DUAL PARAMETERS FOR VOICED-UNVOICED SPEECH SIGNAL DETERMINATION

Dhany Arifianto *

Department of Engineering Physics
Sepuluh Nopember Institute of Technology
Sukolilo Campus, Surabaya 60111, Indonesia
arifianto@ieee.org

ABSTRACT

This paper describes the application of the notion of instantaneous frequency amplitude spectrum (IFAS) to discriminate voiced and unvoiced segment of speech signal. The classification procedures of speech signal into voiced and unvoiced is determined by using harmonicity measure acquired after evaluating instantaneous frequency amplitude spectrum. For accuracy improvement, we use secondary parameter during transition from voiced-to-unvoiced and unvoiced-to-unvoiced to confirm the voiced area estimated by IFAS. Entropy of magnitude spectrum and instantaneous power are considered in this investigation. The performance of the dual method is compared to single thresholding using IFAS and also ESPS, AMDF and TEMPO to demonstrate its effectiveness.

Index Terms— voiced/unvoiced determination, instantaneous frequency, harmonicity measure, entropy, instantaneous power

1. INTRODUCTION

The knowledge of acoustical speech feature in particular voiced or unvoiced segment plays an important role in many speech analysis-synthesis systems. The nature of nonlinearity of speech signal in time and frequency domain makes the exact classification of voiced /unvoiced onset difficult. Numerous approaches have been proposed to address this problem, e.g. in [1].

In recent years, the notion of instantaneous frequency (IF) receives considerable attention for speech signal analysis. Abe, *et.al* [2], reported fundamental frequency estimation based on instantaneous frequency. The original IFAS-based which considered all IF band in the frequency selection is shown inaccurate for voiced /unvoiced detection.

In this paper, we extend our work previously reported in [3]. The voiced/unvoiced determination systems presented herein are based on instantaneous frequency amplitude spectrum (IFAS) to define *harmonicity measure*. The instantaneous frequency is derived from short-time Fourier transform of a signal as a function of time and frequency. The instantaneous frequency amplitude spectrum (IFAS) can represent the harmonic structure of speech signal better than the short time Fourier transform (STFT) amplitude spectrum. Threshold-based of voicing decision relies on difference of harmonicity measure of voiced compared to unvoiced segment.

The major problem is during transition segment between voiced and unvoiced, or vice versa. In this investigation, we used two well-known voice activity detectors, namely entropy of magnitude spectrum and instantaneous power of signal. The error rate can be reduced in the transition regions.

*most of the work was conducted at Tokyo Institute of Technology, Japan and Sepuluh Nopember Institute of Technology, Indonesia

2. INSTANTANEOUS FREQUENCY AMPLITUDE SPECTRUM

2.1. IFAS Derivation

For notation throughout the paper, let $x(t)$ be a function which represents speech signal and $X(\omega)$ be Fourier transform respectively. The STFT of $x(t)$ is rewritten in the form

$$X(\omega,t) = e^{-j\omega t}\int_{-\infty}^{\infty} w(\tau-t)x(\tau)e^{-j\omega(\tau-t)}\,d\tau \quad (1)$$

$$= e^{-j\omega t}G(\omega,t), \quad (2)$$

where $w(t)$ is an analysis window function. Without loss of generality, $w(t)$ is real and of finite duration. The instantaneous frequency estimate is given by the following formula

$$\lambda(\omega,t) = \frac{\partial}{\partial t}\arg[e^{j\omega t}X(\omega,t)] = \omega + \frac{\partial}{\partial t}\arg[X(\omega,t)]. \quad (3)$$

If the Fourier transform of $w(t)$ is a lowpass function, then $G(\omega,t)$ will be the output of a bandpass filter whose impulse response is $w(-t)e^{j\omega t}$ [4]. This bandpass filter has a frequency shifted version of the Fourier transform of $w(t)$ and its passband is centered at frequency ω. For the sake of simplicity, detail derivation can be referred to [5]. The following expression will be used to calculate instantaneous frequency

$$\frac{\partial}{\partial t}\arg[X(\omega,t)] = \frac{\mathrm{Re}[X]\frac{\partial X}{\partial t}(\mathrm{Im}[X]) - \mathrm{Im}[X]\frac{\partial}{\partial t}(\mathrm{Re}[X])}{|X|^2} \quad (4)$$

$$\frac{\partial}{\partial t}X(\omega,t) = \int_{-\infty}^{\infty} -\psi(\tau-t)e^{-j\omega\tau}x(\tau)\,d\tau, \quad (5)$$

where $\psi(t)$ is the derivative of analysis window $w(t)$ in STFT with respect to time. Using the equivalence of $|G(\omega,t)| = |X(\omega,t)|$, the instantaneous frequency amplitude spectrum (IFAS) at the instantaneous frequency is defined by [2]

$$S(\lambda_0,t) = \lim_{\Delta\lambda\to 0}\frac{1}{\Delta\lambda}\int_{\Omega_0}|G(\omega,t)|\,d\omega. \quad (6)$$

At particular time t, integral $|G(\omega,t)|$ on a set of intervals of the frequency is taken along the frequency axis ω such that $\Omega_0 = \{\omega|\lambda_0 \le \lambda(\omega,t) \le \lambda_0 + \Delta\lambda\}$.

The integral limit t_i is spanned across t_i^b to t_i^e, otherwise is nullified the existence of the sinusoidal components. Therefore, (1) can be rewritten into,

$$x_i(t) = \mathcal{R}[a_i(t)e^{(j\theta_i(t))}], \quad (7)$$

where $|a(t)|$ is also called the signal envelope. From the frequency modulated component $\theta_i(t)$, the instantaneous frequency is defined as derivative of phase with respect to time,

$$\phi_i(t) = \frac{d\theta_i(t)}{dt} \tag{8}$$

It should be noted that there is unlimited number of $a_i(t)$ and $\theta_i(t)$ combination which may generate a signal satisfying (3). However, the unique solution in order to fulfill (3) is obtained by using the so-called *analytic signal*, class of signals which satisfy Cauchy-Riemann conditions for differentiability. Let $z_i(t)$ denotes the analytic signal derived from the harmonic component $x_i(t)$,

$$X_i(\omega) = \int_{-\infty}^{\infty} x_i(t) e^{-j\omega t}\, dt, \tag{9}$$

$X_i(\omega)$ is Hilbert transform of $x_i(t)$. Consequently, spectrum of $Z_i(\omega) \geq 0$ is twice of $X_i(\omega)$, while in negative axis $X_i(\omega)$ is vanished. Hence, amplitude component and frequency component can be obtained to complete (3),

$$a_i(t) = |z_i(t)| \tag{10}$$

and accordingly, the frequency is,

$$\phi_i(t) = \frac{d}{dt} \arg[\theta_i(t)] \geq 0 \tag{11}$$

Note, $a_i(t)$ should be bounded and $\phi_i(t)$ should be bandlimited.

2.2. Harmonicity Measure

Let $S(\lambda)$ be the amplitude spectrum of instantaneous frequency from a signal at a fixed time t for notation simplicity. A transform of $S(\lambda)$ is defined as follows

$$\eta(F) = \alpha^{\frac{-\beta}{F}} \int_{\lambda_0}^{\lambda_1} S(\lambda)\Lambda(\lambda, F)\, d\lambda, \tag{12}$$

where α and β are real constants, and

$$\Lambda(\lambda, F) = \begin{cases} 0, & \lambda/F < \pi \\ \frac{1}{2}\big(\cos(\lambda/F) + 1\big), & \lambda/F \geq \pi. \end{cases} \tag{13}$$

If the signal is periodic and $S(\lambda)$ shows harmonic structure with a fundamental frequency of F_0, then $\eta(F)$ has local maxima at the frequencies $F = F_0/n, n = 1, 2, \ldots$. As a result, the value of $\eta(F)$ can be considered to be likelihood where the fundamental frequency of the signal will be F. In (12), the term $\alpha^{-\beta/F}$ works as a weighting constant to give priority to higher fundamental frequencies. The $[\lambda_0, \lambda_1]$ interval of the integral in (12) determines the range used for fundamental frequency estimation. It is important to note that the IFAS is not necessary to calculate the value of $\eta(F)$ because (12) can be expressed by the integral on ω axis of the form

$$\eta(F) = \alpha^{\frac{-\beta}{F}} \int_{\Omega} |X(\omega)|\Lambda(\lambda(\omega, t), F)\, d\omega, \tag{14}$$

where $\Omega = \{\omega | \lambda_0 \leq \lambda(\omega) \leq \lambda_1\}$.

For band selection based on harmonicity measure, suppose interval $[\lambda_0, \lambda_1]$ be on the IF axis. Let Ω be a set of intervals on the ω axis such that $\lambda_0 \leq \lambda(\omega) \leq \lambda_1$ and the measure $m(\Omega)$ exists in Lebesgue's sense.

$$\xi_{\lambda_0,\lambda_1}(F) = \frac{1}{m(\Omega)} \int_{\Omega} \mathcal{C}(\lambda(\omega), F)\, d\omega, \tag{15}$$

where $\Omega = \{\omega | \lambda_0 \leq \lambda(\omega) \leq \lambda_1\}$ and

$$\mathcal{C}(\lambda(\omega), F) = \begin{cases} 0, & \lambda(\omega)/F < \pi/2 \\ \cos(\lambda(\omega)/F), & \lambda(\omega)/F \geq \pi/2. \end{cases} \tag{16}$$

Harmonicity measure is defined as maximum value of $\xi_{\lambda_0,\lambda_1}(F)$ which is denoted by

$$P_{\lambda_0,\lambda_1} = \max_F\, \xi_{\lambda_0,\lambda_1}(F), \tag{17}$$

whose value spans somewhere between

$$-1 \leq \max_F\, \xi_{\lambda_0,\lambda_1}(F) \leq 1.$$

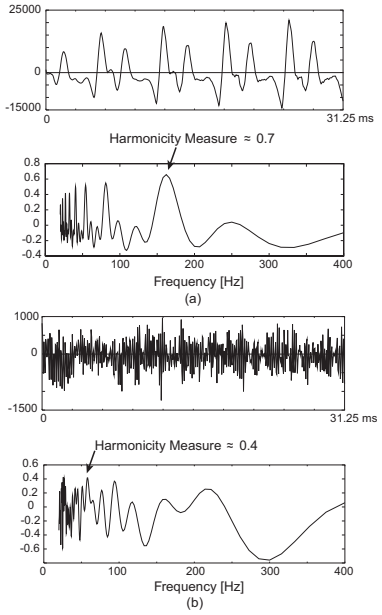

Fig. 1. Example of harmonicity measure $\xi(F)$ for (a) voiced speech and (b) unvoiced speech.

An example of the evaluation function $\xi(F)$ is shown in Fig. 1(a) for voiced whose harmonicity measure value is about 0.7 and 0.2 in Fig. 1(b) for unvoiced part, respectively.

3. APPLICATION

3.1. Voiced/Unvoiced Classification Algorithm

The algorithm of IFAS-based voiced/unvoiced decision can be summarized as follows,

1. Analyze the input signal $x(t)$ using STFT to obtain its spectrum $X(\omega)$.
2. Calculate the instantaneous frequency $\lambda(\omega)$ by using (4) and (5).
3. Select an IF band $[\lambda_0, \lambda_1]$ which maximizes the measure of harmonicity in the IF-domain P_{λ_0,λ_1} in (17).
4. Calculate the $\eta(F)$ of the selected IF band $[\lambda_0, \lambda_1]$ and determine $F = F_0$ which maximizes $\xi(F)$ in (15).
5. Determine a threshold of using techniques explained in Sec. 3.2 for voiced, otherwise marked as unvoiced segment.

6. Compare the IFAS-based estimated voiced region to secondary voice activity measure.

The STFT $X(\omega)$ and the instantaneous frequency $\lambda(\omega)$ are calculated on the frequency of $f_k = kF_s/N$. In the IF calculation, it sometimes occurs that the IF has a meaningless value which means the nonexistence of frequency component within the passband of the bandpass filters centered at each frequency bin. Consequently, if the value of the obtained IF $\lambda(f_k)$ at the n-th frequency bin (i.e n-th bandpass filter) does not exist, the value is excluded from the evaluation of $\xi_{\lambda_0,\lambda_1}(F)$ and $\eta(F)$.

3.2. Voicing Decision Strategy

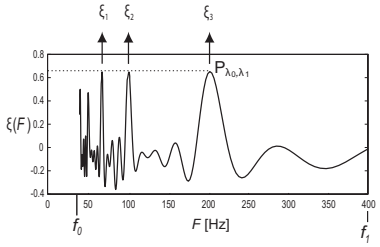

Fig. 2. Voicing decision strategies by using harmonicity measure

The voicing decision techniques is illustrated in Fig. 2. The main objective of thresholding techniques is twofold. Firstly, the voiced/unvoiced "switch" is decided by using the underlying clear structure of harmonicity measure. On the other hand, the harmonicity measure opens many alternatives towards thresholding techniques for voiced or unvoiced boundary marking.

The first strategy for voicing decision is by determining the value of harmonicity measure of each frame, P_{λ_0,λ_1} in (17), in one speech file. The threshold value is selected by examining the overall harmonicity measure to single out the highest possible value for unvoiced speech while otherwise, the value is classified into voiced. Such technique henceforth is called *direct thresholding*. We used variance$\xi_{\lambda_0,\lambda_1}(F)$ from frequency search range f_0 to f_1. The last technique is by ordering the peaks in $\xi_{\lambda_0,\lambda_1}(F)$, then three highest peaks, represented by ξ_1, ξ_2, ξ_3, respectively, of every frame are selected regardless of voiced or unvoiced. These three peaks are then summed to determine a threshold value. This technique is called *peak-picking*. The evaluation results of these thresholding techniques are shown in Table 2.

3.3. Secondary Parameter

Shen et al. [6] proposed an entropy-based parameter for speech detection under adverse conditions where voiced region has higher degree regularity than that of unvoiced. Due to Shannon, it originally measures the average length of bit code per symbol under optimal coding.

$$H(S) = -\sum_{i=1}^{N} P(s(i)) log_2(P(s(i))), \tag{18}$$

where $S = [s(1), \ldots, s(i), \ldots, s(N)]$ represents a source of N symbols, $P(s(i))$ is the probability of symbol i emission. In spectral energy domain

$$H(|X(\omega,t)|^2) = -\sum_{\omega=1}^{\Omega} P(|X(\omega,t)|^2) log_2(P(|X(\omega,t)|^2)), \tag{19}$$

where

$$P(|X(\omega,t)|^2) = \frac{|X(\omega,t)|^2}{\sum_{\omega=1}^{\Omega} |X(\omega,t)|^2}.$$

When the input is purely white noise, $H(|X(\omega,t)|^2)$ will be maximum ($H(X) = log_2(\Omega)$), and minimum ($H(X) = 0$) when it is pure tone.

The second choice is by using maximum of the squared envelope of bandpass filter banks output in each frame, [7]

$$M(\omega,t) = \max \|X(\omega,t)\|^2 \tag{20}$$

The spectrum of speech will have lower maxima in unvoiced segment than that of in voiced part.

4. RESULTS AND DISCUSSION

For experimental purpose, NAIST-CREST clean speech database which contains continuous speech and its corresponding Electroglottograph (EGG) waveforms uttered 84 sentences is incorporated for performance assessment. The whole experimental setup can be referred to [3].

Since window length choice affects the overall classifier performance, experimentally the appropriate window length is four or five times wider than pitch period. In this paper, we define voiced/unvoiced error as one error. If the reference says that the i-th frame is voiced while the i-th estimate is unvoiced or vice versa, it is calculated as one error. Compared to our previous work [3], where we only considered the error within voiced region. The thresholds value to begin voiced (or unvoiced) boundary was obtained by experiment empirically by assigning beforehand a value that gives optimal result for both male and female group.

Methods	window	VUV Error(%)	
		Male	Female
	Allband	6.5	6.4
	Limited	5.7	6.8
Variance	Selected I	5.6	6.9
	Selected II	5.1	6.1
	Selected III	8.9	7.9
	Allband	11.9	15.7
	Limited	8.5	9.9
Threshold	Selected I	8.3	10.4
	Selected II	6.8	5.8
	Selected III	10.0	8.4
	Allband	13.2	10.7
	Limited	5.6	6.3
Peak-Picking	Selected I	6.1	6.6
	Selected II	5.9	6.1
	Selected III	7.7	8.6

Table 1. V/UV Errors of IFAS-based Voiced/Unvoiced Determination

We investigate intrinsic properties of the proposed method with different options available and its accuracy under various conditions. We consider conditions indicated in Table 1,

1. *Allband*. In this case the full band is considered with the lower bound λ_l set to zero while the upper bound $\lambda_u/2\pi$ is 8 kHz.

2. *Limited*. We specify a narrow frequency range of λ_l and $\lambda_u/2\pi$ which are zero and 600 Hz, respectively.

3. *Selected I* means λ_l is zero and $\lambda_u/2\pi$ is moving starting from 600 Hz up to 2 kHz with 100 Hz increment.

4. *Selected II* for variable window length, F_0 candidates are taken from prior consecutive 7 frames with the lowest and the highest frame values elimination. Within these remaining 5 frames, pitch-lags are averaged then multiplied by 4 to provide a window length candidate. If this window length is lower than 400 samples length, the last is used instead.

5. *Selected III* case, we use 400, 450, 500, 600, 800, 1000 samples windows, then the window which maximizes the harmonicity measure value is selected.

The accuracy and reliability of voiced/unvoiced determination presented in Table 2 are solely based on harmonicity measure previously described, hereafter called IFAS-based method. With respect to window choice, *Selected II* gives the lowest error rates. The *Variance* shows the best results with about 5 % for male and 6 % for female speakers.

Methods	**V/UV Error(%)**			
	Male		**Female**	
	Entropy	**Max**	**Entropy**	**Max**
Mean	6.8	5.5	7.0	6.7
Variance	6.5	5.9	6.5	6.1
Direct	6.2	5.7	5.7	5.8
Peak-Picking	6.2	5.1	6.3	5.7

Table 2. V/UV error rates of dual parameters

The effect of using dual parameter is to reduce over-estimated voiced region. Then, we eliminated unvoiced region whose length is less than 10 frames and 13 frames for voiced. Table 2 shows the results of VUV error rate using IFAS-based and secondary parameter where we only consider *Selected II* case. Here, *Entropy* refers to entropy-based and *Max* is instantaneous power spectrum. It is shown that the accuracy for Direct and Pick-Picking are improved particularly using the *Max* method compared to their respective error rate in Table 2. Entropy method tends to over-estimate unvoiced region which consequently lower the original performance, i.e. IFAS-based.

For comparison, we used an open-source speech analysis tool called *Wavesurfer*[8] and speech analysis-synthesis suite written in Matlab called STRAIGHT-TEMPO [9] which are employed with minor modification. *Wavesurfer* uses ESPS-based pitch tracking using normalized cross correlation refined by dynamic programming and the average magnitude difference function (AMDF) [1].

It is clearly shown in Table 3 that IFAS-based (variance with *a priori* window option) and dual method using (Peak-Picking) V/UV classsification technique outperforms the performance of ESPS, AMDF and TEMPO methods for both speakers. The dual method (using the Peak-Picking and Instantaneous Power Spectrum) performs better for female case compared to IFAS.

Methods	**V/UV Error(%)**	
	Male	**Female**
IFAS	5.1	6.1
Dual	5.1	5.7
ESPS	7.4	8.3
AMDF	7.9	10.5
TEMPO	6.9	6.5

Table 3. V/UV error rates of IFAS, ESPS, AMDF, and TEMPO

5. CONCLUDING REMARKS

In this paper, the implementation of the notion of instantaneous frequency to discriminate the voiced and unvoiced segment of speech signal has been investigated, and several extensions to previous research were also presented. In overall, the IFAS-based V/UV classifier performs better to the male speaker group than that of the female speaker groups by error rate roughly about 5%. It is shown that dual parameter method improves *Direct* and *Peak-Picking* accuracy by reducing width of the voiced region.

The IFAS-based voiced-unvoiced classifier, as well as dual parameters voiced-unvoiced classifier, outperforms both ESPS, AMDF and TEMPO particularly in male speaker. By band selection and post-processing, the performance of V/UV discriminator can be further enhanced by lowering V/UV error rate for both male and female speakers. Using similar framework, this research is in progress to deal with embedded noisy speech signal to evaluate its robustness.

6. REFERENCES

[1] W. Hess, *Pitch Determination of Speech Signals*, Springer Verlag, Berlin, 1983.

[2] T. Abe, T. Kobayashi, and S. Imai, "Robust pitch estimation with harmonic enhancement in noisy environment based on instantaneous frequency," in *Proc. 4th ICSLP*, Philadelphia, USA, October 1996, pp. 1277–1280.

[3] D. Arifianto and T. Kobayashi, "Performance evaluation of ifas-based fundamental frequency estimator in noisy environments," in *Proc. EUROSPEECH '03*, Geneva, Switzerland, September 2003, pp. 2877–2880.

[4] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, New Jersey, USA, 1993.

[5] T. Tanaka, T. Kobayashi, D. Arifianto, and T. Masuko, "Fundamental frequency estimation based on instantaneous frequency amplitude spectrum," in *Proc. 2002 ICASSP*, Florida, USA, May 2002, pp. 329–332, IEEE.

[6] J.W. Hung J.L. Shen and L.S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *Proc. ICSLP-98*, 1998.

[7] O.D. Grace, "Instantaneous power spectra," *J. Acoust. Soc. Am.*, vol. 69, no. 1, pp. 191–198, January 1981.

[8] "http://www.speech.kth.se/wavesurfer/," .

[9] H. Kawahara, H. Katayose, A. de Cheveigne, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f_0 and periodicity," in *Proc. EUROSPEECH'99*, Budapest, Hungary, September 1999, pp. 2781–2784.