# An Improvement on Vertical Phase Coherence of Scaled Phase Locking Phase Vocoder for Speech Signal

Ahmad Reza Eskandari, Zahra Kouchaki

**Abstract**— Phase vocoder is a popular algorithm for speech time scale modification. The performance of phase vocoder is highly depends on the accuracy of finding the place of instantaneous frequencies. Recent work that addressed this issue proposed Multi-resolution peak picking method. This method assumes instantaneous frequencies happen in local maxima of frequency spectrum whereas they are distributed exponentially. We improved this method by introducing a non-stationary peak-picking algorithm. Simulation results indicate remarkable increases in output quality according to PESQ, LLR and MOS criteria.

**Index Terms**— phase vocoder, scale phase locking, peak picking, multiresolution peak picking, non-stationary multiresolution peak picking.

## 1 Introduction

Audio time-scale modification is the process of changing the duration of an input audio signal while retaining the signal local frequency content, resulting in the overall effect of speeding up or slowing down the perceived playback rate of a recorded audio signal without affecting quality, pitch, timbre or naturalness of the original signal. Audio time-scale modification has many applications as language and music learning, fast playback for telephone answering machines and audio-video synchronization in broadcasting applications.

The Phase Vocoder is a popular method of audio time-scale modification due to its ability to achieve high quality modification on a variety of signals within a wide range of time-scaling factor. This technique was introduced by Flanagan and Golden [1] and is put into its efficient FFT based form in [2], by making use of the short-time Fourier transform (STFT). A tutorial on the phase vocoder is presented in [2], [3] stated that most of the phase Vocoder problems are due to the loss of what they call vertical phase coherence. They proposed the identity phase locking and the scaled phase-locking techniques to preserve the vertical phase coherence in the time-stretched audio. Their phase-locking technique is based on the explicit identification of peaks in the frequency spectrum and the assumption that they are sinusoids. In this case, they assumed that the sinusoids, which can be found in each peak, have transmitted from the peak in the nearest channel of the previous frame. Various definitions for peaks can be mentioned. For example, the peak is the member that is greater than two adjacent members. This type of definition is not dependent on the frequency and a clear definition is used for every frequency channel to search the sinusoidal. This method of peak picking is named as the constant resolution peak-picking [5]. Although this method is simple, unfortunately it has undesirable effects on the output signal that is heard similar to the shallow bass or musical over tone. Such as when an audio files is not correctly compacted to mp3 file. The method of peak picking that proposed by Laroche and Dolson [4] has taken under further improvement by the work of Karrer et al [5]. As it could be seen in Table.1, Karrer et al [5] proposed to divide the frequency spectrum into seven frequency channels .As a result, the channel bandwidth increases exponentially. In each channel, the constraint of peak–picking is different which is explained in detail in section 3. Their method is inspired by the nonlinear property of the speech signal and the human auditory system**.**

In the proposed method by Karrer et al [5], the channel division method is considered constant for all the frames. This kind of channel division is inconsistent with the non-stationary property of audio signals. Moreover, the peak distribution has simple exponential form in their method. Although this model approximates the non-linear properties of the speech signal, it is different with the standard methods that model the non-linear properties of the human ear.

In this work, we proposed a non-stationary constraint for peak picking using a standard nonlinear model of the human auditory system. In our method, the speech signal is divided to voiced and unvoiced parts. For unvoiced parts, we select more peaks than the voiced parts. In other words, by applying looser constraint for the peak selection in the unvoiced part, the numbers of sinusoidals that are used for the modeling of the voiced part are more than voiced parts. The simulation results indicate the remarkable increases in the output quality according to PESQ, LLR and MOS criteria.

The rest of this paper is organized as follows. In section 2, the

- *Digital Media Lab, AICTC Research Center, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, E-mail: eskandari@dml.ir*
- *Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran, E-mail: parisakouchaki@gmail.com*

scale phase locking and its improvement by multiresolution peak picking are explained. The disadvantageous of the multiresolution peak-picking and our improvement are presented in detail in section 3. The experimental results are discussed in section 4. Finally, we conclude the paper in section 5.

## 2 Phase Vpcoder

### 2.1 Scaled Phase-Locking

In this section, we briefly summarize the scaled phase-locking phase vocoder that is presented by Laroche and Dolson [4]. This technique starts by dividing the input signal x (t) into the overlapping windows of the length N that start every $R_a$ samples. Each window grabs a weighted part of the signal that is called frame. Then FFT of the resulting frames are separately taken and finally we will have a discrete time-frequency spectrum. The process of getting discrete time-frequency spectrum is called the short time Fourier transform (STFT). We assume here that the signal could be analyzed as a set of sinusoids which are not necessarily harmonics of each others. Desired time-scale modification can be obtained by changing the amount of the overlapping between frames in synthesis section. With assuming $\alpha$ as the scale factor, the distance of the windows during the synthesis ($R_s$) will be as follows:

$$R_s = \alpha R_a \quad (1)$$

where $R_a$ and $R_s$ are called the analysis hop factor and the synthesis hop factor, respectively. The phases of the STFT must be adjusted prior to this replacing to avoid phase jumps between the re-spaced windows. Phase adjusting starts by identifying the bins that contain peaks in the amplitudes of the STFT spectrum, assuming that those bins correspond to the most important sinusoids in the signal. The synthesis phases of these bins are computed by considering that a sinusoid might switch from channel $k_0$ at frame $u-1$ to the channel $k_1$ at the frame $u$, and uses the appropriate analysis phases when calculating the instantaneous frequency. Let $t_a^u = R_a$ and $t_s^u = R_s$ $(u \in N)$ be the corresponding analysis and synthesis time points. Moreover, $\Omega_k = \frac{2\pi k}{N}$ denotes the center frequency of the k$_{\text{th}}$ FFT-bin, $X(t_a^u, \Omega_k)$ denotes the STFT of the input signal x(t) at the time t using a window function h(t) of the length N, and $Y(t_s^u, \Omega_k)$ denotes the modified STFT from which the output signal y(t) will be synthesized. To calculate the correct synthesis phases, the instantaneous frequency $\widehat{\omega}_k(t_a^u)$ has to be determined for each FFT-bin $k$. In particular, for bin $k_1$ we have:

$$\Delta\phi_{k_1}^u = \angle X(t_a^u, \Omega_{k_1}) - \angle X(t_a^{u-1}, \Omega_{k_0}) - R_a \Omega_{k_1} \quad (2)$$

$$\widehat{\omega}_{k_1}(t_a^u) = \Omega_{k_1} + \frac{\Delta_p \phi_{k_1}^u}{R_a} \quad (3)$$

Where $\Delta_p \phi_{k_1}^u$is the principal determination ($\in [-\pi, \pi]$) of the phase deviation$\Delta\phi_{k_1}^u$. Now, the synthesis phases $\angle Y(t_s^u, \Omega_{k_1})$ can be found by advancing from the synthesis phases in the last time frame $Y(t_s^{u-1}, \Omega_{k_0})$ for the duration of one synthesis hop $R_s$ at the rate of the FFT-bin's instantaneous frequency$\widehat{\omega}_{k_1}(t_a^u)$.

$$\angle Y(t_s^u, \Omega_{k_1}) = \angle Y(t_s^{u-1}, \Omega_{k_0}) + R_s \widehat{\omega}_{k_1}(t_a^u) \quad (4)$$

To determine which peak in the frame $u-1$ corresponds to the peak at $k_1$ in the frame u, the peak in the frame $u-1$ that is closest to channel $k_1$ is identified. The non-peak bins are said to belong to the closest peak's region of influence, and are then phase-locked to that peak using the phase-locking equation.

$$\angle Y(t_s^u, \Omega_k) = \angle Y(t_s^u, \Omega_{k_1}) + \beta\left[\angle X(t_a^u, \Omega_k) - \angle X(t_a^u, \Omega_{k_1})\right] \quad (5)$$

Where $\beta$ is a phase scaling factor.

In this way, the phase relations between each peak channel and its neighboring nonpeak channels are carried over from the original signal to the time-stretched signal which helps to preserve the vertical phase coherence.

### 2.2 Multiresolution Peakpicking (MP)

Laroche and Dolson's scaled phase-locking technique [4], identifies the peaks of the signal that are to be phase-locked using a simple local maximum that search over a fixed interval in the frequency spectrum. This peak picking technique is referred as *constant resolution peak-picking*. Unfortunately, such a scheme which is simple, introduces artifacts in the resulting audio signal that can be heard as shallow bass and musical overtones, similar to a badly compressed MP3 file. In order to solve these two issues, Karrer et al proposed the multiresolution peak-picking method [5]. In their work, based on the non-uniform characteristics of the human auditory system and audio signals (especially music and speech), a non-uniform peak-picking method is presented. They suggested that the peaks in the frequency axis distributed exponentially. In contrast with constant resolution peak-picking, the multiresolution case, not only increases the quality of the output signals but also decreases the computation time. In multiresolution peak-picking, the detector function selects the peaks with high resolution in lower frequency band and with low resolution in higher frequencies. As could be seen in Table.1, in implementation proposed in [5], the 4096 bins of the FFT are divided into seven sub-bands. They assumed that the lowest 16 frequency bins contain a peak corresponding to a sinusoid that is audible to the human ear. For the next 16 bins, a bin is considered to contain a peak if its amplitude is larger than both of its neighboring bins. For the next 32 bins, the amplitude must be larger than its two neighboring bins to either side, and so on. The maximum distance in the frequency for searching the peak predecessors in those sub-bands is equal to the index number of each sub-band. Thus, if a peak was present in bin 142 at time u, it would continue the sinusoidal trajectory of the closest peak at time u − 1 in the region between bins 129–256. If there were no peaks in this region at time u – 1, the peak in bin 142 would be considered to start a new sinusoidal trajectory.

Table1 The spectrum division proposed by (Karrer et al, 2006); the band width increases exponentially.

| Sub-band Index | FFT Bin Range |
|---|---|
| 1 | Bins 0 -16 ≃ 0 Hz – 172 Hz |
| 2 | Bins 17 -32 ≃ 183 Hz – 345 Hz |
| 3 | Bins 33 -64 ≃ 355 Hz – 689 Hz |
| 4 | Bins 65 -128 ≃ 700 Hz – 1378 Hz |
| 5 | Bins 129 -256 ≃ 1389 Hz – 2756 Hz |
| 6 | Bins 257 -512 ≃ 2767 Hz – 5513 Hz |
| 7 | All remaining bins |

## 3 Non Stationary Multiresolution Peak Picking (NSMP)

### 3.1 Disadvantageous of multi-resolution peak-picking

As mentioned before and could be seen in Table.1, in multi-resolution peak-picking technique that proposed by [5], the frequency spectrum is divided into seven frequency channels and the channel band width increases exponentially. In each channel, the constraint of peak–picking is different which is explained in proceeding of this section. This model is used because of the nonlinear properties of the human auditory system and frequency spectrum of the sound signal. Following terms is proposed to improve Multi-resolution peak-picking algorithms:

- In this approach, the bandwidth distribution is in exponential form. Although this model approximates the non-linear properties of the speech signal, it is different with the standard models that model non-linear properties of the human auditory system. Here, we used *Mel distribution model* instead of using *the exponential model* [6].
- In the approach presented by Karrer et al [5], the channel division is considered stationary for all the frames which is not consistent with the non-stationary property of the audio signals. Here, we propose two different style of peak picking for voiced and unvoiced parts of the speech signal.

In proceeding, by dividing the speech signal into voiced and unvoiced parts and assigning different peak selection constraint for each type of parts, we cope with these two issues. We called this technique as non-stationary multi resolution peak picking technique which is explained in the next section.

### 3.2 Voiced and unvoiced

One of the main characteristics of the speech signals is the type of excitation. We have two main excitations which according to that excitation, the output signal is divided into voiced and unvoiced parts. The Voiced and unvoiced parts have different frequency spectra [7]. The voiced parts are created by the fluctuation of the vocal cords that produce pitch and its harmonics. So that, the output signal shape is periodic whereas the mouth act as a resonator that causes reinforcement or weakening of some harmonics. In other words, the vocal track will change the spectrum of the output signal of the vocal cords. This variation depends on the status of the jaw, tongue and lips. In producing unvoiced phonemes, the vocal cords do not fluctuate and the emissive output air of the larynx, which is not affected by vocal cords, would be the source for producing unvoiced phonemes. Finally, unvoiced parts are produced through vocal track that acts as filter. The important note is that in unvoiced parts, the source of excitation can be modeled with random noise that passes through vocal track (filter). Although in this model unvoiced phonemes are produced by noise, but previous research has shown that unvoiced parts of the speech signal can also be modeled by some sinusoids [8].

### 3.3 Improvement of Multiresolution Peak-Picking Algorithm

According to the discussions about voiced and unvoiced parts of speech signal, one can ask this question: can we treat equally with two parts of a speech signal that are originated with different excitation? In other words, regardless of the classification of the speech signal into voiced and unvoiced parts, can we use a constant process of peak selection for an inherently non-stationary signal? In this work, we found that by improving the method of peak-picking, better results of speech time stretching can be achieved. The presented idea was inspired by the work in [9], which states more sinusoids should be used to model noisy signals. We found that the number of peaks that selected for the modeling of unvoiced parts should be higher than the number of such peaks that needed for the modeling of the voiced parts.

Although source signal of the unvoiced parts is noise-like [8], unlike the treatment with noisy signals (fewer number of peaks should be considered for noisy signals [5], unvoiced part should be modeled with more sinusoids. For this reason, we propose a looser constraint for unvoiced parts respect to the voiced parts.

It should be mentioned that, we used Mel model for the channel division. This filter bank has logarithmic design so that it acts linearly in the frequencies below 1 KHZ and then acts as the following logarithmic formula.

$$\mathrm{Mel}(f)= 1127\log_e(1+f/700) \qquad (6)$$

*f* is in hertz and Mel (f) is the frequency in Mel. Using Mel criteria, new channel division for the voiced and unvoiced parts of the speech signal are proposed as shown in Table. 2 and Table. 3.

First, the limits of low frequency and high frequency for the channel division, which containing the most information, is transferred into the Mel domain. Then, considering the number of needed channels, bandwidth in the Mel domain is computed as:

$$\delta_f = (f_{mh} - f_{ml})/(n-1) \qquad (7)$$

Where n is the number of channels, $f_{ml}$ and $f_{mh}$ are the limits of

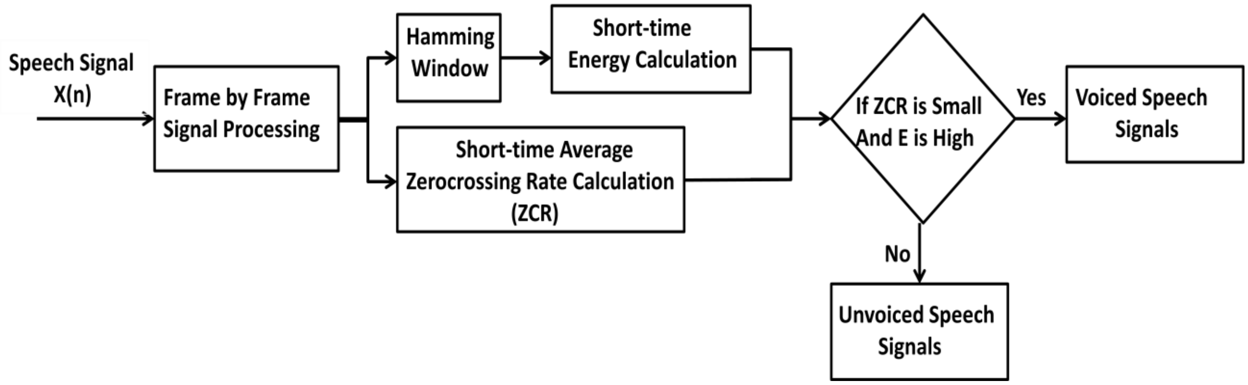

Figure 1: Blick diagram of the voiced/unvoiced classification

low frequency and high frequency in Mel domain respectively and $\delta_f$ is the bandwidth of the sub-bands of Mel domain. Frequency division is performed by dividing the frequency spectrum of Mel domain into uniformly distributed sub-band channels and then returning to Hertz domain. In our work, the limit of high frequency is considered 8 kHz and the limit of low frequency is considered 150Hz which they are experimentally obtained and can be changed depending on the type of the sound. This method of peak selection is named the non-stationary multi-resolution peak-picking (**NSMP**). In this work, we found that by assigning different channel divisions of the frequency domain for the voiced and unvoiced parts of the speech signal, better results for time-scale modification can be obtained. Simulation results indicate that this kind of channel division for the voiced and unvoiced parts improves the quality of the output sound according to PESQ, LLR and MOS criteria.

Table 2 : Channel division for voiced part

| Sub-band Index | Frequency Range (Hz) |
|---|---|
| 1 | 0 Hz – 148.4 Hz |
| 2 | 148.4 Hz – 483.3 Hz |
| 3 | 483.3 Hz – 950.2 Hz |
| 4 | 950.2 Hz – 1601.2 Hz |
| 5 | 1601.2 Hz – 2508.8 Hz |
| 6 | 2508.8 Hz – 3774.3 Hz |
| 7 | 3774.3 Hz – 5.5387 Hz |
| 8 | 5.5387 Hz – fs/2 Hz |

Table 3 : Channel division for unvoiced part

| Sub-band Index | Frequency Range (Hz) |
|---|---|
| 1 | 0 – 148.4 |
| 2 | 148.4 – 2018.3 |
| 3 | 2018.3 – fs/2 |

In these tables, fs is the sampling frequency.

### 3.4 Classification of voiced and unvoiced parts

The classification of the voiced and unvoiced parts is done during STFT operation. So that, after multiplying signal by a window, the resulting frame will be located in one of the voiced or unvoiced classes. Different approaches are presented for the division of speech to voiced and unvoiced parts. For example, a statistical approach has been presented in [10]. A classification network has been used in [11] for this purpose. An approach which is based on the Gaussian mixture model has been used by the authors in [12]. The approach which is used in this work is based on [13]. They combined zero crossings rate and energy calculation for voiced/unvoiced classification. The frames that deal with voiced and unvoiced parts are illustrated in Figure. 1. First, each frame is classified into voiced and unvoiced. As shown in the Figure. 1, with using two threshold values for the zero crossing and the amount of energy of the frame, we determine whether the frame is voiced or unvoiced.

## 4 Experimental Results

In this section we prove our assertion by presenting the results of the simulations. We used PESQ, LLR and MOS criteria for the quality measurement. PESQ (Perceptual Evaluation of Speech Quality) criterion has been presented in the recent years by which perceptual quality of the sound can be measured through the calculations. LLR (likelihood ratio) is a statistical test used to compare the fit of two models and is a criterion that objectively measures the quality of the output. Finally, MOS (mean opinion score) is a criterion that provides a numerical indication of the perceived quality of the output signal.

### 4.1 Using PESQ and LLR criteria for the quality comparison

Phase Vocoder is known as a high fidelity algorithm. This means that if an input signal got under process with the time scale factor of 1, the output will be very similar to the input [14]. The contribution of this work is the proposing a method for searching instantaneous frequencies. In other words, applied improvement is concerned with the method of modeling the signal. Hence, to inspect the quality of the proposed method, we measured the fidelity of the algorithm. With higher fidelity of the algorithm, the algorithm will operate better in modeling the signal and as a result, a better quality in time-scale modification of sound will be expected. For this purpose, for the scale factor of 1, the outputs of time stretches speech using NSMP and MP are obtained. Also, accordance with the PESQ and LLR criteria, the outputs are compared to the input signal. The higher value of PESQ and lower value of LLR mean the less distortion in the output signal. As a result, the algorithm for modeling the signal is more capable. The test was investigated on a few samples of the speech. Sounds are selected from standard database named noizeus. The results of this simulation are represented in Table. 4. The voiced and unvoiced parts are separated from each other per ThrE = 0.018 and ThrZ = 165. (ThrE and ThrZ are the thresholds related to the zero crossing and energy of the signal respectively). As could be seen in the table, the proposed method led to an average time scale factor of 0.03 according to PESQ criterion and time scale factor of 0.02 according to LLR criterion. It has been observed that with adjusting threshold value, these values are doubled. Please do not include footnotes in the abstract and avoid using a footnote in the first column of the article. This will cause it to appear of the affiliation box, making the layout look confusing.

Table 4: Channel division for voiced part comparison of the output of time scale modification using NSMP and MP Based on PESQ and LLR criteria

| | **TSM using NSMP** | | **TSM using MP** | |
|---|---|---|---|---|
| | LLR | PESQ | LLR | PESQ |
| Sp01.wav | 0.029 | 4.43 | 0.047 | 4.41 |
| Sp12.wav | 0.024 | 4.31 | 0.042 | 4.26 |
| Sp13.wav | 0.030 | 4.42 | 0.039 | 4.38 |
| Sp16.wav | 0.015 | 4.41 | 0.042 | 4.39 |
| Sp19.wav | 0.012 | 4.40 | 0.037 | 4.37 |
| Sp24.wav | 0.027 | 4.44 | 0.037 | 4.40 |

### 4.2 Using PESQ and LLR criteria for quality comparison

In the next experiment, the quality of the outputs based on MOS criterion for the phase vocoder using MRPP and NSMRPP are evaluated. We had a group of 30 adult humans between 15 to 30 years that compare the quality of the methods. For this purpose, 6 samples of FARSIDOT dataset were time stretched by 4 different stretching factors: 1.43, 1.2, 0.78, and 0.5. Each listener assigned a score between one and five points, five being the best. Finally, the mean score on a given stretch ratio is calculated. Each sample contains enough unvoiced parts that facilitate the judgment of the listeners. The results of the test are set in Table. 5. As shown in Table. 5, our improvement made a visible effect on the quality of the output sound.

Table 5: The Comparision of the output of time scale modification using NSMP and MP based on MOS criteria

| | Time scale factor (α) | | | |
|---|---|---|---|---|
| | 1.43 | 1.2 | 0.78 | 0.5 |
| **TSM using MP** | 3.10 | 3.90 | 4.57 | 3.43 |
| **TSM using NSMP** | 3.33 | 4.13 | 4.83 | 3.47 |

## 5 CONCLUSION

In this work, we have shown that if the frequency spectrum division in multi resolution peak picking assigned differently according to the voiced and unvoiced parts of the speech signal, better result of time scale modification can be achieved. Resolution rate for the voiced and unvoiced parts was performed according to the Mel scale. This method of the frequency spectrum division is different from previous methods that are performed based on the exponential model according to Mel-scale. Modified algorithm was tested and investigated

on different types of human voice. The simulation results indicate the remarkable increases in the output quality according to PESQ and LLR criteria.

## ACKNOWLEDGMENT

Database is available at:
http://www.utdallas.edu/_loizou/speech/noizeus/.

## REFERENCES

[1] Flanagan, J. L and Golden, R. M., 1966. Phase vocoder. The Bell System Technical Journal 45, 1493–1509

[2] Portnoff, M. R., 1976. Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(3), 243-248.

[3] Dolson, M., 1986.The phases vocoder: A tutorial. Computer Music Journal, 10,145-27.

[4] Laroche, J. and M. Dolson., 1999. Improved phase vocoder time-scale modification of audio. IEEE Transactions on Speech and Audio Processing, 7, 323–332.

[5] Karrer, T., Lee, E., Borchers, J., 2006. PhaVoRIT: A phase vocoder for real-time interactive time-stretching. International Computer Music Conference (ICMC) Proceedings, New Orleans, USA.

[6] Stevens, S. S., Volkman, J. and Newman, E., 1937. A scale for the measurement of the psychological magnitude of pitch. Journal of the Acoustical Society of America, 8 (3), 185-190

[7] Rabiner and Juang, 1993.Fundamentals of speech recognition, Prentice. Hall

[8] Macon, M. and Clements, M., 1997. Sinusoidal Modeling and Modification of Unvoiced Speech. IEEE Trans. on Speech and Audio, 557-560.

[9] Serra, X.; Smith, J. O., 1990. Spectral modeling synthesis: A sound analysis /synthesis system based on a deterministic plus stochastic decomposition. Computer Music Journal, 14(4), 12-24

[10] Al-Hashemy, B.A.R. and Taha, S.M.R. Voiced-unvoiced-silence classification of speech signals based on statistical approaches. Applied Acoustics, 25, 169-179.

[11] Qi, Y. and Hunt, B.R., 1993. Voiced-Unvoiced-Silence Classifications of Speech using Hybrid Features and a Network Classifier. IEEE Trans. Speech Audio Processing.1(2), 250-255.

[12] Jashmin K. Shah, Ananth N. Iyer, Brett Y. Smolenski, and Robert E. Yantorno. Robust voiced/unvoiced classification using novel features and Gaussian Mixture model. IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, 2004, pp. 17-21

[13] Bachu, R.G., Kopparthi, R.G., Adapa, B. and Barkana, B.D, 2010. Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and EnergyAdvanced Techniques in Computing Sciences and Software Engineering, 279-282..

[14] Sussman, R., and Laroche, J., 1999. Application of the phase vocoder to pitch-preserving synchronization of an audio stream to an external clock. In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20.

[15] Yi Hu, P. C. Loizou. 2008. evaluation of objective quality measures for speech enhancement, IEEE Transactions on In Audio, Speech, and Language Processing,16(1), pp229-238 .