-151001, India

Dr. Rakesh Kumar Bansal bansal_r_k@yahoo.com
Asstt Professor
ECE Deptt, GZSCET,
Bathinda-151001, India

Dr. Savina Bansal hodecegzs@yahoo.com
Professor & Head
ECE Deptt, GZSCET,
Bathinda-151001, India

Abstract

In this paper, three existing techniques, Signed Power-of-Two (SPT), Steepest decent and Coefficient segmentation, for power reduction of FIR filters are analyzed. These techniques reduce switching activity which is directly related to the power consumption of a circuit. In an FIR filter, the multiplier consumes maximum power. Therefore, power consumption can be reduced either by making the filter multiplier-less or by minimizing hamming distance between the coefficients of this multiplier as it directly translates into reduction in power dissipation [8]. The results obtained on four filters (LP) show that hamming distance can be reduced upto 26% and 47% in steepest decent and coefficient segmentation algorithm respectively. Multiplierless filter can be realized by realizing coefficients in signed power-of-two terms, i.e. by shifting and adding the coefficients, though at the cost of shift operation overhead.

Keywords: FIR, SPT, Steepest decent, Coefficient segmentation, low pass filter.

1. INTRODUCTION

The need for higher battery lifetime is ever increasing for portable devices like cellular phones, laptops, calculators, hearing aids and numerous other such devices. Due to large scale component integrations in such devices, power dissipation has become a major issue demanding special measures (heat sinks, special circuitry etc) to protect the chip from thermal runaway making the system complex and costly. So minimizing power dissipation of an application chip is a vital issue before researchers that needs to be addressed. In FIR filters, multipliers play an important role, and in a well designed CMOS circuit, switching component is a dominant term [5]. Input level transitions (0 to 1 or 1 to 0) are always associated with a power loss, which can be reduced by reducing toggling in the input of the multiplier. Power dissipation in a CMOS circuit is given as [5]

$P_{total} = P_{dynamic} + P_{short} + P_{leakage}$,

with $P_{dynamic} = \alpha.C_{load}.V^2_{dd}.f_{clk}$, $P_{short} = I_{sc}*V_{dd}$, and $P_{leakage} = I_{leakage}*V_{dd}$, where V_{dd} is the supply voltage, f_{clk} is the clock frequency, C_{load} is the load capacitance, α is the node transition factor i.e., the average number of 0 → 1 transitions for the equivalent electric node per clock cycle. I_{sc} and $I_{leakage}$ are the short circuit and leakage current respectively. This paper aims at reducing the switching factor, α, out of these factors while maintaining the throughput of the FIR filter.

2. RELATED WORK

Many methods have been reported for the reduction of power dissipation in FIR filter. Power reduction can be done at four levels- process, circuit, architectural, and algorithmic level. A multiplier in a FIR filter is most power consuming component and its implementation in VLSI is also very expensive. The coefficients of filter can be represented as sum of signed power-of-two terms making it multiplierless. Several algorithms are available for designing filters with signed power-of-two (SPT) coefficients [1-4, 12, 13]. Thus, multiplier can be replaced by adders and shifters, making the filter multiplierless. Also, adders can further be reduced by extracting common subexpressions from the coefficients [18].

Transition density of the inputs of multiplier can be reduced by minimizing hamming distance between the inputs [8, 11]. Mahesh Mehendale et. al. [9] presented seven transformations to reduce power dissipation which together operate at algorithmic, architectural, logic and layout levels of design abstraction. Power can be reduced by the way of arithmetic operators which use coding as a method of decreasing the switching activity [10]. In [14], the authors presented different architectures for implementation of low power FIR filtering cores. This included coefficient segmentation, block processing and combined segmentation and block processing algorithms. The work in [15] presents low power FIR filter implementations which are based on processing coefficients in a non-conventional order using both direct form and transposed direct form FIR filters.

3. FIR FILTER IMPLEMENTATION

FIR filters are widely used in digital signal processing (DSP) systems that are characterized by the extensive sequence of multiplication operations. These are also used in IF processing block of wireless communication systems. It can be represented by difference equation as given below:

$$y(n) = \sum_{k=0}^{M-1} h(k)\, x(n-k)$$

where h(k) are the coefficients of the filter, x(n) is the input , y(n) is the output and M is the order of the filter. Such a difference equation can be implemented using a transversal filter as shown in Fig 1. The transversal filter, which is also referred to as a tapped delay line filter, consists of three basic elements: (1) unit-delay element, (2) multiplier, and (3) adder. The number of delay elements used in the filter determines the finite duration of its impulse response. The number of delay elements, shown as M in Fig.1 is commonly referred as the filter order. The role of each multiplier in the filter is to multiply the tap input by a filter coefficient called tap weight.

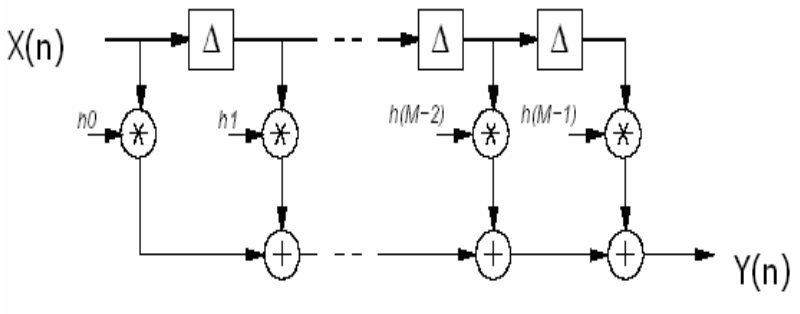


FIGURE 1: Transversal FIR Filter

4. SIGNED POWER-OF-TWO

The coefficients of fixed point FIR filter can be represented in sum of signed power-of-two (SPT) terms. In binary, multiply by 2 is implemented by simply shifting and adding the term. So in FIR, multiplier may be implemented by shifting the coefficients and then adding to input data. In this way, the power and complexity of multiplier can be minimized. As complexity of fixed point FIR filter is directly related to the number of total SPT terms[13], this number should be limited if complexity is also one of the constraint. The number of SPT terms per coefficient may be same. In the past decades, it has been shown that significant advantage can be achieved if the coefficient values are allocated with different number of SPT terms while keeping the total number of SPT terms for the filter fixed. Each coefficient value represented as a sum of signed power-of-two terms is given by the following equation [4]:

$$h(n) = \sum_{r=1}^{R} a(r)\, 2^{g(r)}$$

where a(r)=(-1, 0 or 1), h(n) is the nth coefficient value, g(r) is a positive integer and R is the number of terms.

The following algorithm [4] finds an approximation of a SPT term for a number x:

(i) Initialize m = 1 and $S_0 = x$.

(ii) Find $y(m)*2^{g(m)}$ which minimizes $| S_{m-1} - y(m)2^{g(m)} |$

(iii) If either y(m) = 0 or m = u, go to Step (vi), else go to Step (iv).

(iv) Update $S_m = S_{m-1} - y(m)*2^{g(m)}$

(v) Increment m. Go to Step 2.

(vi) $[x]u = \sum_{i=1}^{M} y(i)*2^{g(i)}$. Stop.

5. STEEPEST DECENT TECHNIQUE

Power dissipation can be reduced by minimizing the hamming distance between the coefficients of the filter by using steepest decent technique. It is a local search method in which a minima of the function can be found in its neighborhood. Hamming distance gives the measurement of the number of bit changes in the successive coefficients. As switching activity is proportional to hamming distance, it can be reduced by minimizing hamming distance. Algorithm for hamming distance minimization of FIR filters using steepest decent technique is as under [8]:

Step 1:- For a given FIR filter coefficients H[i] (i =1, N-1) and given pass band ripple and stop-band attenuation, calculate the Hamming Distance between successive coefficients.

Step 2:- Now perturb each coefficient (increase the value of each coefficient one by one by 1) and calculate new hamming distance between the coefficients H[i+], H[i-1] such that HD(H[i], H[i-1]) > HD(H[i+], H[i-1])

Step 3:- Replace H[i] with H[i+] to get a new set of coefficients.

Step 4:- Now again perturb each coefficient (decrease the value of each coefficient one by one by 1) and calculate new hamming distance between the coefficients H[i-], H[i-1] such that HD(H[i], H[i-1]) > HD(H[i-], H[i-1])

Step 5:- Replace H[i] with H[i-] to get a new set of coefficients.

Step 6:- Compare the two sets and replace original coefficients with new value i.e. which having smaller hamming distance.

Step 7:- Again compute hamming distance between new coefficients and also find pass band ripple and stop band attenuation for this new set of coefficients.

Using this steepest decent strategy, for every coefficient, it's nearest higher and nearest lower coefficient values are identified. A new set of coefficient is formed by replacing one of the coefficients with it's nearest higher or nearest lower value i.e. having minimum hamming distance. Hamming distance is then calculated for this new set of coefficients. Also passband ripples and stopband attenuation is calculated.

6. COEFFICIENT SEGMENTATION

Using this algorithm, a coefficient is divided into two parts. For each coefficient the nearest signed power-of-two term is find and is subtracted from the original coefficient. This SPT term can be implemented using a shift operation and the rest of the coefficient value can be given to the multiplier in filter. In this way, the bits required to represent the coefficients becomes less. So the hamming distance between the successive coefficients, applied to multiplier, is reduced which in turn reduces switched capacitance. Hence, power consumption is reduced by reducing switching power. The main algorithm for segmentation is described as follow [11]:

Let H = (h_0, h_1,...., h_{N-1}), where N is the order of the filter. For a coefficient h_k, it is divided in such a way that $h_k = s_k + m_k$, where s_k is the component to be implemented using shift operation and the second component m_k is applied to the multiplier. To reduce the switched capacitance of the hardware multiplier, consecutive values of m_k, applied to the multiplier input must be of the same polarity, to minimize switching, and have the smallest value possible, to minimize the effective wordlength. In this case, m_k is chosen to be the smallest positive number. For a small positive m_k, s_k, must be the largest power of two number closest to h_k. So by

checking the polarity of h_k, if it is a positive number then s_k is chosen as the largest power of two number smaller then h_k. If it is negative, s_k is chosen as smallest power-of-two number larger than $|h_k|$. In both cases m_k is $h_k - s_k$.

7. RESULTS

The three algorithms (SPT, coefficient segmentation and steepest decent) were implemented under MATLAB environment on four different FIR low pass filters, with varying parameters (table 1) and number of coefficients. These filters were designed initially using window method. Coefficient values, obtained using these algorithms and quantized to 16-bit 2's complement representation form the initial set of coefficients for optimization. Hamming distance is first calculated for a particular set of coefficients using window method. These coefficients are then calculated by using these algorithms and the hamming distance is recalculated. It is gathered from these results that upto 26% and 47% reduction in hamming distance is achievable using steepest decent and coefficient segmentation algorithm as shown in table 2. The passband ripple and stopband attenuation is also compared for all the cases as shown in table 3 and 4. Performance parameters degrade somewhat. The frequency response of filter 1 obtained for all the algorithms is shown in Fig. 2.

Filter	Passband (kHz)	Stopband (kHz)	Passband ripple (dB)	Stopband attenuation (dB)	Window function	Filter length
Filter 1	0-1.5	2-4	0.1	50	Hamming	53
Filter 2	0-1.2	1.7-5	0.01	40	Kaiser	71
Filter 3	0-1	1.5-5	0.0135	56	Kaiser	61
Filter 4	0-1.5	2-4	0.1	50	Blackman	89

TABLE 1: Low Pass Fir Filter Specifications

Filter	Initial Hamming Distance (HD)	Coefficient Segmentation		Steepest Decent	
		HD	Reduction (%)	HD	Reduction (%)
Filter 1	326	216	33.74	278	14.72
Filter 2	378	256	32.27	300	20.63
Filter 3	326	244	25.15	244	25.15
Filter 4	538	284	47.21	396	26.39

TABLE 2: Hamming distance obtained using these algorithms for different low pass filters

Filter	Performance parameters	Original	Coefficient Segmentation	Steepest Decent	SPT
Filter 1	Max (δp)	1.0042	1.0041	1.0044	1.0134
	Min (δp)	1.0000	1.0000	1.0001	1.0074
Filter 2	Max (δp)	1.0003	1.0002	1.0004	0.9937
	Min (δp)	0.9987	0.9987	0.9982	0.9919
Filter 3	Max (δp)	1.0012	1.0012	1.0056	1.0019
	Min (δp)	0.9985	0.9985	0.9833	0.9988
Filter 4	Max (δp)	1.0002	1.0001	1.0076	1.0127
	Min (δp)	0.9999	0.9998	0.9922	1.0085

TABLE 3: Comparison in terms of passband ripples

Filter	Original	Coefficient Segmentation	Steepest Decent	SPT
Filter 1	-52	-52	-54	-43
Filter 2	-57	-57	-56	-45
Filter 3	-56	-56	-36	-40
Filter 4	-75	-75	-41	-46

TABLE 4: Comparison in terms of stopband attenuation(in db)

8. CONCLUSION

In this work, Signed power-of-two, Steepest decent and coefficient segmentation algorithms have been discussed for the low power realization of FIR filters. As, a multiplier is the major component for power dissipation, power can be reduced by minimizing the hamming distance between successive filter coefficients which are being fed to the multiplier. Steepest Decent optimization algorithm is presented so as to minimize the Hamming distance and the analysis shows that the total Hamming distance can be reduced upto 26%. But the penalty paid in this case is the degradation of performance parameters like. passband ripples and stopband attenuation. The results of coefficient segmentation shows that the hamming distance can be reduced upto 47%. Here the performance parameters are not degraded but an additional overhead, in terms of extra hardware and the power dissipated for shift and add operation, caused by shift operation is to be added. Also as reported in literature [11] power dissipation can be reduced considerably (63%) using SPT, as compared to these algorithms. However, the penalty is again in the form of additional overhead of adders and shifters. There exist a tradeoff between hamming distance reduction and degradation of performance parameters.

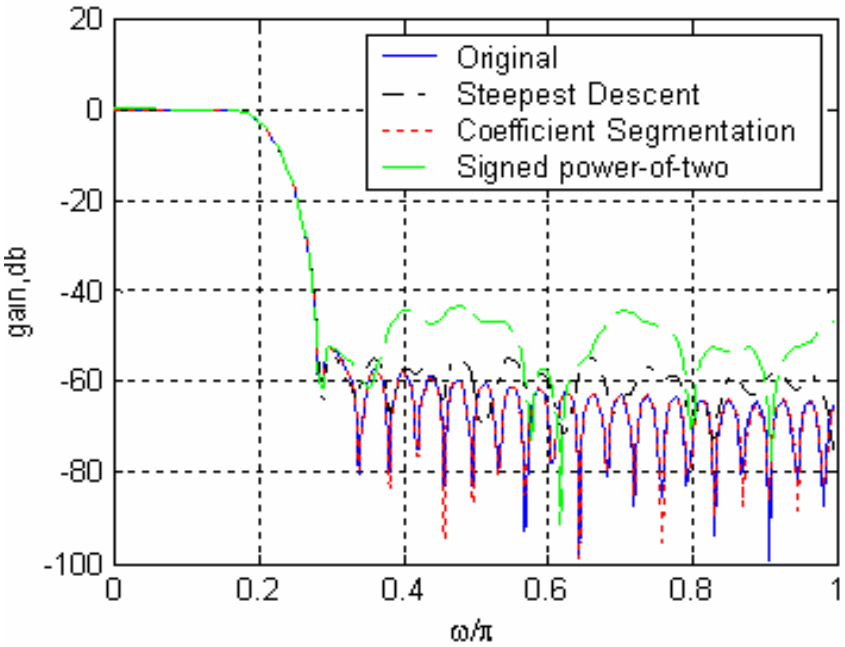


FIGURE 2: Frequency response of low pass FIR filter for example 1

9. REFERENCES

[1] Y. C. Lim and Sydney R Parker, *"FIR Filter Design over a Discrete Powers-of-Two Coefficient Space"*, IEEE Trans., Vol. ASSP-31, No. 3, pp. 583-591, June 1983.

[2] Quangfu Zhao and Yoshiaki Tadokoro, *"A Simple Design of FIR Filters with Powers-of-Two Coefficients"*, IEEE transactions circuits and systems, Vol. 35, No. 5, pp. 566-570, May 1988.

[3] Henry Samuel1, *"An Improved Search Algorithm for the Design of Multiplier less FIR Filters with Powers-of-Two Coefficients",* IEEE transactions on circuits and systems, Vol. 36, No. 7, pp. 1044-1047, July 1989.
[4] Yong Ching Lim, Joseph B. Evans and Bede Liu, *"Decomposition of Binary Integers into Signed Power-of-Two terms",* IEEE transactions on circuits and systems, Vol. 38, No. 6, pp. 667-672, June 1991.
[5] Anantha P. Chandrakasan, Samuel Sheng, and Robert W. Brodersen, *"Low Power CMOS Digital Design"*, IEEE Journal of Solid Stare Circuits, vol. 27, no. 4, pp. 473-484, Jan 1992.
[6] Farid N. Najm, *"Transition density, a new Measure of Activity in Digital circuits"*, IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems, vol. 12, No. 2, pp. 310-323, Jan 1993.
[7] Anantha P. Chandrakasan, Miodrag Potkonjak, Renu Mehra, Jan Rabaey, and Robert W. Brodersen, *"Optimizing Power Using Transformations",* IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems, vol. 14, No. 1, pp. 12-31, January 1995.
[8] M. Mehendale, S. D. Sherlekar, and G. Venkatesh, *"Coefficient optimization for low power realization of FIR filters,"* in Proc. IEEE Workshop VLSI Signal Processing, pp. 352–361, 1995.
[9] Mahesh Mehendale, S.D. Sherlekar and G.Venkatesh, *"Low Power Realization of FIR Filters on Programmable DSPs",* IEEE transactions on VLSI Systems, Vol. 6, No. 4, pp. 546-553, Jan 1998.
[10] Eduardo Costa, Sergio Bampi, Jose Monteiro, *"FIR filter design using low power arithmetic operators"*, Eleventh International conference on VLSI design, pages 12-17, 1998.
[11] A. T. Erdogen and T. Arslan, *"Low Power Coefficient Segmentation Algorithm for FIR filter Implementation",* IEEE Electronics letters, Vol. 34, Issue 19, pp. 1817-1819, Sept.17, 1998.
[12] Yong Ching Lim, Rui Yang, Dongning Lia and Jianjian Song, *"Signed Power-of-Two term allocation scheme for the design of digital filters",* IEEE transactions on Circuits and Systems—II: Analog and Digital Signal Processing, Vol. 46, No. 5, pp. 577-584, May 1999.
[13] Chia-Yu Yao and Chiang-Ju Chien, *"A Partial MILP algorithm for the Design of Linear Phase FIR filters with SPT coefficients",* IEICE Transactions fundamentals, vol. E85-A, no. 10, October 2002.
[14] A.T. Erdogan, M. Hasan and T. Arslan, *"Algorithmic Low Power FIR Cores",* IEE Proceedings of Circuit, System and Devices, Vol. 150, No. 3, pp. 23-27, June 2003.
[15] A.T. Erdogan and T. Arslan, *"Low power FIR filter implementation based on Coefficient ordering algorithm",* proc. of IEEE computer society annual symposium on VLSI emerging trends in VLSI systems design, September 2004.
[16] Emmanuel C. Ifeachor and Barrie W. Jervis, *"Digital Signal Processing – A practical approach"*, Second Edition, Pearson Education, 2004.
[17] Mohamed Al Mahdi E**s**htawie and Masuri Bin Othman, *"An algorithm proposed for FIR filter coefficients representation",* IJAMCS, vol. 4, no. 1, 2007.
[18] Ya Jun Yu and Y.C. Lim, *"Design of linear phase FIR filters in subexpression space using mixed integer linear programming",* IEEE transactions on circuits and systems-I, vol.54, no. 10, October 2007.

CALL FOR PAPERS

Journal: Signal Processing: An International Journal (SPIJ)
Volume: 4 **Issue:** 1
ISSN:1985-2339
URL: http://www.cscjournals.org/csc/description.php?JCode=SPIJ

About SPIJ

The International Journal of Signal Processing (SPIJ) lays emphasis on all aspects of the theory and practice of signal processing (analogue and digital) in new and emerging technologies. It features original research work, review articles, and accounts of practical developments. It is intended for a rapid dissemination of knowledge and experience to engineers and scientists working in the research, development, practical application or design and analysis of signal processing, algorithms and architecture performance analysis (including measurement, modeling, and simulation) of signal processing systems.

As SPIJ is directed as much at the practicing engineer as at the academic researcher, we encourage practicing electronic, electrical, mechanical, systems, sensor, instrumentation, chemical engineers, researchers in advanced control systems and signal processing, applied mathematicians, computer scientists among others, to express their views and ideas on the current trends, challenges, implementation problems and state of the art technologies.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for SPIJ.

SPIJ List of Topics

The realm of International Journal of Signal Processing (SPIJ) extends, but not limited, to the following:

- Biomedical Signal Processing
- Communication Signal Processing
- Detection and Estimation
- Earth Resources Signal Processing
- Industrial Applications
- Optical Signal Processing
- Acoustic and Vibration Signal Processing
- Data Processing
- Digital Signal Processing
- Geophysical and Astrophysical Signal Processing
- Multi-dimensional Signal Processing
- Pattern Recognition

- Radar Signal Processing
- Signal Filtering
- Signal Processing Technology
- Software Developments
- Spectral Analysis
- Stochastic Processes
- Remote Sensing
- Signal Processing Systems
- Signal Theory
- Sonar Signal Processing
- Speech Processing

CFP SCHEDULE

Volume: 4
Issue: 2
Paper Submission: March 31 2010
Author Notification: May 1 2010
Issue Publication: May 2010

CALL FOR EDITORS / REVIEWERS

CSC Journals is in process of appointing Editorial Board Members for ***Signal Processing: An International Journal (SPIJ)***. CSC Journals would like to invite interested candidates to join **SPIJ** network of professionals/researchers for the positions of Editor-in-Chief, Associate Editor-in-Chief, Editorial Board Members and Reviewers.

The invitation encourages interested professionals to contribute into CSC research network by joining as a part of editorial board members and reviewers for scientific peer-reviewed journals. All journals use an online, electronic submission process. The Editor is responsible for the timely and substantive output of the journal, including the solicitation of manuscripts, supervision of the peer review process and the final selection of articles for publication. Responsibilities also include implementing the journal's editorial policies, maintaining high professional standards for published content, ensuring the integrity of the journal, guiding manuscripts through the review process, overseeing revisions, and planning special issues along with the editorial team.

A complete list of journals can be found at http://www.cscjournals.org/csc/byjournal.php. Interested candidates may apply for the following positions through http://www.cscjournals.org/csc/login.php.

Please remember that it is through the effort of volunteers such as yourself that CSC Journals continues to grow and flourish. Your help with reviewing the issues written by prospective authors would be very much appreciated.

Feel free to contact us at coordinator@cscjournals.org if you have any queries.

Contact Information

Computer Science Journals Sdn BhD
M-3-19, Plaza Damas Sri Hartamas
50480, Kuala Lumpur MALAYSIA

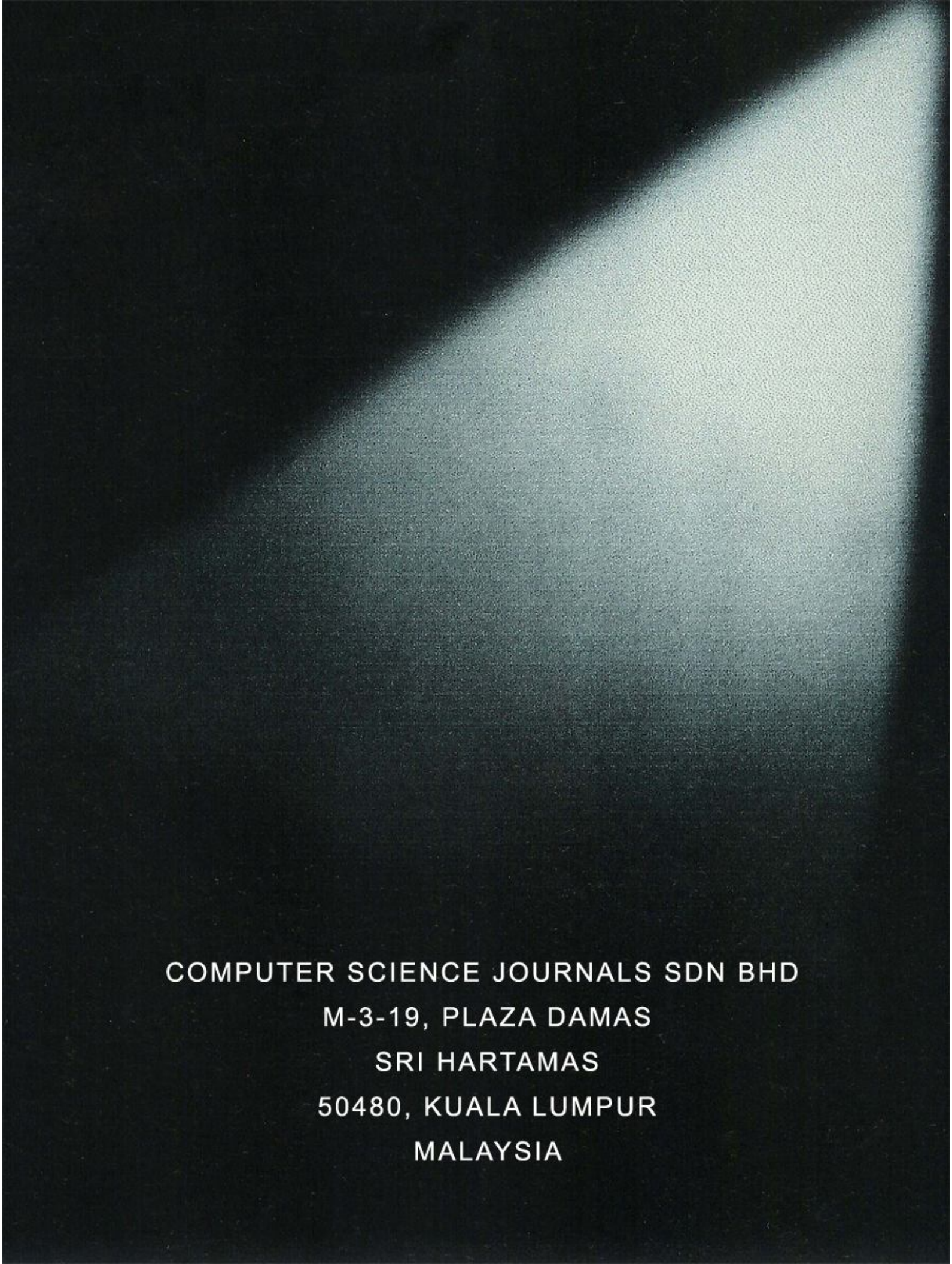
Phone: +603 6207 1607
+603 2782 6991
Fax: +603 6207 1697

BRANCH OFFICE 1
Suite 5.04 Level 5, 365 Little Collins Street,
MELBOURNE 3000, Victoria, AUSTRALIA

Fax: +613 8677 1132

BRANCH OFFICE 2
Office no. 8, Saad Arcad, DHA Main Bulevard
Lahore, PAKISTAN

EMAIL SUPPORT
Head CSC Press: coordinator@cscjournals.org
CSC Press: cscpress@cscjournals.org
Info: info@cscjournals.org



COMPUTER SCIENCE JOURNALS SDN BHD
M-3-19, PLAZA DAMAS
SRI HARTAMAS
50480, KUALA LUMPUR
MALAYSIA