Dominance spectrum based V/UV classification and F_0 estimation

Tomohiro Nakatani[†], Toshio Irino[†‡], and Parham Zolfaghari[†]
[†]NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
[†]nak@cslab.kecl.ntt.co.jp
[‡]Faculty of Systems Engineering, Wakayama University

Abstract

This paper presents a new method for robust voiced/unvoiced segment (V/UV) classification and accurate fundamental frequency (F_0) estimation in a noisy environment. For this purpose, we introduce *the degree of dominance* and *dominance spectrum* that are defined by instantaneous frequency. The degree of dominance allows us to evaluate the magnitude of individual harmonic components of speech signals relative to the background noise. The V/UV segments are robustly classified based on the capability of the dominance spectrum to extract the regularity in the harmonic structure. F_0 is accurately determined based on fixed points corresponding to dominant harmonic components easily selected from the dominance spectrum. Experimental results show that the present method is better than the existing methods in terms of gross and fine F_0 errors, and V/UV correct rates in the presence of background white and babble noise.

1. Introduction

In speech signal processing applications, the robust and accurate fundamental frequency (F_0) estimation and voicing detection are very important in resolving their problems especially in adverse noise conditions. For example, voiced/unvoiced segment (V/UV) classification is essential for obtaining the durations of user utterances in a human-machine dialogue system. F_0 has also been used as a major clue for sound source separation. By employing F_0-based sound separation as a preprocessor to a speech recognizer, we could obtain higher word recognition accuracy [1]. Recently, a very high quality vocoder, STRAIGHT [2], was developed based on F_0 adaptive processing. With such applications, F_0 estimation accuracy is very important since any error in F_0 has a detrimental effect on the system performance.

A number of F_0 estimation techniques and V/UV classification methods have been proposed, each having decision measures in different domains. The measures are categorized into two types: time-domain measures and frequency-domain measures. The former includes measures based on the autocorrelation function, the average magnitude difference function (AMDF), and their combinations [3]. The latter includes cepstrum coefficients, fundamentalness [2], and the instantaneous frequency (IF) amplitude spectrum [4]. Zero-crossing rates and signal power are also used for V/UV classification [5].

Although some measures are essential for accurate estimation and some are suitable for robust estimation in the presence of background noise, they still have certain limitations in real world applications where various adverse noise conditions can occur simultaneously. YIN developed by Alain de Cheveigné [3] is a highly accurate F_0 estimator based on a measure combining autocorrelation and AMDF, but it is not sufficiently robust in the presence of background noise. The IF amplitude spectrum is a robust measure to background noise, however, its property is sensitive to the spectral distortion caused, for example, by the frequency properties of microphones such as telephone handsets, and by spatial acoustics. In robust V/UV classification methods [5], robustness deteriorates when the recording conditions change such as changes in the signal level and variation in background noise. This deterioration is mainly a result of tuning for specific conditions.

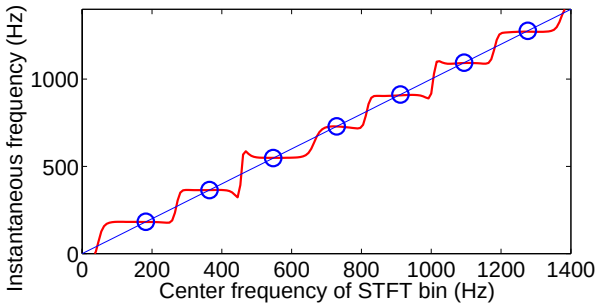

Figure 1: Instantaneous frequency (thick line) and fixed points (circles) of a speech sound.

To overcome the above problems, we proposed the dominance spectrum for F_0 estimation [6]. The dominance spectrum is an IF based measure for evaluating the magnitude of a harmonic component relative to the background noise at each STFT bin, and is shown to be robust in the presence of both background noise and spectral distortion. In this paper, we propose an integrated method based on the dominance spectrum, including V/UV classification and F_0 refinement. The robustness and accuracy of the method was evaluated using a large database of simultaneous recordings of speech and electro-glottal graph (EGG) signals [7]. We used white noise and babble noise which simulates a "cocktail party". As a result, the effectiveness of the proposed methods can be reliably confirmed under various kinds of adverse noise conditions.

2. Robust V/UV classification and accurate F_0 estimation

2.1. Definition of dominance

Figure 1 shows a mapping between the center frequency of each STFT bin, ω_c, and its derived instantaneous frequency (IF), $\dot{\phi}_\tau(\omega_c)$, for a voiced speech signal. Each harmonic component, which is an integer multiple of F_0, produces each flat region[1]. The best estimates of the harmonic frequencies are the frequencies at each *fixed point* where the IF coincides with the

[1]By contrast, the flatness disappears and the IF increases in proportion to the center frequencies of STFT bins when the component is severely smeared by white Gaussian background noise.

center frequencies of the bins as $\dot{\phi} = \omega_c$. Then, F_0 is estimated as the difference between the frequencies of adjacent fixed points since the harmonic frequencies are integer multiples of F_0. The use of IF is superior for F_0 estimation to conventional approaches, such as cepstrum methods, as it can provide a precise frequency estimate of each harmonic component.

The degree of dominance is calculated by evaluating the IF concentration in a specific frequency region around a fixed point. This degree is defined as $D_0(\tau, \omega_c)$ in Eq. (1) for each frequency bin centered at ω_c at frame time τ.

$$D_0(\tau, \omega_c) = 10 \log_{10}(1/B(\tau, \omega_c)^2), \quad (1)$$

$$B(\tau, \omega_c)^2 = \frac{\int_{\omega_c - \triangle\omega/2}^{\omega_c + \triangle\omega/2} (\dot{\phi}(\tau, \omega) - \omega_c)^2 S(\tau, \omega)^2 \mathrm{d}\omega}{\int_{\omega_c - \triangle\omega/2}^{\omega_c + \triangle\omega/2} S(\tau, \omega)^2 \mathrm{d}\omega}. \quad (2)$$

where $\dot{\phi}(\tau, \omega)$ represents IF for a frequency bin centered at ω and $S(\tau, \omega)^2$ is the power spectrum. $B(\tau, \omega_c)^2$ is derived as the weighted average of the squared difference between the center frequency of the frequency bin, ω_c, and the IF, $\dot{\phi}(\tau, \omega)$. The averaging was performed over the frequency range, $\triangle\omega$, with the weighting function being the power spectrum, $S(\tau, \omega)^2$.

The dominance spectrum has sharp peaks only at fixed points corresponding to dominant frequency components. This can be understood as follows: the value of $B(\tau, \omega_c)^2$ reaches minimum when the dominant frequency component of a signal coincides with the center frequency, ω_c, because the value of $\dot{\phi}(\tau, \omega)$ for the adjacent frequency bins approaches ω_c. Then the degree of dominance, $D_0(\tau, \omega_c)$, reaches its maximum value since it is defined as the logarithm of the inverse of $B(\tau, \omega_c)^2$ in Eq. (1). By contrast, the dominance value becomes smaller and does not have a sharp peak when the frequency component is greatly affected by noise because $\dot{\phi}(\tau, \omega)$ increases as ω increases and thus the difference between $\dot{\phi}(\tau, \omega)$ and ω_c becomes larger.

The dominance spectrum was shown to be a robust measure in the presence of both background noise and spectral distortion [6] since, 1) it enhances only dominant harmonic components as well as reduces background noise, and 2) it normalizes the spectral envelope reducing the influence of spectral distortion.

2.2. Robust V/UV classification method

2.2.1. Advantages of dominance spectrum

In addition, the dominance spectrum has characteristics that are advantageous for V/UV classification. The left panel of Fig. 2 shows an example of a voiced segment. The power spectrum, dominance spectrum and fixed points are shown with a thin line, a thick line, and circles, respectively. The figure clearly shows that the dominance spectrum has sharp peaks corresponding to harmonics, and they coincide exactly with the positions of the fixed points. Moreover, some spurious peaks in the power spectrum caused by background noise are completely absent from the dominance spectrum. By contrast, the right panel shows an example of an unvoiced segment. In this panel, the peaks in the dominance spectrum are not as clear and do not coincide with the fixed point positions. This regularity of the dominance spectrum in accordance with the harmonic structure is very useful for classifying V/UV segments.

Furthermore, the dominance spectrum has a property that is advantageous for classifying V/UV segments; its value is independent of the signal level. For example, the degree of dominance of a dominant harmonic component almost always has a

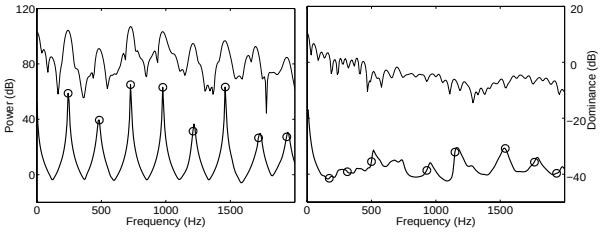

Figure 2: Dominance spectrum (thick lines), logarithmic power spectrum (thin lines), and fixed points (circles) of voiced (left panel) and unvoiced (right panel) segments

value between -20 and 0 dB even if the power in the target signals changes. On the other hand, it has smaller values between -40 and -20 dB between the dominant harmonic components. Therefore, the sum of the dominance values corresponding to the harmonic components has a larger value within a certain limited region than that corresponding to non-harmonic components without being affected by signal power.

2.2.2. V/UV classification method

To classify V/UV segments based on the degree of dominance, we define a decision measure, referred to as harmonic dominance. It is defined as the sum of the dominance values corresponding to fixed points near individual harmonic components, that is, multiples of the estimated F_0 at each frame time τ.

$$D_h(\tau) = \sum_{l=1}^{n} D_{0,F}(\tau, 2\pi l f_0(\tau)),$$

$$D_{0,F}(\tau, \omega) = \begin{cases} D_0(\tau, \omega) - \mathrm{E}(D_0(\tau, \omega_c)) & \text{if } \omega \in \dot{\Phi}(\tau), \\ 0 & \text{otherwise.} \end{cases}$$

Here, l is the index of a harmonic component, $f_0(\tau)$ is the estimated F_0, and $\dot{\Phi}(\tau)$ is the set of fixed points at frame time τ. $D_{0,F}(\tau, 2\pi l f_0)$ is the degree of dominance corresponding to l-th harmonic component, or zero if there is no fixed point at the frequency. $E(\cdot)$ is a function that calculates the average dominance value across frequencies, and is subtracted from each dominance value. Then, a median filter is applied to the time series of harmonic dominance to obtain a smoothed time sequence. V/UV segment is determined by thresholding the obtained value at each frame time.

$$\text{V/UV}(\tau) = \begin{cases} \text{voiced} & \text{if } M(D_h(\tau)) > \theta, \\ \text{unvoiced} & \text{otherwise,} \end{cases}$$

where $M(D_h(\tau))$ is a median value of the harmonic dominance series, $D_h(\tau)$, over frames included in $\tau \pm 30$ msec.

2.2.3. V/UV threshold determination

Figure 3 shows histograms of the harmonic dominance values after they had been filtered with a median filter $M(\cdot)$, for two speech databases. The left panel is a histogram for an adult clean speech database [7], and the right panel is a histogram for an infant utterance database recorded in daily child care settings [8]. Each plot clearly contains two peaks: one corresponding to unvoiced segments, and the other to voiced segments. Although each database contains a different F_0 range and the infant utterances contain varieties of background noise, the harmonic

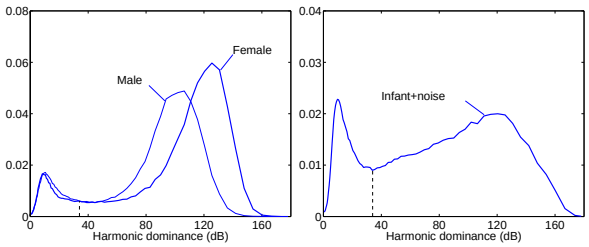

Figure 3: Histograms of harmonic dominance values normalized with the number of samples for adult male (left panel, thin line) and adult female (left panel, thick line) voices with no background noise, and for infant voices (right panel) with various background noises.

dominance values for the troughs between the two peaks are almost the same, that is, about 34 dB. Therefore, we assume 34 dB to be a desirable threshold for V/UV classification.

2.3. Accurate F_0 estimation method

Dominance spectrum based F_0 estimation method is composed of two steps: the initial rough estimate of F_0, and the refinement of this roughly estimated F_0. We have already proposed two methods for the initial estimate for fundamental frequency [6]; a dominance spectrum based method, and an alternative method based on the ripple-power spectrum[2]. In this paper, we will focus on the second refinement step, namely improving the accuracy of the F_0 estimation.

2.3.1. Accurate F_0 estimation method

In order to improve the accuracy of F_0 estimates, we introduce a F_0 refinement stage based on the IFs at fixed points. As fixed points with large dominance values are expected to be derived from dominant harmonic components, the IFs at such fixed points are considered as good candidates of the harmonic frequencies. Therefore, reliable F_0 candidates can be obtained by dividing their harmonic frequencies by their harmonic numbers. With our method, F_0 is determined as the weighted average of the F_0 candidates derived from fixed points using the degree of dominance as the weight. Because of this weight, F_0 is determined mainly based on fixed points of dominant harmonic components, and therefore, the obtained F_0 is expected to be reliable.

The original idea behind the F_0 refinement method based on fixed points was first introduced by Atake *et al.* [7] where Cohen's bandwidth equation was used to evaluate the reliability of fixed points. In this paper, this method is modified in that the degree of dominance is used.

Let F_0' be the rough estimate of the fundamental frequency given in the first step. The refined F_0 is then defined as follows:

$$F_0 = \frac{1}{2\pi} \frac{\sum_{i=1}^{n} \sum_{\dot{\phi} \in \Omega(\tau, i \cdot F_0')} (\dot{\phi}/i)\{D_0(\tau, r(\dot{\phi})) - c\}}{\sum_{i=1}^{n} \sum_{\dot{\phi} \in \Omega(\tau, i \cdot F_0')} \{D_0(\tau, r(\dot{\phi})) - c\}}, \qquad (3)$$

[2]Both are robust in the presence of background noise, and only the former is robust against spectral distortion. The latter has another advantage for F_0 estimation; its F_0 search region can easily be extended so that, for example, it can estimate the F_0 of infant utterances which inherently have a wide range (50 to 2000 Hz) [9].

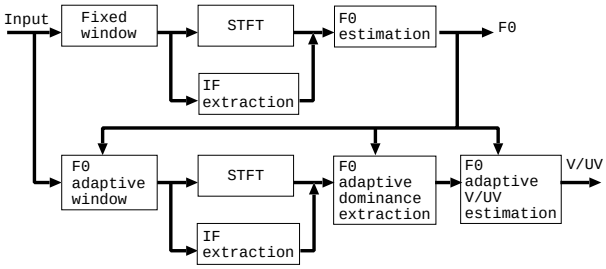

Figure 4: Processing flow

$$c = \min_{\dot{\phi} \in \Omega(\tau, i \cdot F_0'), i=1 \sim N} (D_0(\tau, r(\dot{\phi}))) - \epsilon. \quad (\epsilon > 0) \qquad (4)$$

Here, $\Omega(\tau, i \cdot F_0')$ is a set of IFs at fixed points that are located within $\pm 10\%$ of the i-th multiple of F_0', and N is the number of harmonics. Each $\dot{\phi}$ is a candidate frequency of the i-th harmonic component derived from a fixed point, so $\dot{\phi}/i$ is a candidate for F_0, and F_0 is calculated as the average $\dot{\phi}/i$ value weighted by the degree of dominance. $\mathrm{r}(\omega)$ is used to transform a continuous frequency to its nearest STFT center frequency. The term c enables the weights for all i to be greater than zero.

2.4. Processing flow

Figure 4 shows the processing flow of our integrated F_0 estimation and V/UV classification method. First, the F_0 is roughly estimated from the dominance/ripple-power spectrum and then refined by the method described in section 2.3. Then, the estimated F_0 is used to classify V/UV segment based on the method proposed in section 2.2. Here, we employed F_0 adaptive windowing in order to extract the instantaneous frequency more accurately [7]. The optimum integration length, $\triangle\omega$, for calculating the dominance spectrum is also determined based on F_0 [6]. Furthermore, the harmonic dominance is calculated according to the F_0.

3. Experiments

We used a Japanese speech database consisting of 30 utterances by 14 male and 14 female speakers (total of 840 utterances, 16 kHz sampling and 16 bit quantization) for the evaluation. The background noise was white noise and babble noise. The babble noise consisted of the utterances of 10 speakers randomly selected from the speech database and mixed so that the average power of each utterance was the same. The correct values of F_0 and V/UV were calculated from electro glottal graph (EGG) signals that were collected at the same time as the speech recordings [7]. Since EGG is a signal derived directly from the glottal pulses, it is considered to be an almost ideal signal for calculating the correct values. In our evaluation, we calculated the correct F_0 and F_0 of the target speech from the EGG and from the speech with background noise, using the same F_0 estimation method. Then, we compared these values to evaluate the performance of the method. We used the same procedure to evaluate the V/UV method.

3.1. Gross and fine F_0 errors

We evaluated the performance of our F_0 estimation method in terms of gross F_0 errors and fine F_0 errors. The F_0 gross error is the ratio between the number of frames giving "incorrect" F_0^{est} values and the total number of the frames. The "incorrect" F_0^{est} value is defined as the value beyond the range

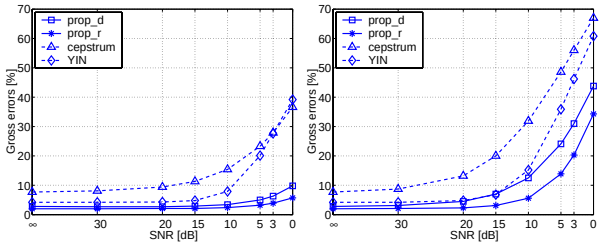

Figure 5: Gross F_0 errors with white noise (left panel) and babble noise (right panel).

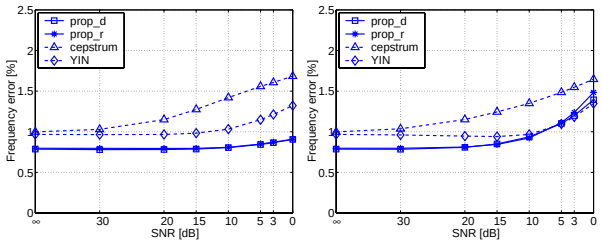

Figure 6: Fine F_0 errors with white noise (left panel) and babble noise (right panel).

of $\pm 5\%$ of the F_0^{cor} value. The fine F_0 error is the normalized root mean square error between the F_0^{cor} value and F_0^{est} value, which is not judged to be "incorrect" in the gross error measurement. The formulation is the root mean square of $(F_0^{\mathrm{cor}} - F_0^{\mathrm{est}})/F_0^{\mathrm{cor}}$. We compared our proposed method using dominance spectrum based initial F_0 estimation (prop_d), our proposed method using ripple-power spectrum based initial F_0 estimation (prop_r), the F_0 estimation method used in YIN [3], and the cepstrum method.

Figures 5 and 6 dipict the gross and fine F_0 errors obtained by individual F_0 estimation methods. Figure 5 clearly shows the robustness of the proposed methods as is reported in [6]. Furthermore, Fig. 6 shows the superior accuracy in F_0 estimation with the proposed methods to the other existing methods.

3.2. V/UV correct rate

We evaluated the performance of our V/UV classification method in terms of two types of correct rates, rates at which the voiced segments are correctly estimated as voiced segments (V), and rates at which the unvoiced segments are correctly estimated as unvoiced segments (UV). For comparison, we used the speech analysis toolkit (SPTK) version 2.0 developed by Tokuda et al. Figure 7 shows that our proposed method could provide superior results under all SNR conditions. For example, when the SNR was higher than 15 dB, our proposed method gave much more accurate results than SPTK. When the SNR was lower than 10 dB, SPTK completely mis-determined almost all the frames as V in both panels, while our method could, to some extent, distinguish V/UV correctly.

4. Conclusion

We proposed a robust V/UV classification scheme and an accurate F_0 estimation method based on the dominance spectrum in the presence of background noise. The dominance spectrum is useful for extracting the regularity of the harmonic structure, and therefore, is advantageous for detecting voicing robustly. In addition, the dominance spectrum makes it easy to select dominant harmonic components, which is essential in improving the accuracy of F_0 estimation using fixed points. Exprimental results showed that the proposed V/UV classification and F_0 estimation methods were more robust and accurate than existing methods in the presence of background noise consisting of white noise and babble noise.

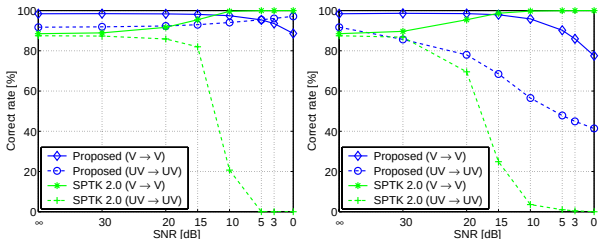

Figure 7: V/UV correct rate with white noise (left panel) and babble noise (right panel).

We express our gratitude to H. Kawahara of Wakayama Univ. for providing the database of speech with EGG, S. Katagiri for research support, and members of Speech Open Lab. of NTT for helpful discussions.

5. References

[1] Nakatani, T., and Okuno, H. G. "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Communications*, Vol. 27, Nos. 3-4, pp. 209–222, Elsevier, 1999.

[2] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communications*, Vol. 27, Nos. 3-4 pp. 187–207, Elsevier, 1999.

[3] de Cheveigné, A., and Kawahara, H. (2002), "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.* vol. 111, pp. 1917–1930, 2002.

[4] Abe, T., Kobayashi, T., and Imai, S. (1997), "The IF spectrogram: A new spectral representation," *Proc. ASVA 97*, pp. 423–430, 1997.

[5] Ahmadi, S., and Spanias, A. S., "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. SAP*, vol. 7, no. 3, pp. 333–338, 1999.

[6] Nakatani, T., and Irino, T., "Robust fundamental frequency estimation against background noise and spectral distortion," *Proc. ICSLP-2002*, vol. 3, pp. 1733–1736, Denver, Sep., 2002.

[7] Atake, Y., Irino, T., Kawahara, H., Lu, J., Nakamura, S., and Shikano, K. (2000), "Robust fundamental frequency estimation using instantaneous frequencies of harmonic components," *Proc. ICSLP-2000*, vol. II, pp. 907–910, 2000.

[8] Amano, S., Kato, K., and Kondo, T., "Development of Japanese infant speech database and speaking rate analysis," *Proc. ICSLP-2002*, vol. 1, pp. 317–320, Denver, Sep., 2002.

[9] Nakatani, T., Amano, S., and Irino, T., "An estimation method for fundamental frequency and voiced segment in infant utterance," *JASA*, Vol. 112, No. 5, Pt. 2 of 2, p. 2322 (144th Meeting of ASA), Nov., 2002.