ABSTRACT

This paper presents a new method for robust fundamental frequency ($F_0$) estimation in the presence of background noise and spectral distortion. We define degree of dominance and a dominance spectrum based on instantaneous frequencies. The degree of dominance allows us to evaluate the magnitude of individual harmonic components of speech signals relative to background noise while eliminating the influence of spectral distortion. The fundamental frequency is robustly estimated from reliable harmonic components easily selected from the dominance spectra. Experiments are performed using white and multi-talker background noise with and without spectral distortion produced by a SRAEN filter. Results show that the present method is better than the commonly-used methods in terms of correct $F_0$ rates.

1. INTRODUCTION

Robust fundamental frequency ($F_0$) estimation is an important problem in relation to speech signal processing. By employing an $F_0$-based sound segregation system as a pre-processor, we were able to demonstrate an improvement in the recognition rate of a speech recognition system [1]. In this case, robust $F_0$ estimation is essential under adverse conditions since any error in $F_0$ is fatal as regards sound separation. Accurate $F_0$ estimation is also important for the STRAIGHT system which was recently developed as a high quality vocoder [2].

In the real world, the detrimental factors in $F_0$ estimation are mainly divided into two types: background noise and spectral distortion. Spectral distortion is often caused by, for example, spatial acoustical effects and the frequency properties of microphones, such as telephone handsets. Although a large number of $F_0$ estimation methods have been proposed [3, 4, 5], their robustness in the face of these two kind of factors has not been evaluated. Only robustness against background noise has usually been evaluated using white Gaussian noise. However, its spectral density and temporal regularity are very different from those of speech signals making it insufficient for a comprehensive evaluation of $F_0$ estimation. Furthermore, these methods were not necessarily evaluated with a large and reliable database to enable a reliable estimate to be made of the “correct value” of the fundamental frequencies estimated independently of the speech signal. The limits of the evaluation preclude the possibility of making a final decision as to the best method.

We evaluated our new method using a large database of simultaneous recordings of speech and electro glottal graph (EGG) signals. The robustness was evaluated with a multi-talker noise which simulates a “cocktail party” as well as the white noise conventionally used. We also applied spectral distortion to the speech signals using a SRAEN filter recommended by ITU-T [6]. This is a high pass filter above 300 Hz which we used to simulate a telephone handset.

The present method is basically a frequency domain method using instantaneous frequency (IF). The use of IF in $F_0$ estimation was initiated by Charpentier [7] and, recently, its effectiveness was reported for $F_0$ estimation under noisy conditions [2, 8, 9]. When a dominant frequency component of a sound source nearly coincides with a specific short-time Fourier transform (STFT) frequency bin, the IFs for the adjacent frequency bins concentrate at almost the same value. This provides a good estimate of the frequency of the dominant component. For speech sounds, it is a harmonic frequency of a multiple of the fundamental frequency, $F_0$. When we draw the mapping function between the center frequency of the STFT bin, $\omega_c$, and the derived IF, $\phi(\omega_c)$, it has the appearance of a regularly spaced staircase for $F_0$ as shown in Fig. 1. So the harmonic frequencies are estimated as the frequencies when the IFs coincide with the center frequencies of the bins as $\phi = \omega_c$. The point is referred to as a fixed point. Then $F_0$ is estimated as the difference between the frequencies of adjacent fixed points.

Abe et al. proposed an “IF amplitude spectrum”. This is a variation of an amplitude spectrum, in which amplitude values $S(\omega_c)$ in a usual amplitude spectrum are reorganized according to IF, $\phi(\omega_c)$, instead of $\omega_c$ [8]. Since the IF amplitude spectrum enhances the harmonic structure of speech sounds, it is consid-
ered applicable to \( F_0 \) estimation in the presence of background noise. However, the \( F_0 \) estimation performance using this spectrum deteriorates severely in the presence of spectral distortion caused by, for example, telephone handsets because the IF amplitude spectrum inherits spectral distortion from the amplitude spectrum. Moreover, the IF spectrum does not always reduce noise components since it is defined using a differential operation which usually enhances the noise.

The purpose of this study is to develop a new method to overcome the problems posed by the conventional methods. For this purpose, we define the dominance spectrum in which both background noise and spectral distortion are efficiently reduced. We describe \( F_0 \) estimation based on dominance in section 2 and we evaluate the method in section 3.

2. METHOD

2.1. Definition of dominance

We introduce a new measure to evaluate the magnitude of a harmonic component relative to the other components at the output of an STFT frequency bin. We call this “the degree of dominance” and define it as \( D_0(\omega_c) \) in Eq. (1) for the frequency bin centered at \( \omega_c \).

\[
D_0(\omega_c) = \log(1/B(\omega_c)^2),
\]

\[
B(\omega_c)^2 = \frac{\int_{\omega_c-\Delta \omega}^{\omega_c+\Delta \omega} (\phi(\omega) - \omega_c)^2 S(\omega)^2 d\omega}{\int_{\omega_c-\Delta \omega}^{\omega_c+\Delta \omega} S(\omega)^2 d\omega},
\]

where \( \phi(\omega) \) represents the IF for the frequency bin at \( \omega \) and \( S(\omega) \) is the amplitude spectrum. \( B(\omega_c)^2 \) is derived as the weighted average of the squared difference between the center frequency of the frequency bin, \( \omega_c \), and IFs, \( \phi(\omega) \), over the frequency range, \( \Delta \omega \). The weighting function is the power spectrum, \( S(\omega)^2 \).

The value of \( B(\omega_c)^2 \) becomes minimum when the dominant frequency component of a signal almost coincides with the center frequency, \( \omega_c \), because \( \phi(\omega) \) becomes flat and close to \( \omega_c \) around the fixed point as shown in Fig. 1. Then the degree of dominance, \( D_0(\omega_c) \), becomes maximum since it is defined as the logarithm of the inverse of \( B(\omega_c)^2 \) in Eq. (1). By contrast, the dominance value becomes smaller when the frequency component is greatly affected by noise because \( \phi(\omega) \) increases in proportion to \( \omega \) and thus the difference between \( \phi(\omega) \) and \( \omega_c \) becomes larger.

2.1.1. Robustness against background noise

When the degree of dominance, \( D_0(\omega_c) \), is calculated for all the STFT bins, it is represented as a sort of spectrum (Fig. 2 (a)). It is referred to as a “dominance spectrum”. The dominance spectrum includes sharp peaks which correspond to the harmonic components. The peaks are much sharper than those for a usual logarithmic power spectrum (Fig. 2 (b)). When the signal is smeared with background noise, neither the dominance spectrum nor the logarithmic power spectrum (Fig. 2 (d) and (e)) have sharp peaks corresponding to the harmonic components above 500 Hz. For this frequency range, the peak to trough ratio is smaller in the dominance spectrum than in the logarithmic power spectrum. It is opposite at lower frequencies. So the dominance spectrum enhances the peaks for the harmonic components and suppresses the variation produced by noise. This property is particularly useful for robust \( F_0 \) estimation. By contrast, Fig. 2 (c) and (f) show that the power spectrum is also a robust as regards background noise.

2.1.2. Robustness against spectral distortion

The dominance spectrum also has a property that whitens the spectral envelope including formant peaks or eliminates the effect of spectral distortion. This can be clearly demonstrated by using a speech signal with high pass filtering for a telephone handset. Figure. 2(a) and (g) show the dominance spectra for an original speech signal and the signal filtered with a SRAEN filter. The first and the second peaks of the dominance spectrum in Fig. 2(g) are smaller than those in Fig. 2(a), but the difference is slight. Figure. 2(c) and (i) show the power spectra. The peaks below 300 Hz are clearly distorted by the filtering, and this shows the power spectrum is much more sensitive to spectral distortion than the dominance spectrum. By contrast, in the logarithmic power spectrum, the peaks in lower frequency regions are reduced as shown in Fig. 2(b) and (h), however, the shape of the peaks clearly remains. This shows that the logarithmic power spectrum is not as sensitive to spectral distortion as the power spectrum.

Consequently, only the dominance spectrum is robust against both background noise and spectral distortion, and thus must be useful for \( F_0 \) estimation.

2.2. \( F_0 \) estimation based on dominance

Now, we define a decision measure which summarizes the degree of dominance for all harmonic components. It is called harmonic dominance and is defined as \( D_{10}(\omega_0) \) in Eq. (3).

\[
D_{10}(\omega_0) = \sum_i \{D_0(l \cdot \omega_0) - E(D_0(\omega_c))\},
\]

\[
F_0 = \frac{1}{2\pi} \arg \max_{\omega_0} \{D_{10}(\omega_0)\},
\]

where \( l \cdot \omega_0 \) corresponds to the frequency of the \( l \)-th harmonic component when \( \omega_0 \) is the fundamental frequency. \( E(D_0(\omega_c)) \) is the
average value of $D_0(\omega_c)$ over the number of frequency bins, i.e.,
$$\sum_{\omega_c} D_0(\omega_c)/N.$$ $E[D_0(\omega_c)]$ is a term that ensures that the measure
is unbiased and thus reduces specific errors known as “double pitch” and “half pitch” errors. A frequency, $\omega_0$, that maximizes
the harmonic dominance as in Eq. (4) is a good estimate of the
fundamental frequency $F_0$.

Figure 3 shows the $F_0$ estimation flow based on the degree of
dominance. First, an input signal was down-sampled (to 4 kHz),
and converted to a signal in the frequency domain by STFT (512
points using a 42 ms Hanning window). Next, we calculated
the dominance spectrum $D_0(\omega_c)$ based on IF as explained in the
previous section. We used Flanagan’s method [10] for the IF extraction.
Then, we estimated $F_0$ using Eq. (4). Incremental dynamic pro-
gramming (incremental DP, a modified version of DP that works
incrementally), is effective as regards avoiding discontinuous transi-
tion error in $F_0$ [11]. Finally, $F_0$ is further refined based on the
degree of dominance of fixed points [11]. Because of this refine-
ment, our method can also provide very accurate $F_0$ estimation.
We will discuss about this in another article.

2.2.1. Frequency range for integration

The optimum frequency range, $\Delta \omega$, for the integration in Eq. (2)
depends on the $F_0$ of the input signal. $F_0$ can be estimated to
be more robust if the optimum range is given based on a rough
estimate of $F_0$. For this purpose, we introduce two methods by
which to decide the range. One method provides a fixed range with
preliminary information and the other provides an adaptive range
without preliminary information. With the fixed range method, the
gender of the target speech is assumed to be known in advance, and
the integration range that is the optimum with regard to the gen-
der is applied to the signal. In our experiment, the optimum range
was about 130 Hz for a man and 260 Hz for a woman. By con-
trast, with the adaptive range method, the harmonic dominance is
maximized twice, first for a rough estimate, and then to obtain a
more precise value. We use an integration range applicable to both
men and women for the rough estimate, and we use the optimum
range for the rough estimate of $F_0$ for the precise estimation.
In our experiment, the 260 Hz range (the same as the optimum value
for a woman) was suitable for the rough estimate, and about 67% ~ 110% of the rough $F_0$ estimate was suitable for the precise es-
timation.

3. EXPERIMENTS

3.1. Evaluation method

Obtaining the correct $F_0$ of target speech is very important when
evaluating $F_0$ estimation performance. In conventional evaluation
methods, it is sometimes determined by observing the logarithmic
power spectrum of the speech, and the $F_0$ estimation performance
depended on the way in which an individual determined the cor-
rect $F_0$. In other cases, the correct $F_0$ is determined from the $F_0$
estimated from target speech without noise using a certain $F_0$ esti-
mation method. The influence of the vocal tract in speech might
affect the correct $F_0$ estimation with this method. We believe the
correct $F_0$ should be determined more carefully.

With our method, the correct $F_0$ is calculated from electro
glottal graph (EGG) signals that were collected at the same time
as the speech utterances were recorded [9]. Since EGG is a signal
derived directly from the glottal pulses, it is considered to be an
almost ideal signal for calculating the correct $F_0$. In our evaluation,
we calculated the correct $F_0$ and $F_0$ of the target speech using the
same $F_0$ estimation algorithm from EGG and from the speech with
background noise, respectively. Then, we compared these values
in terms of correct $F_0$ rates. The correct $F_0$ rate is the rate where
the estimated $F_0$ is within ±5% of the correct $F_0$. Of course, this
evaluation method does not completely define the correct $F_0$, but
it can provide an adequate basis for examining the robustness of a
$F_0$ estimation against noise and the influence of the vocal tract.

We used a Japanese speech database consisting of 30 utter-
ances by 14 male and 14 female speakers (total 840 utterances,
16 kHz sampling and 16 bit quantization) for the evaluation. The
background noise was white noise (noise-1) and multi-talker noise
(noise-2). The multi-talker 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. For ex-
ample, target speech can be distinguished by the human ear with
multi-talker noise of 0 dB SNR, although this is a very severe con-
dondition for $F_0$ estimation. About half the target speech can still
be distinguished with multi-talker noise of ~5 dB SNR, but noth-
ing can be distinguished when the SNR is ~10 dB. We compared
and evaluated our proposed method with a fixed integration range
(proposed-1), our proposed method with an adaptive integration range
(proposed-2), the $F_0$ estimation method used in STRAIGHT [2], and the cepstrum method.

3.2. Robustness against background noise

Figure 4 shows the correct rates of $F_0$ estimation with background
noise. It shows that the proposed-1 and proposed-2 methods are
more robust in terms of correctly estimating $F_0$ than existing meth-
ods for background noise with different SNRs in the infinite dB range.
Moreover, the proposed-2 method is as effective as the
proposed-1 method, despite not using preliminary information.
The proposed-1 method is only superior to the proposed-2 method
when the multi-talker noise has an SNR of greater than 10 dB.

3.3. Use of power spectrum

The power spectrum was examined for $F_0$ estimation compared
with dominance spectrum. Since the power spectrum is robust
against background noise as shown in section 2.1.1, it may be able
to be used as an alternative to the dominance. We examined a normalized version of the power spectrum, $R(\omega_c)$, and substituted it for $D_0(\omega_c)$ in Eq. (3) leaving the other processing unchanged. Normalization was employed to eliminate the spectral envelope in the power spectrum, $S(\omega_c)^2$, by applying discrete Fourier transformation (DFT) to $S(\omega_c)^2$, substituting zeros for its lower frequency bins, and applying inverse DFT to it. The $F_0$ estimation performance was improved in almost every case by this normalization in our preliminary experiments.

The thick solid lines in Fig. 5 show correct $F_0$ rates with white and multi-talker noise using the power spectrum (PowerSpec-1). The results were not good, particularly under high SNR conditions, compared with the results obtained using the dominance spectrum shown in Fig.4. This is because there was a large error in the correct $F_0$ estimation from the EGG signal. The spectral power in the region lower than $F_0$ often exceeds the power of the $F_0$ components in the EGG signal, and that precludes exact $F_0$ estimation. To eliminate the influence of the spectral property in the EGG signal, it was effective to apply pre-emphasis to input signals in advance (PowerSpec-2), and the result of this method is shown by the thin solid lines in Fig.5. The correct rate under less noisy conditions is improved but rapidly worsens as the background noise increases. This is because the robustness of target speech $F_0$ estimation is degraded by the pre-emphasis. Finally, the thick dashed lines in Fig. 5 show the correct rate when the correct $F_0$ is calculated with pre-emphasis from EGG signal and $F_0$ of the target speech is calculated without pre-emphasis. The correct rate of this method (PowerSpec-3) is much better than with the above two methods, and even slightly better than the proposed-1 method. These results show that the power spectrum is very robust against additive noise but much less robust against spectral distortion. Although the power spectrum can potentially provide good performance if the pre-processing of the target signal is adequately chosen, it is difficult to specify a particular type of pre-processing that is appropriate for all spectral conditions. By contrast, the degree of dominance provides almost the best $F_0$ estimation without employing any additional pre-processing.

3.4. Robustness against spectral distortion

Figure 6 shows the correct $F_0$ rates of target speech convoluted with a SRAEN filter under white and multi-talker noise conditions. A SRAEN filter simulates the ideal microphone property of a telephone [6]. The figures show that the two proposed methods can still provide the best performance. By contrast, the correct $F_0$ rates of the PowerSpec-3 method deteriorated severely because of the spectral distortion.

4. CONCLUSION

We proposed a robust method for $F_0$ estimation using a new measure, namely the degree of dominance. The degree of dominance is defined to allow us to evaluate the magnitude of harmonic components of speech signals relative to background noise while eliminating the influence of spectral distortion. Our proposed $F_0$ estimation method was more robust than conventional methods with background conditions consisting of white noise and multi-talker noise. The proposed methods were also proven to be robust against spectral distortion caused by a SRAEN filter.

Further improvement is required for such applications as sound segregation under multi-talker or cocktail party conditions. For example, it would be possible to use a dual $F_0$ estimation method using sound source localization information when a multi-microphone is available [1].

Acknowledgments The authors express their gratitude to H. Kawahara of Wakayama Univ. for providing the speech and EGG database, S. Katagiri for research support, and members of the Speech Open Lab. of NTT for helpful discussions. This work was partially supported by CREST of JST.

5. REFERENCES