VOICED/UNVOICED DETERMINATION OF SPEECH SIGNAL IN NOISY ENVIRONMENT USING HARMONICITY MEASURE BASED ON INSTANTANEOUS FREQUENCY

Dhany Arifianto, Takao Kobayashi

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

ABSTRACT

This paper presents a voiced/unvoiced determination algorithm using instantaneous frequency amplitude spectrum (IFAS) in adverse environment. The proposed algorithm measures the degree of periodicity of speech signal, defined as harmonicity measure, where the difference between voiced part and unvoiced speech can be quantitatively obtained. We describe a new technique for voicing decision using IFAS-based F_0 evaluation function with variable window length and IF band selection. The proposed technique is evaluated with speech signal corrupted by additive white Gaussian, pink, and traffic noises. The results show that the proposed method outperforms ESPS, AMDF and TEMPO for both female and male speakers in all simulated conditions.

1. INTRODUCTION

Voiced/unvoiced determination is an essential technique in many applications of speech processing, such as speech coding, speech synthesis, and speech enhancement. A significant amount of research has been conducted on finding reliable and accurate voicing determination in the past recent decades (see in a recent review reported in [1]).

Recently, the notion of instantaneous frequency (IF) has been found to be attractive for speech signal analysis. Abe, *et.al* [2], reported a fundamental frequency (F_0) estimation method based on instantaneous frequency. In [3], it is reported the use of instantaneous frequency to estimate F_0 with modification of [2] in weighting procedure and post-processing stage for F_0 refinement. We also proposed an improved F_0 estimation method based on instantaneous frequency amplitude spectrum (IFAS) with introducing an idea of harmonicity measure [4] and showed its robustness in noisy environment [5]. Furthermore, it was demonstrated that the IFAS-based approach with the harmonicity measure is a potential method to address the problem in voiced/unvoiced determination of speech signals [6].

In this paper, we refine our previous work reported in [6] and show the robustness of the proposed technique in the presence of background noise. The voiced/unvoiced determination is conducted in two steps. Rough estimates are obtained using F_0 contour continuity information [7]. The F_0 estimate in each analysis frame is obtained using the IFAS-based technique [4]. The key idea of this technique is the use of the harmonicity measure which provides quantitative degree of regularity of periodicity. Then, another voicing decision is made by using an IFAS-based F_0 evaluation function with a prescribed threshold. This two-step algorithm consequently refines rough estimates in the first step by removing the artifacts that may exist in the transition segment between voiced and unvoiced regions.

The IFAS and harmonicity measure is revisited briefly in the second section, then followed by the algorithm of the voicing decision based on the IFAS-based F_0 evaluation function. To demonstrate its effectiveness, the proposed method is evaluated with several experimental conditions and its performance comparison are discussed respectively.

2. IFAS AND HARMONICITY MEASURE

Let $x(t)$ and $X(\omega, t)$ be a function which represents speech signal and its short-time Fourier transform (STFT), respectively.

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

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

where $w(t)$ is an analysis window function. 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 $w(-t)e^{j\omega t}$ [9].

The instantaneous frequency at frequency ω and at instant time t is defined by

$$\lambda(\omega,t) = \frac{\partial}{\partial t}\arg[G(\omega,t)]$$

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

The following expression will be used to calculate instantaneous frequency

$$\frac{\partial}{\partial t}\arg[X(\omega,t)] = \frac{a\frac{\partial b}{\partial t} - b\frac{\partial a}{\partial t}}{a^2+b^2}, \quad (4)$$

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

where $X(\omega,t) = a + jb$ and $\psi(t)$ is the derivative of analysis window $w(t)$ with respect to time.

In the following, it is considered that all derivations are at instant t, and t will be omitted for notation simplicity. Let $S(\lambda_0)$ be the IFAS at the instantaneous frequency λ_0 defined by the following equation [2]

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

where $\Omega_0 = \{\omega | \lambda_0 \leq \lambda(\omega, t) \leq \lambda_0 + \Delta\lambda\}$.

Then F_0 estimate is given by the value of F that maximizes the following function

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

where α and β are real constants and

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

In (7), λ_l and λ_u are lower and upper bounds of IF band respectively, and the term $\alpha^{-\beta/F}$ is a weighting constant to give priority to higher fundamental frequencies. Consider an interval $[\lambda_l, \lambda_u]$ on the IF axis λ, and let Ω be a set of intervals on the frequency axis such that $\lambda_l \leq \lambda(\omega) \leq \lambda_u$. A harmonicity evaluation function is defined as follows

$$\xi_{\lambda_l,\lambda_u}(F) = \frac{1}{m(\Omega)} \int_{\Omega} C(\lambda(\omega), F)\, d\omega, \tag{9}$$

where $m(\Omega)$ be the measure of Ω in Lebesgue's sense, i.e., the total length of intervals, and

$$C(\lambda(\omega), F) = \begin{cases} 0, & \lambda(\omega)/F < \pi/2 \\ \cos(\lambda(\omega)/F), & \lambda(\omega)/F \geq \pi/2. \end{cases} \tag{10}$$

We define harmonicity measure in the instantaneous frequency domain [4] by

$$P_{\lambda_l,\lambda_u} = \max_F \xi_{\lambda_l,\lambda_u}(F). \tag{11}$$

The harmonicity measure lies somewhere between

$$-1 \leq P_{\lambda_l,\lambda_u} \leq 1. \tag{12}$$

If the harmonic structure is perfect, that is, the Fourier spectrum of the signal has only F_0 and its multiple components, then P_{λ_l,λ_u} becomes unity. On the other hand, if the harmonic structure is not clear, P_{λ_l,λ_u} is about zero.

3. ALGORITHM

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 (3) - (5).
3. Select an IF band $[\lambda_l, \lambda_u]$ which maximizes the harmonicity measure in the IF-domain P_{λ_l,λ_u} of (11).
4. Calculate the IFAS-based F_0 evaluation function $\eta(F)$ of the selected IF band $[\lambda_l, \lambda_u]$ and determine $F_0 = F$ which maximizes $\eta(F)$ in (7).
5. Compare the value of $\eta(F_0)$ with a threshold value and mark the frame voiced if it exceeds the threshold, otherwise unvoiced (in detail, see **3.2**).

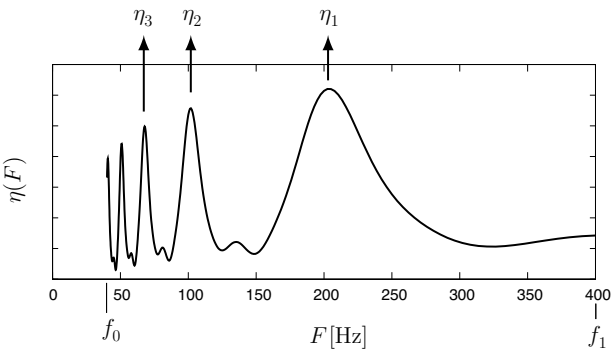

Fig. 1. V/UV determination strategy by using evaluation function $\eta(F)$.

The STFT $X(\omega)$ and the instantaneous frequency $\lambda(\omega)$ are calculated on the frequency of $f_k = kF_s/N$, where N is the window length and k is frequency bin index. 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 k-th frequency bin (i.e k-th bandpass filter) does not exist, the value is excluded from the evaluation of $\xi_{\lambda_l,\lambda_u}(F)$ and $\eta(F)$. In addition, step 3 becomes a maximization problem with respect to F, λ_u, and λ_l. We simplify this problem by fixing λ_l to a prescribed value and restricting λ_u to a finite set of frequencies.

We use the variable window length analysis proposed in [4]. F_0 candidates are taken from seven prior consecutive frames with the lowest and the highest values eliminated. Within these remaining five frames, pitch-lags are averaged then multiplied by four to provide a window length candidate. If resulting window length is lower than 400 samples in length, 400-point window length is used instead. This will enhance the accuracy and reliability of voicing decision since the window length will be adapted according to the input properties (periodic or non-periodic). We prescribed that λ_l is zero and $\lambda_u/2\pi$ is shifted from 600 Hz up to 2 kHz with every 100 Hz increments.

3.2. Voicing Decision Strategy

Voiced/Unvoiced determination part consists of two steps. A preprocessing stage is performed by using what so-called pitch continuity tracking, suggested in [7], to roughly estimate the voiced and unvoiced regions. In [4], we showed that our proposed technique can estimate F_0 contour smoothly without any doubling or halving. In the algorithm, firstly, the continuity of the F_0 contour is used to pre-determine voiced/unvoiced region. If the differences between current F_0 ($F_0[i]$) and those at the previous and next frames, $F_0[i-1]$ and $F_0[i+1]$, where i is frame index, are less than 15%, then it is considered to be continuous curve (i.e., voiced region). This step also consequently removes possible discontinuity which may occur in-between voiced and unvoiced regions. Indeed, there exists continuity in the 'true' unvoiced region less than five consecutive frames that will be eliminated in the second stage.

From our direct observation, the voiced region using pitch

continuity tends to be larger than the V/UV reference and suffers from many small voiced region in unvoiced area. Therefore, it is necessary to introduce the second step for refinement. We determine voiced/unvoiced region based on the IFAS-based F_0 evaluation function by thresholding. This is conducted by ordering the peaks in $\eta(F)$ from frequency search range f_0 to f_1, then the three highest peaks, represented by η_1, η_2, and η_3 as shown in Fig. 1, at every frame are selected regardless voiced or unvoiced. These three peaks are summed up as

$$\eta_p[i] = \eta_1[i] + \eta_2[i] + \eta_3[i]. \quad (13)$$

If the value of $\eta_p[i]$ is larger than a predetermined threshold, then it is classified into voiced, otherwise unvoiced.

In order to eliminate the requirement to adjust threshold in every environmental condition, it was assumed that the first 15-frame period of input was non-speech region. We determine the largest value of $\eta_p[i]$ during these 15 frames, then it is set as the threshold.

Finally, voiced regions obtained in the second step are compared to the voiced regions obtained in the first step. If the region formed by the $\eta(F)$ is unvoiced then the final decision is unvoiced. Similarly, if the region determined by F_0 contour tracking is unvoiced then the voiced region classified in the second step is flipped into unvoiced.

4. EXPERIMENTS

4.1. Experimental Condition

NAIST-CREST clean speech database which contains continuous speech and its corresponding Electroglottograph (EGG) waveforms uttered by 14 male and 14 female speakers was incorporated for performance assessment. We randomly selected three Japanese sentences of each speaker from the database for evaluation, 84 sentences in total. The VUV reference was developed automatically then corrected by hand and eye-inspection. The input signal was sampled with 16 kHz then analyzed by using Blackman window shifted in every 1 ms. The constants α and β in (7) were set to 10 and 8 Hz, respectively.

The signal to noise ratio (SNR) was set spanned from clean to 0 dB. This was done by adding generated white noise from Matlab, pink noise obtained from the Signal Processing Information Base (SPIB) [11] and traffic noise from the JEIDA Noise Database[8]. The traffic noise was recorded at the Hachiko-crossing located in Shibuya-ward, Tokyo, a popular spot crowded with people, passing cars, and advertising sounds. For evaluation rule, if one frame in the VUV reference is voiced while the output of the corresponding frame is unvoiced (or vice-versa), then it is counted as one error.

For performance comparison, we used the latest version of an open-source speech analysis tool called *Wavesurfer* [10] after window shift adjusted to 1 ms instead of 10 ms and a MATLAB based software called STRAIGHT-TEMPO (hereafter called TEMPO) [3] with adjustment to the frequency sampling from 20 kHz to 16 kHz. *Wavesurfer* used ESPS-based pitch tracking using normalized cross correlation refined by dynamic programming and the other method was AMDF which stands for average magnitude difference function.

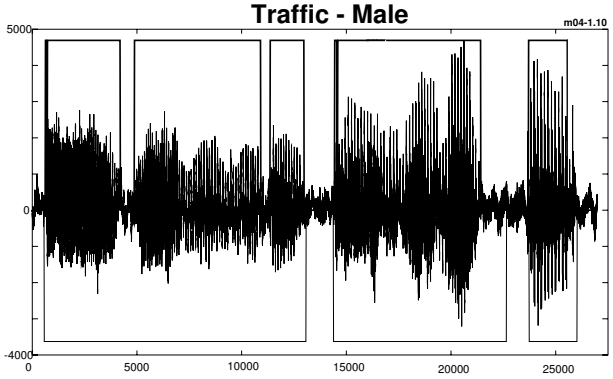

Fig. 2. Example of two-step V/UV classification.

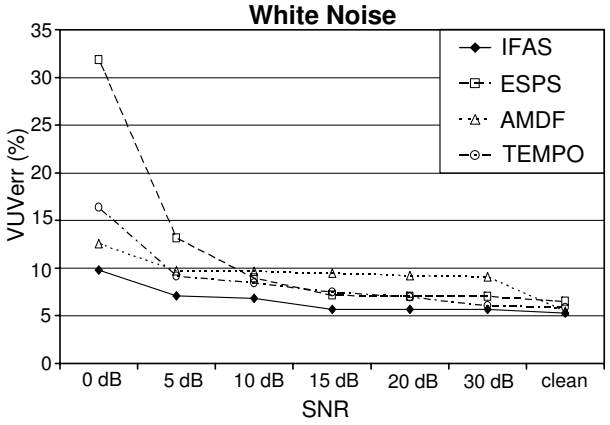

Fig. 3. Performance comparison in white noise condition.

4.2. Results

Fig. 2 illustrates the visualization of the two-step procedure aforementioned in **3.2** of a sentence in the database uttered by a male speaker. The environmental condition was set to 10 dB SNR in traffic noise. The horizontal axis shows sample index. For direct comparison, the first step rough voiced/unvoiced region estimates are located at the lower part of the figure. The upper (thicker line) part is the output of the second step as the final result. It is obviously shown that V/UV determination using the F_0 contour tracking tend to enlarge the voiced region. Moreover, there exists continuity in the non-voiced region. The IFAS-based F_0 evaluation function based procedure can reduce the falsely determined region as voiced, while the first step is necessary to reduce the discontinuities region in-between voiced region. However, it may obvious that the first step failed to remove completely some discontinuities particularly in the very adverse environment. For the second step, falsely classified regions particularly in the transition from unvoiced to voiced can be seen in the upper part (shown in darker region).

Figs. 3, 4, and 5 show the overall performance of V/UV error rate for both male and female speakers. The first comparison results is in white noise shown in Fig. 3. As can be seen, IFAS-

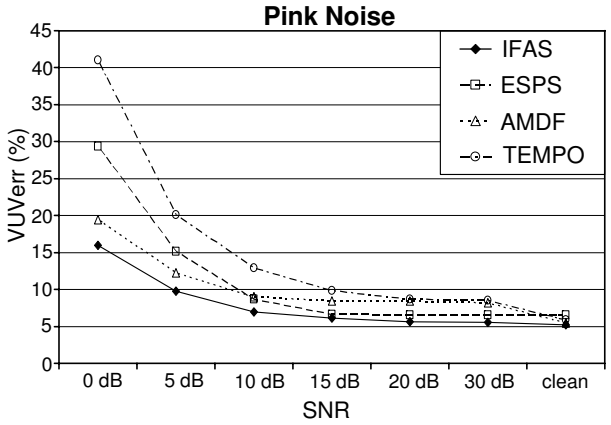

Fig. 4. Performance comparison in pink noise condition.

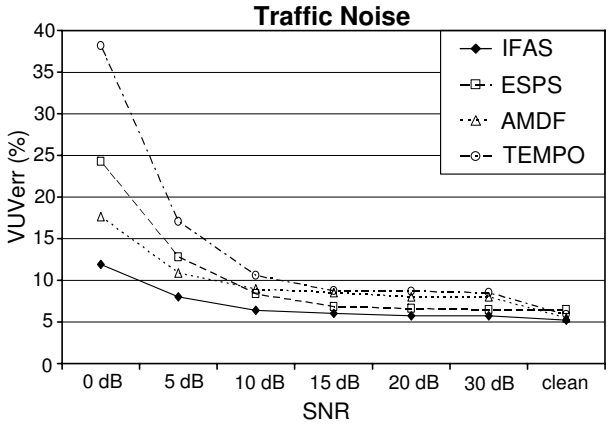

Fig. 5. Performance comparison in traffic noise condition.

based method has the lowest error rate from clean up to 0 dB less than 10%. TEMPO performance is close to IFAS but only in mild noise condition. Fig. 4 shows the evaluation results in the presence of pink noise. The proposed method maintained its performance with low error rates. The IFAS error rate is about 15% at 0 dB. Unlike in white noise, the error rates tend to enlarge rapidly. In real traffic noisy environment, the IFAS shows the best performance in all conditions as depicted in Fig. 5. The error rate is about 12% at 0 dB. The overall error rates of all methods in the traffic noise are lower than that of in the pink noise.

In general, the proposed method performs satisfactorily where from clean speech up to 10 dB SNR, the error rate remains almost the same. We also found that female speakers group has higher error rates than that of male speaker group in all cases. From clean to 0 dB, the error rate enlargement in male case is insignificant with respect to female case. ESPS performance is slightly lower than IFAS from clean to 10 dB while AMDF performance is better than ESPS and TEMPO from 5 dB to 0 dB. From 5 dB to 0 dB, ESPS performance is the lowest in white noise, while TEMPO has the largest error rates in pink and traffic noises.

5. CONCLUDING REMARKS

In this paper, a robust voiced/unvoiced determination algorithm in adverse environment has been investigated. In the algorithm, voiced/unvoiced region is firstly pre-determined by using F_0 contour continuity tracking. Then, these rough estimates are refined with a technique using IFAS-based F_0 evaluation function as the second step. The performance of the proposed technique were compared against ESPS and AMDF via Wavesurfer, and TEMPO. The IFAS-based voiced/unvoiced determination method outperformed ESPS, AMDF and TEMPO in white, pink and traffic noises ranged from clean to 0 dB for both female and male groups. For future, we will be working on the implementation of IFAS-based F_0 estimator and voicing decision for multipitch tracking. Using other database is also a possible future direction since Japanese sentence does not contain many fricatives or plosives as English. This would give another insight to deal with transition from voiced to unvoiced and vice versa.

6. ACKNOWLEDGMENTS

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research (B) 15300055.

7. REFERENCES

[1] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech and Audio Processing*, vol. 8, no.4, pp. 478-482, July 2000.

[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, Oct. 1996.

[3] 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, Budapest, Sept. 1999.

[4] 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.

[5] D. Arifianto and T. Kobayashi, "Performance evaluation of IFAS-based fundamental frequency estimator in noisy environments," *Proc. EUROSPEECH '03*, vol.IV, pp.2877-2880, Geneva, Sept. 2003.

[6] D. Arifianto and T. Kobayashi, "IFAS-based voiced / unvoiced classification of speech signal," *Proc. ICASSP*, vol.I, pp.812-815, Hong Kong, April 2003.

[7] 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, Sept. 2001.

[8] S. Itahashi, "A noise database and Japanese common speech data corpus," J. Acoust. Soc. Jpn., vol. 47, no. 12, pp. 951-953, Dec. 1991. (in Japanese)

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

[10] http://www.speech.kth.se/wavesurfer/

[11] http://spib.rice.edu/spib/select_noise.html