# IFAS-BASED VOICED/UNVOICED CLASSIFICATION OF SPEECH SIGNAL

*Dhany Arifianto, Takao Kobayashi*

Interdisciplinary Graduate School of Science ang Engineering
Tokyo Institute of Technology
4259 Nagatsuta, Midori-ku, Yokohama, Japan 226-8502
{dany.arifianto,takao.kobayashi}@ip.titech.ac.jp

**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. Several voicing decisions based on harmonicity measure are described to provide possible modifications including post-processing stage to improve the classifier accuracy. The performance of the proposed method is also compared against ESPS and AMDF to demonstrate the effectiveness of the proposed method.

## 1. INTRODUCTION

The knowledge of acoustical speech feature in particular voiced or unvoiced segment plays important role in many speech analysis-synthesis systems. A significant amount of research has been conducted on finding reliable and accurate voicing determination in the past recent decades. Despite numerous approaches have been proposed to addresss this problem which can be traced back in the standard literature by Hess [3] including cited references therein, it remains an active research area because of its difficulties dealing with nonstationary properties of speech signal.

In recent years, the notion of instantaneous frequency (IF) receives considerable attention for speech signal analysis. Abe, *et.al* [1], [2], reported fundamental frequency estimation based on instantaneous frequency. In [4], it is also reported the use of instantaneous frequency to estimate fundamental frequency with slight modification. However, both authors do not explicitly describe to perform voiced/unvoiced classification in those papers.

In this paper, a new feature is added to IFAS-based fundamental frequency estimation previously reported in [7]. The development of voiced/unvoiced determination systems presented herein is based on instantaneous frequency amplitude spectrum (IFAS) to define *harmonicity measure*. By means of the clear underlying structure of harmonicity measure, the voiced speech boundary can be discriminated accurately.

Threshold-based of voicing decision relies on perspicuous difference of harmonicity measure of voiced against unvoiced segment respectively. The output of voiced/unvoiced determination block was refined in the post-processing part to remove the artifacts that may exist for example in the transition segment between voiced and unvoiced, or vice versa. 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 for period boundary (epoch). For the reason that the instantaneous frequency is a local variable at specific time, more accurate and reliable voiced segment discrimination can be obtained. The periodic structure of speech, called harmonicity measure, is applied to discriminate between voiced and unvoiced regions by detecting specified threshold value exploited from harmonicity measure.

## 2. INSTANTANEOUS FREQUENCY AMPLITUDE SPECTRUM

### 2.1. IFAS Derivation

Speech signals can be considered as an additive mixture of periodic and/or quasiperiodic signals modulated in both amplitude and frequency. 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 $w(-t)e^{j\omega t}$ [8]. This bandpass filter has a frequency shifted version of the Fourier transform of $w(t)$ and its passband is centered at frequency $\omega$. For the sake of simplicity, detail derivation can be referred to [7]. 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]

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

At particular time $t$, integral $|G(\omega, t)|$ on a set of intervals of the frequency is taken along the frequency axis $\omega$ such that $\Omega_0 = \{\omega | \lambda_0 \le \lambda(\omega, t) \le \lambda_0 + \Delta\lambda\}$ in Lebesgue's sense.

### 2.2. Harmonicity Measure

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

$$\eta(F) = \alpha^{\frac{-\beta}{F}} \int_{\lambda_0}^{\lambda_1} S(\lambda)\Lambda(\lambda, F)\, d\lambda, \tag{7}$$

where $\alpha$ and $\beta$ are real constants, and

$$\Lambda(\lambda, F) = \begin{cases} 0, & \lambda/F < \pi \\ \frac{1}{2}\big(\cos(\lambda/F) + 1\big), & \lambda/F \ge \pi. \end{cases} \tag{8}$$

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 (7), 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 (7) 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 (7) can be expressed by the integral on $\omega$ axis of the form

$$\eta(F) = \alpha^{\frac{-\beta}{F}} \int_{\Omega} |X(\omega)|\Lambda(\lambda(\omega, t), F)\, d\omega, \tag{9}$$

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

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

$$\xi_{\lambda_0,\lambda_1}(F) = \frac{1}{m(\Omega)} \int_{\Omega} \mathcal{C}(\lambda(\omega), F)\, d\omega, \tag{10}$$

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

$$\mathcal{C}(\lambda(\omega), F) = \begin{cases} 0, & \lambda(\omega)/F < \pi/2 \\ \cos(\lambda(\omega)/F), & \lambda(\omega)/F \ge \pi/2. \end{cases} \tag{11}$$

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

$$P_{\lambda_0,\lambda_1} = \max_F\ \xi_{\lambda_0,\lambda_1}(F), \tag{12}$$

whose value spans somewhere between

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

An example of the evaluation function $\xi(F)$ is shown in Fig. 1. We calculated the STFT using a 500-point Blackman window

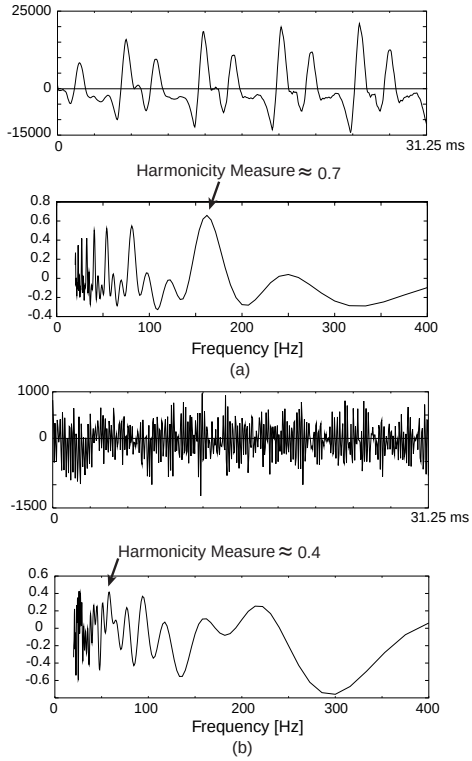


**Fig. 1**. Example of harmonicity measure $\xi(F)$ for (a) voiced speech and (b) unvoiced speech.

(31.25 ms at 16 Hz sampling) and 1024-point DFTs. The evaluation of both $|X(\omega)|$ and $S(\lambda)$ is carried out at only uniformly spaced frequency points. The input signal shown in Fig. 1(a) is a portion of vowel /i/ uttered by a male speaker. The fundamental frequency of the signal located at about 160 Hz whose harmonicity measure value is about 0.7. Fig. 1(a) depicts the corresponding $\xi(F)$ with condition of $[\lambda_0, \lambda_1] = [0, 1]$ kHz. It is apparent that $\xi(F)$ has a maximum value when $F$ is equal to the fundamental frequency of the signal. Similar to Fig. 1(a), Fig. 1(b) depicts for noise-like speech waveform. Contrasting to Fig. 1(a), it is clear that the harmonicity measure or other harmonics part can not be discriminated since the peaks are almost in the same height whose harmonicity measure value is lower than voiced case.

## 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 $P_{\lambda_0,\lambda_1}$ in (12).

4. Calculate the $\eta(F)$ of the selected IF band $[\lambda_0, \lambda_1]$ and determine $F = F_0$ which maximizes $\xi(F)$ in (10).
5. Determine a threshold value for voiced boundary, otherwise marked as unvoiced segment.

Last step is applicable when voiced/unvoiced decision is required since the algorithm implementation allows ones to select only for fundamental frequency estimation or only voiced/unvoiced classification or both processes simultaneously with slight modification. 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

The voicing decision is taken by relying on the harmonicity measure as discussed in section 2.2. In the following, several thresholding methods will be described with harmonicity measure as its stand-point. In fact, there is a significant value difference in the voiced frame and unvoiced frame, as well as in the frame which analyze transition region from voice to unvoice and vice versa. The threshold value can be determined by examining value of the harmonicity measure of male and female speaker respectively.

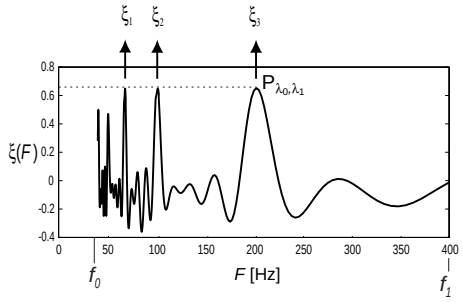


**Fig. 2**. Voicing decision strategies by using harmonicity measure

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 value of harmonicity measure of each frame, $P_{\lambda_0,\lambda_1}$ in (12), 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*. In the second and third scheme, mean and variance are employed to yield threshold value by means of numerical value of each sample of $\xi_{\lambda_0,\lambda_1}(F)$ from frequency search range $f_0$ to $f_1$. In Fig. 2, the dashed line illustrates the case of thresholding by using mean where in fact the position may vary than shown. Third method is by ordering the peaks in $\xi_{\lambda_0,\lambda_1}(F)$, then three highest peaks, represented by $\xi_1, \xi_2, \xi_3$, respectively, of every frame are selected regardless voiced or unvoiced. These three peaks are then summed to determine a threshold value. This technique is called *peak-picking*.

Only in the direct thresholding, the threshold value is gender dependent that we should set by experimentally adjusting the value between male and female speaker. Thus, it is difficult to find threshold value that maximazes the classification accuracy for both. By examining one arbitrary selected speaker, threshold value in the other methods is relatively easy to determine by setting one particular threshold value once. It is not only gender independent but also gives minimum voicing determination error rate for all speakers. The evaluation results of these thresholding techniques are shown in Table 2.

## 4. RESULTS AND DISCUSSION

For experimental purpose, NAIST-CREST clean speech database which contains continuous speech and its coresponding Electroglottograph (EGG) waveforms uttered by 14 male and 14 female speakers is incorporated for performance assessment. We selected three Japanese sentences from the database for evaluation, 84 sentences in total. The whole experimental setup can be referred to Table 1.

| Database | NAIST-CREST with EGG |
|---|---|
| True Pitch & Voiced/Unvoiced | Manual |
| Speaker | 14 male and 14 female |
| Sentences | Three sentences each speaker |
| Sampling Frequency | 16 kHz |
| Analysis Window type | Blackman |
| Window Shift | 1 msec |
| $F_0$ search range | 40 - 400 Hz |
| $\alpha$,$\beta$ from eq.(9) | 10, 8 Hz |

**Table 1**. Summary of experimental parameters

The accuracy and reliability of voicing decision are influenced by many factors among which are the proper detection of the onset and the end of voiced/unvoiced segment. The accuracy and reliability of voiced/unvoiced determination presented herein are mainly based on harmonicity measure previously described. Since window length choice affects the overall classifier performance, experimentally the appropriate window length is four or five times wider than pitch period. For further refining the estimated voiced / unvoiced boundaries, post-processing stage is performed by using what so-called pitch-continuity tracking, suggested in [5], to eliminate possible discontinuity may occur in-between voiced or unvoiced region.

In Table 2, *LimitedBand* which means the IF band $[\lambda_0, \lambda_1]$ is prescribed prior to process to single fixed value which in this case 600 Hz is set as $\lambda_1$ while $\lambda_0$ is zero. On the other hand, *AllBand* scheme means IF band $\lambda_1$ was assigned to 8 kHz while $\lambda_0$ is fixed to zero. Either in $F_0$ estimation or voiced/unvoiced determination, *LimitedBand* case exceeds the performance of *AllBand* as the effect of narrower analysis filter bandwidth for all scheme of experiments. Direct thresholding has lower accuracy since the threshold value is set for both male and female. It is well-known that fundamental frequency of female is higher than male which implies fix threshold value may fit to one group concurrently inaccurate for other group. It can be noticed that both statistical

| Methods | V/UV Error(%) | | | |
|---|---|---|---|---|
| | AllBand | | LimitedBand | |
| | Male | Female | Male | Female |
| Mean | 4.174 | 3.538 | 2.810 | 2.393 |
| Variance | 4.082 | 3.677 | 3.518 | 3.046 |
| Direct Thresholding | 5.322 | 8.447 | 4.450 | 7.640 |
| Peak-picking | 3.009 | 2.245 | 1.426 | 1.102 |

**Table 2**. V/UV Errors of IFAS-based Voiced/Unvoiced Determination

measures performs better than direct thresholding specifically in *LimitedBand* case. The peak-picking technique is superior to that of other thresholding techniques, the three-peak value of unvoiced frame may not surpass the voiced frame including when the frame falls in transition region. However, it is slightly more complex procedure than direct thresholding since the harmonicity value should be rearranged in either ascending or descending order.

The performance evaluation is conducted at every 1 ms of the speech signal for all techniques. An open-source speech analysis tool called *Wavesurfer*[6] is employed for comparison after window shift adjusted to 1 ms instead of 10 ms. *Wavesurfer* uses ESPS-based pitch tracking using normalized cross correlation refined by dynamic programming and the other method is AMDF which stands for average magnitude difference function [3].

| Methods | V/UV Error(%) | |
|---|---|---|
| | Male | Female |
| IFAS | 1.426 | 1.102 |
| ESPS | 7.860 | 6.631 |
| AMDF | 5.233 | 5.402 |

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

It is clearly shown in Table 3 that IFAS-based V/UV classsification technique outperforms the performance of ESPS and AMDF methods. The IFAS-based boundary detector performance has been adjusted in post-processing stage by using pitch continuity tracking. However, the V/UV decision of ESPS and AMDF is merely relied on the existence of tracked $F_0$ without further modification to enhance its potential performance respectively.

| Methods | GPE(%) | |
|---|---|---|
| | Male | Female |
| IFAS | 1.485 | 1.761 |
| ESPS | 9.146 | 5.888 |
| AMDF | 11.412 | 5.039 |

**Table 4**. Gross pitch error rates of IFAS, ESPS, and AMDF

In addition to V/UV classifier performance benchmarking, we also compare the performance of $F_0$ estimation. Each point at the existence of $F_0$ reference was compared to the evaluated method. In IFAS-based, *LimitedBand* option, 600 Hz cut-off frequency is used without any post-processing method to smooth pitch contour. As can be seen in Table 4, the IFAS-based $F_0$ estimation performs better for both male and female case than ESPS or AMDF. However, the wavesurfer is used without any modification or post-processing which may lower its estimation accuracy. The IFAS-based $F_0$ estimator generally performs better to male than female. On the contrary, better estimation accuracies in female group are resulted in ESPS and AMDF.

## 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. Several harmonicity-measure-based voicing decisions were described to improve the accuracy. Furthermore, post-processing what so-called pitch continuity tracking was utilized to boost overall performance of the voicing discriminator. In overall, the IFAS-based V/UV classifier performs better to the female speaker group than that of the male speaker groups by error rate roughly about $2\%$. 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. The results of the proposed technique was contrasted against ESPS, and AMDF via Wavesurfer. The IFAS-based voiced-unvoiced classifier, as well as IFAS-based fundamental frequency estimator, outperforms both ESPS and AMDF. 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 in adverse environment.

## 6. REFERENCES


[1] T. Abe, T. Kobayashi, and S. Imai, "Harmonics Estimation Based on Instantaneous Frequency and its Application to Pitch Determination of Speech", *IEICE Trans., Information and Systems*, vol.E78-D, No.9, pp. 1188-1194, September 1995.

[2] T. Abe, T. Kobayashi, and S. Imai, "Robust Pitch Estimation with Harmonic Enhancement in Noisy Environment Based on Instantaneous Frequency," *Proc. 4th ICSLP*, pp.1277-1280, Philadelphia, USA, Oct. 1996.

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

[4] H. Kawahara, H. Katayose, A. de Cheveigne, R. D. Patterson, "Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity", *Proc. EUROSPEECH'99*, Vol. 6, pp. 2781-2784, 1999.

[5] D.J. Liu and C.T. Lin, Fundamental Frequency Estimation Based on the Joint Time-Frequency Analysis of Harmonic Spectral Structure, *IEEE Trans., Speech and Audio Proc.*, vol. 9, no. 6, pp. 609-621, September 2001.

[6] `http://www.speech.kth.se/wavesurfer/`

[7] T. Tanaka, T. Kobayashi, D. Arifianto, T. Masuko, "Fundamental Frequency Estimation Based on Instantaneous Frequency Amplitude Spectrum", *Proc. ICASSP*, vol-I, pp.329-332, Orlando, Fl., May 2002.

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