The Journal of the Acoustical Society of America

ASA ACOUSTICAL SOCIETY OF AMERICA

DECEMBER 24 2024

Robust fundamental frequency-detection algorithm unaffected by the presence of hoarseness in human voice

Itsuki Kitayama; Kiyohito Hosokawa; Shinobu Iwaki; Misao Yoshida; Akira Miyauchi; Toshihiro Kishikawa; Hidenori Tanaka; Takeshi Tsuda; Takashi Sato; Yukinori Takenaka; Makoto Ogawa; Hidenori Inohara

Check for updates

J. Acoust. Soc. Am. 156, 4217–4228 (2024)

https://doi.org/10.1121/10.0034624

View Online

Export Citation

JASA ARTICLE

Robust fundamental frequency-detection algorithm unaffected by the presence of hoarseness in human voice

Itsuki Kitayama,[1] Kiyohito Hosokawa,[1,2,a)] Shinobu Iwaki,[3] Misao Yoshida,[4] Akira Miyauchi,[5] Toshihiro Kishikawa,[1] Hidenori Tanaka,[1] Takeshi Tsuda,[1] Takashi Sato,[1] Yukinori Takenaka,[1] Makoto Ogawa,[1] and Hidenori Inohara[1]

[1]*Department of Otorhinolaryngology and Head & Neck Surgery, Osaka University Graduate School of Medicine, Osaka 565-0871, Japan*

[2]*Department of Otorhinolaryngology, Osaka International Medical & Science Center, Osaka 543-0035, Japan*

[3]*Department of Rehabilitation, Kobe University Hospital, Hyogo 650-0017, Japan*

[4]*Department of Rehabilitation, Sakai Heisei Hospital, Osaka 599-8236, Japan*

[5]*Department of Surgery, Kuma Hospital, Hyogo 650-0011, Japan*

ABSTRACT:

The fundamental frequency (f_o) is pivotal for quantifying vocal-fold characteristics. However, the accuracy of f_o estimation in hoarse voices is notably low, and no definitive algorithm for f_o estimation has been previously established. In this study, we introduce an algorithm named, "Spectral-based f_o Estimator Emphasized by Domination and Sequence (SFEEDS)," which enhances the spectrum method and conducted comparative analyses with conventional estimation methods. We analyzed 454 voice samples and used conventional methods and SFEEDS to calculate f_o. The ground truth of f_o was determined as the lowest frequency within the most dominant harmonic complex observed on the spectrogram. Subsequently, we assessed the concordance between each f_o-estimation method and the f_o ground truth. We also examined the variations in the accuracy of these methods when analyzing speech with hoarseness. Regardless of hoarseness, the f_o-estimation accuracy was significantly greater by SFEEDS than by conventional methods. Moreover, whereas the conventional methods impaired f_o-estimation accuracy in samples with roughness, the SFEEDS algorithm was robust and significantly reduced subharmonic errors. The SFEEDS f_o-estimation algorithm accurately estimated the f_o of both normal and hoarse voices. © *2024 Acoustical Society of America.*

https://doi.org/10.1121/10.0034624

(Received 22 June 2024; revised 5 October 2024; accepted 25 November 2024; published online 24 December 2024)

[Editor: B. Yegnanarayana] Pages: 4217–4228

I. INTRODUCTION

Pitch is defined as a subjective sensation of the frequency of a sound perceived through a human auditory system. In purely periodic sound signals, pitch aligns with the fundamental frequency (f_o), which is inversely related to the period of a signal. Pitch and f_o are often confused, and the process of estimating f_o is often referred to as the "pitch-detection algorithm" (Hess, 1983). However, the relationship between pitch and f_o is not simple due to harmonic components and the frequency characteristics involved in voice production, so pitch and f_o may not correlate consistently. The presence of subharmonics can cause a divergence between f_o and its perceived pitch, resulting in the simultaneous perception of dual pitches (Cavalli and Hirson, 1999). Accurate and robust estimation algorithms can be used in various applications. For example, the accuracy of speech recognition can be improved by recognizing speech tones to identify homonyms (Wang, 2001) and emotions by speech signals (Kwon *et al.*, 2003). Accurate f_o estimation is also very important for other applications, such as voice-quality evaluation, speech synthesis, speech coding, and speaker recognition, which are fundamental to a wide variety of speech-processing applications.

In real-world environments, the quality of the input audio signal can be significantly compromised by noises originating from background or recording equipment. Therefore, f_o-estimation methods should be tolerant to these environmental noises. To address this issue, a number of pitch detection algorithms have been developed to date. Especially for voice-quality evaluation, other than environmental noises, a laryngeal noise due to a disturbed vocal-fold vibration is also a problem when attempting to accurately estimate f_o in a hoarse voice and can lead to misunderstanding or underestimation of the degree of a hoarseness (Dejonckere *et al.*, 2011). Existing f_o-estimation methods were not designed to be robust to voice samples with a laryngeal noise.

A. Existing f_o-estimation methods

Existing f_o-estimation algorithms are categorized into three principal groups: those that predominantly use time-domain characteristics, those that use frequency-domain characteristics, and hybrid approaches that integrate both

[a)]Email: khosokawa@ent.med.osaka-u.ac.jp time- and frequency-domain characteristics (Titze and Liang, 1993).

The prevalent method for leveraging the time-domain characteristics involves signal correlation (Ross *et al.*, 1974; Rabiner *et al.*, 1976). Autocorrelation (AC) and cross correlation (CC) are widely used as standard f_o-estimation methods in Praat software (Paul Boersma and David Weenink, the Institute of Phonetic Sciences, University of Amsterdam, Amsterdam, The Netherlands: http://www.praat.org/) (Boersma, 2001). Alternative methods for estimating f_o in the time domain include the peak-picking method and zero-crossing method, which involve detecting specific time-domain events (Hess, 1983). However, it has been established that their f_o-estimation accuracy is less reliable than that of methods employing signal correlation, such as AC and CC (Titze and Liang, 1993). The algorithm named YIN, which is derived from the speech CC function, significantly reduces unwanted peaks and superfluous operations, achieving an error reduction of a third or less compared with the reduction in methods prior to 2002 (De Cheveigné and Kawahara, 2002).

The power spectrum of a highly periodic speech signal forms a harmonic complex with spectral peaks at integer multiples of f_o (Baken, 1987). The spectral method estimates f_o by identifying the frequency of the lowest peak in the power spectrum and estimating the intervals between peaks of the harmonic structure (Noll, 1964). However, the power spectrum-based estimation is influenced by articulatory filters, such as formants, necessitating a method to mitigate these effects. Sawtooth waveform-inspired pitch estimator (SWIPE) and Sawtooth waveform-inspired pitch estimator prime (SWIPE′), developed in 2008, considering the harmonic structure of the power spectrum, enhancing the precision of f_o estimation through innovative error-reduction techniques (Camacho and Harris, 2008). The cepstrum method, developed by Noll, uses a parameter known as the quefrency, derived by inverse Fourier transformation of the log-power spectrum (Noll, 1967). This method effectively separates the cepstrum from the influence of articulation filters, making it a pivotal parameter for both speech analysis and f_o estimation. The cepstrum method calculates f_o by finding a quefrency of the highest peak of the cepstrum waveform (f_o is the reciprocal of the quefrency). Nonetheless, the cepstrum method's f_o-estimation accuracy is limited and is prone to noise interference (Ba *et al.*, 2012).

BaNa is a hybrid approach involving harmonic frequency ratios and the cepstrum approach (Ba *et al.*, 2012). An evaluative study on the precision of various leading-edge f_o-estimation algorithms using online speech and noise databases revealed that BaNa outperformed other algorithms across all examined noise types and signal-to-noise ratio (SNR) levels (Sukhostat and Imamverdiyev, 2015).

B. Limitations of existing f_o-estimation algorithm

The advancement of the various f_o-estimation methodologies outlined above has achieved exceptional accuracy in differentiating between harmonic structures and nonperiodic noise. However, these algorithms were developed mainly for normal voices without hoarseness.

Hoarseness, i.e., voices containing laryngeal noise, often makes it difficult to estimate f_o. Breathiness and roughness are two important perceptual features of hoarseness (Dejonckere and Lebacq, 1996). Breathy voices have various causes, such as glottal closure failure, and their acoustic complexity has already been studied to some extent (Latoszek *et al.*, 2017; Hosokawa *et al.*, 2019b). A rough voice also has various causes, such as vocal-fold lesions, involving complex acoustic properties, such as subharmonics or structures that have not yet been clarified (van Latoszek *et al.*, 2018). Titze (1994) reported that asymmetric vocal-fold oscillations cause subharmonic structure in acoustic waveforms over two to three cycles when the two states alternate in period and amplitude. In a spectrum, these subharmonic structures emerge as distinct peaks situated between successive harmonic structures aligned with f_o, customarily partitioning the harmonic interval into multiple equal segments (e.g., 1/2, 1/3, and 1/4), thereby complicating the estimation of f_o by use of conventional methods (Baken, 1987).

A few evaluation methods have been developed to quantify subharmonic. The Degree of Subharmonics measure, part of the Multidimensional Speech Program (KayPENTAX, USA) acoustic analysis package, assesses the temporal dominance of subharmonics, although its precision is dependent upon successful f_o detection (Deliyski, 1993). The Diplophonia Diagram, which evaluates the qualities of both single and multiple oscillators, faces limited clinical applicability due to its computational demands (Aichinger *et al.*, 2017). Additionally, validation experiments comparing the method proposed by Awan and Awan (2020) which involves conducting a two-stage cepstrum analysis by segmenting the analysis frequency band, revealed the challenges in quantifying subharmonics and emphasized the critical role of f_o estimation in assessing hoarse voice (Kitayama *et al.*, 2023). Considering the potential presence of subharmonics, there is a clear need to develop a robust f_o-estimation algorithm that accounts for the temporal- f_o transitions in speech waveforms generated during the oral reading of texts.

C. Development of a novel f_o-estimation algorithm

To develop a novel algorithm for accurate f_o estimation in hoarse voices, we created an algorithm named Spectral-based f_o Estimator Emphasized by Domination and Sequence (SFEEDS), which incorporates two features named the Dominant Spectrum Test and the Sequential Spectrum Test, based on the spectral method. This algorithm was developed as a script file in the free software, Praat.

The primary objective of this study was to develop a new f_o-estimation algorithm capable of providing reliable f_o estimates not only in normal speech but also in speech containing subharmonics. The second objective was to define the ground truth for f_o in voice samples containing subharmonics, which was previously challenging when using human pitch perception or laryngography. The third objective was to use the defined ground truth for f_o to compare SFEEDS with traditional f_o-estimation methods and investigate their validity and effectiveness. The study outcomes will offer new alternatives for f_o estimation, a fundamental aspect of speech processing, and contribute to the development of key technologies for identifying the acoustic characteristics of not yet fully understood hoarse voices.

II. METHODS

A. Dataset

To enhance the versatility of the algorithm, continuous speech (CS) was included in this study, which differs from an earlier study in which only the conventional sustained vowel (SV) was included (Zraick *et al.*, 2005; Maryn *et al.*, 2010). A total of 454 recordings consisting of SV and CS were obtained from a dataset used in a former study (Hosokawa *et al.*, 2019b). The incorporated data comprised 288 voice recordings of individuals with diverse types of voice disorders of varying degrees of dysphonia, 55 voice recordings of individuals with no speech-related complaints, and 111 voice recordings that were >3 months post-treatment (see supplementary material). A head-worn microphone (SE50; Samson Technologies Corp., Hicksville, NY) was used to record all samples in a sound-treated room using and digitized at a sampling rate of 44.1 kHz with 16-bit resolution. A linear PCM recorder H4n (Zoom Corp., Tokyo, Japan) was used for the recording. The samples were confirmed to meet the generally required level of SNR (>30 dB) (Deliyski *et al.*, 2005; Deliyski *et al.*, 2006). The participants were instructed to sustain the vowel /a:/ for a minimum of 3 s and to read the Japanese translation of "The North Wind and the Sun" at a comfortable pitch, loudness, and pace. The procedures used to prepare the CS and SV samples were identical to those required to calculate the Acoustic Voice Quality Index for Japanese speakers (Hosokawa *et al.*, 2019a). For the CS sample, 30 syllables were selected from the first sentence to the eighth syllable of the second sentence (/aruhi kitakaze to taiyoː ga chikara kurabe wo shimashita tabibito no gaitoː wo/). For the SV samples, the middle vowel was extracted for 3 s, excluding the beginning and ending parts, except for patients who were unable to sustain the vowel for >3 s. The CS and SV samples were then concatenated for analysis. The data included information on sex, age, and diagnosis as well as the GRBAS scale (auditory-perceptual judgment of the degree of hoarseness levels) in SV and CS by three raters whose intra- and inter-rater reliabilities were established in a previous study (Hosokawa *et al.*, 2019b). The GRBAS scale consists of a score indicating the overall rating of hoarseness (G: grade) and four basic elements (R: roughness, B: breathiness, A: asthenia, and S: strain) (DeBodt *et al.*, 1997; Yamaguchi *et al.*, 2003). Roughness and breathiness are two particularly important perceptual features of hoarseness (Dejonckere and Lebacq, 1996; DeBodt *et al.*, 1997; Yamaguchi *et al.*, 2003), and each of these parameters (G score, R score, and B score) is rated on a scale from 0 to 3: 0, normal; 1, mildly abnormal; 2, moderately abnormal; 3, severely abnormal.

The G-scores of the CS and SV samples assessed by the three raters were totally averaged for each individual, generating the score of G_{total}. Similarly, the R_{total} and B_{total} were calculated. The presence of hoarseness, roughness, and breathiness was defined as G_{total}, R_{total}, and $B_{total} > 0.5$, respectively, which are used as general threshold values (Barsties and Maryn, 2016; Hosokawa *et al.*, 2017; Latoszek *et al.*, 2017; Hosokawa *et al.*, 2019a). For each of the 454 voice samples, the distributions of G_{total}, R_{total}, and B_{total} are shown in Fig. 1.

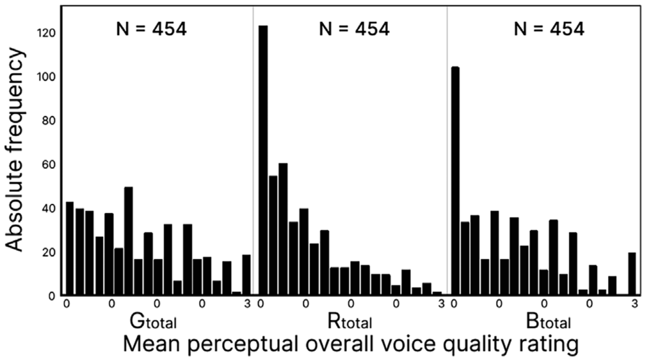

FIG. 1. Frequency distributions of auditory-perceptual judgments of the G_{total}, R_{total}, and B_{total} on the 454 total voice samples.

B. SFEEDS

SFEEDS is an algorithm designed for accurate estimation of hoarse voice. Its two main features are the Dominant Spectrum Test and the Sequential Spectrum Test. The following provides an overview and further details about these features. This f_o-estimation algorithm consists of a hybrid approach that integrates time- and frequency-domain properties (see supplementary material).

We have provided a GitHub repository of the SFEEDS scripts running on Praat (Kitayama, 2024). A patent was applied for the SFEEDS algorithm in the Japanese Patent Office (Japanese Patent Application Number 2024-087318).

1. Dominant Spectrum Test

The Dominant Spectrum Test was designed to identify the dominant harmonics present in speech waveforms. If we considered the f_o component of the glottal sound (not speech sound radiated from lips), estimating f_o is relatively straightforward because the lowest-frequency harmonic component (f_o) always has the highest spectral intensity within the audible frequency range [Fig. 2(a)]. However, the sound that emitted from the lips undergoes modifications owing to resonance within the vocal tract, reverberations in the nasal cavity, and lip radiation. These alterations lead to the formation of formants whose frequency components may be amplified or attenuated (Dejonckere and Lebacq, 1996). The enhancement of the formants in the frequency component

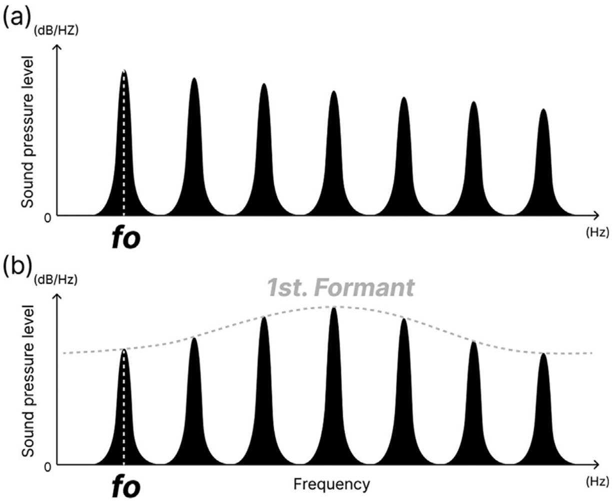

FIG. 2. (a) Schematic diagram of spectral waveform of speech originating from the glottal sound source. (b) Spectral waveform of speech filtered by the vocal tract and emitted through the lips.

makes it difficult to estimate f_o solely through spectral analysis [Fig. 2(b)].

Moreover, subharmonics appear in a spectrum at certain fractions, such as 1/2 or 1/3 of the f_o, under the vibration patterns with slight differences in the left and right vocal folds (Omori *et al.*, 1997). Given that subharmonics are periodic, much like the harmonic structures associated with vocal-fold vibration, existing methodologies have yet to fully address the subharmonics error, wherein subharmonics are erroneously identified as f_o (Fig. 3).

Consequently, it is essential to search for dominant harmonic structures within a short timeframe analysis. To begin this process, we first identified the peak with the highest spectral intensity within the 50–400 Hz range of the spectral waveform [Fig. 4(a)]. The low-frequency region was subsequently subdivided to identify spectral peaks within each frequency range [Fig. 4(b)]. On the basis of the assumption that subharmonics, background noise, and turbulence noise are negligible in comparison with the frequency components of f_o, the spectral peak with the most significant intensity relative to the highest peaks in the 50–400 Hz range was identified as the f_o candidate in the short-term analysis [Fig. 4(c)]. This process reduces the risk of false detection of spectral components other than f_o (e.g., subharmonics and environmental noise).

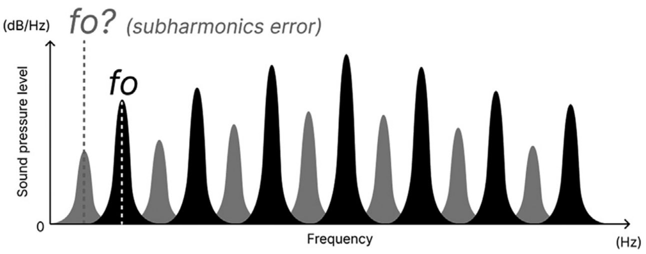

FIG. 3. Schematic diagram of spectral waveform of voice includes subharmonics. The spectral waveform of speech containing subharmonics presents a notable issue for traditional approaches, i.e., subharmonic error, for which regions containing subharmonics are inaccurately identified as the f_o.

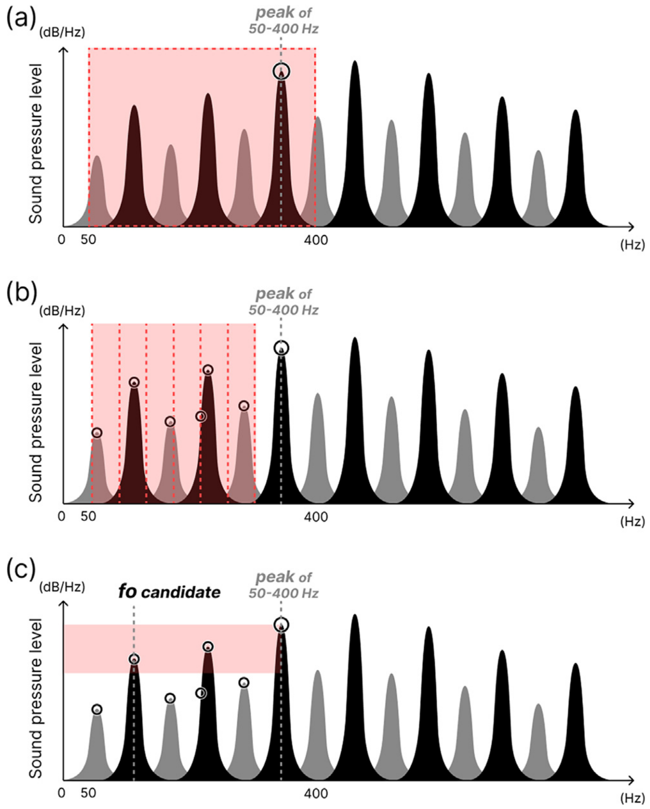

FIG. 4. (Color online) Search process for f_o candidates in the Dominant Spectrum Test. (a) Extract the highest spectral peaks between 50 and 400 Hz. (b) Subdivision of the low-frequency domain and extraction of spectral peaks within each designated frequency band. (c) The f_o candidate is determined by identifying the spectral peak with the lowest frequency that exhibits a spectral intensity above a certain threshold when compared to the spectral peak obtained in (a).

However, the f_o in these dominant harmonics structure may mis-detect the instantaneous enhancement of subharmonics or chaotic noise as f_o. Therefore, the following Sequential Spectrum Test is used to supplement the f_o false detection.

2. Sequential Spectrum Test

This algorithm focuses on the f_o transition across the time series of voice samples, counteracting f_o false positives identified in the Dominant Spectrum Test and minimizing the f_o-estimation error within the same phrase. The algorithm leverages the observation that f_o gradually varies during reading of sentences, ensuring the uniformity of f_o shifts over time. When a frequency peak possessing comparable spectral intensity and frequency appears in the immediately succeeding temporally adjacent frame to the f_o determined in a given frame, it is preferentially chosen as the f_o for the following frame [Fig. 5(a)]. Conversely, at junctures where the temporal continuity of f_o is disrupted due to sentence onset and offset, abrupt pitch transitions, alterations in vibration modes, and similar factors, the continuity in the Sequential Spectrum Test is reset, and the f_o estimate

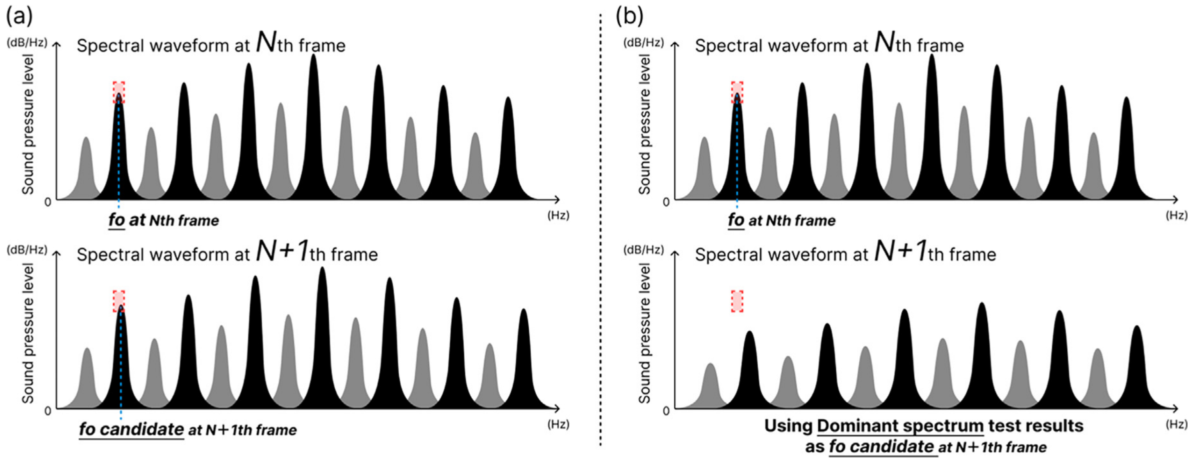
(a)
Spectral waveform at Nth frame
(dB/Hz)
Sound pressure level
0
fo at Nth frame
(Hz)
Spectral waveform at N+1th frame
fo candidate at N+1th frame
(b)
Spectral waveform at Nth frame
fo at Nth frame
Spectral waveform at N+1th frame
Using Dominant spectrum test results
as fo candidate at N+1th frame

FIG. 5. (Color online) The process of selecting f_o candidates in the Sequential Spectrum Test. (a) If a frequency peak exhibiting similar spectral intensity and frequency is present in the N + 1th frame, based on the f_o identified in the Nth frame, it is selected as the f_o for the N + 1th frame. (b) If a frequency peak with an approximate spectral intensity and frequency is absent, the f_o candidate determined by the Dominant Spectrum Test is used.

derived from the Dominant Spectrum Test is chosen [Fig. 5(b)].

C. Ground truth for f_o

In general, the accuracy of f_o-estimation algorithms has been evaluated by comparing them to the ground truth of f_o defined by specific methodologies, such as pitch perception. However, Bechtold (2021) emphasized the ambiguity of using pitch as the ground truth for f_o when assessing the accuracy of these algorithms. Furthermore, pitch tends to vary significantly in CS samples. Therefore, in this study, we focused on the narrow-band spectrogram of speech, rather than pitch perception, to establish the ground truth of f_o.

With regard to the voice samples used for evaluation, studies focusing on the f_o-estimation of hoarseness with subharmonics were very limited and were restricted to evaluation with SV (Camacho and Harris, 2008). To enhance versatility, it is imperative to extend evaluations beyond SV to include text reading (Zraick *et al.*, 2005; Maryn *et al.*, 2010). The present study therefore established the ground truth of f_o as an evaluation criterion in speech samples containing hoarseness consisting of CS and SV.

Figure 6 presents a spectrogram of a concatenated CS and SV sample from a participant in this study. In this spectrogram, several segments with enhanced spectral intensity can be observed between harmonic frequency bands, particularly at the onset and end of sentences, as well as in the middle of the SV. Even when subharmonic signals are present, subharmonic bands are easily distinguishable from f_o bands in the spectrogram. Consequently, two laryngologists (K.H. and I.K.) manually and visually determined the ground truth of f_o by identifying the lowest frequency band among the dominant harmonic complexes in each spectrogram. For this process, the spectrogram was binarized, and the identified f_o band was extracted using ImageJ software (National Institutes of Health, Bethesda, MD), a free software for image analysis (Fig. 7).

D. Accuracy validation of algorithms

Following the ground truth determination, the accuracy of the algorithms was examined by use of the "f_o concordance rate." Figure 8 illustrates the calculation of the f_o concordance rate. In this process, the contour of f_o (red line) estimated using each f_o-estimation algorithm was overlaid on the binarized spectrographic plane of the ground truth f_o [Fig. 8(a), bottom] and overlapped segments were subsequently extracted [Fig. 8(a), top]. The f_o concordance rate indicated a percentage of the duration in which the estimated f_o contour overlapped with the f_o band of the ground truth [Fig. 8(b)]. The f_o contour was delineated in Praat or MATLAB (version 2022a, The MathWorks, Inc., Natick, MA) and the f_o concordance rate was calculated by performing luminance analysis in ImageJ software. The algorithms examined in this study are presented in Table I.

These algorithms were selected due to their widespread and well-regarded usage, as well as reports of superior

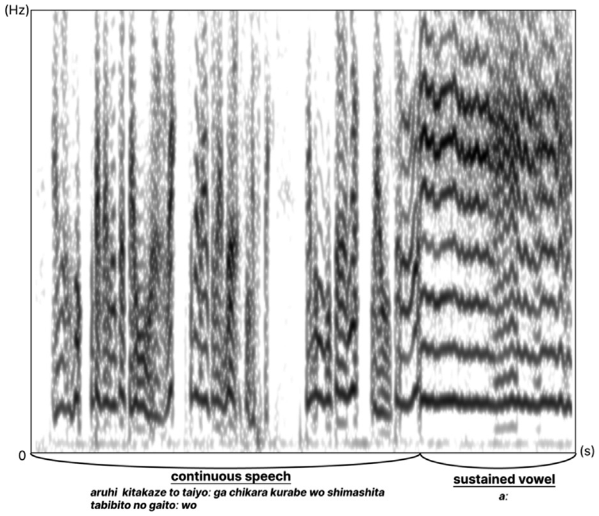

FIG. 6. Speech sample consists of CS and SVs.

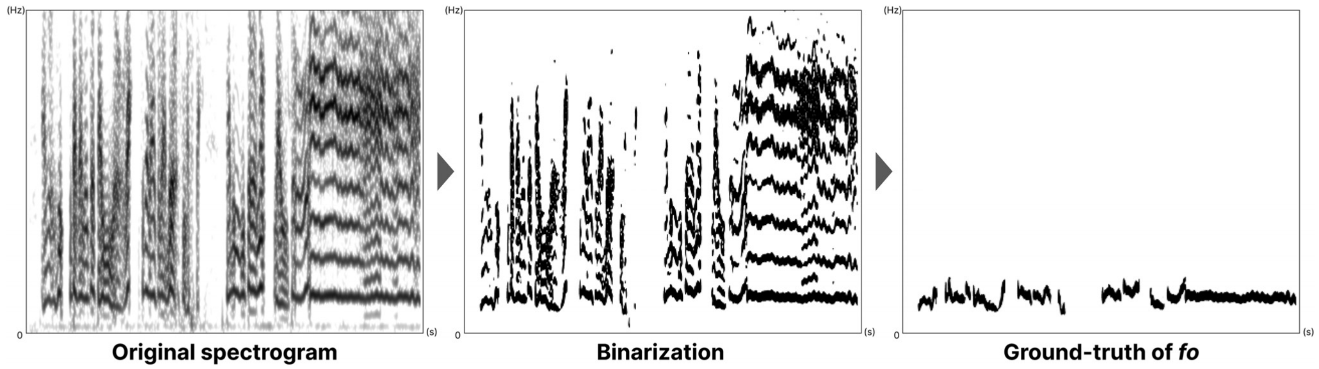

FIG. 7. Process of extracting the ground truth of f_o from a spectrogram. The ground truth of f_o was determined as the lowest frequency within the most dominant harmonic complex observed on the spectrogram.

performance in f_o-estimation accuracy compared with alternative approaches (Camacho and Harris, 2008; Ba *et al.*, 2012). The peak-picking and zero-crossing methods were excluded from the analysis because their accuracy has been reported to be lower for AC and CC (Titze and Liang, 1993). Additionally, the cepstrum method was excluded as previous studies have concluded that its f_o estimation accuracy is low (Sukhostat and Imamverdiyev, 2015). All parameter settings for the algorithms were set to the default values recommended by their developers.

E. Statistical analysis

First, the Shapiro–Wilk normality test showed that the agreement between the estimated f_o calculated by each f_o-estimation algorithm and the ground truth violated the normality assumption ($p < 0.001$), necessitating a nonparametric test. Therefore, Wilcoxon's signed-rank test with Bonferroni's correction and effect size of Cliff's delta (Cliff, 1996) was used to compare the f_o-estimation accuracy of SFEEDS with those of the other f_o-estimation algorithms. The magnitude of the effect sizes is assessed using of the thresholds provided by Romano *et al.* (2006) (i.e., $|d| < 0.147$ "negligible," $|d| < 0.33$ "small," $|d| < 0.474$ "medium," otherwise "large" differences. For the multiple comparisons, statistical significance was set at $p < 0.0125$. Statistical analyses were performed by use of R version 4.4.1 (R Core Team, Vienna, Austria) for Cliff's delta and JMP version 16.0.0 software package (SAS Institute, Cary, NC) for all other statistical analyses. Except the multiple comparison, all results were considered to be statistically significant at $p < 0.05$.

III. RESULTS

A. Distribution of the accuracy of f_o estimation and degree of hoarseness

Figure 9 presents the f_o concordance rates indicating the accuracy of f_o estimation for all 454 concatenated voice samples across different degree of hoarseness: G_{total}, R_{total}, and B_{total}. These plots illustrate the extent to which the accuracy of f_o estimation is influenced by the degree of hoarseness for each algorithm. In all plots, the smoothed spline curves indicate a general trend where the f_o concordance rates decrease as the degree of hoarseness increases. Despite this trend, the curves for SFEEDS exhibit the smallest decline among the algorithms evaluated. The density distribution in the plots shows that SWIPE′ and SFEEDS have smaller areas, whereas BaNa, raw CC, and filtered AC have larger areas, in that order. Notably, the BaNa plots tended to cluster in the lower-left quadrant, indicating a lower f_o concordance rate even at lower hoarseness levels, which differs from the other algorithms.

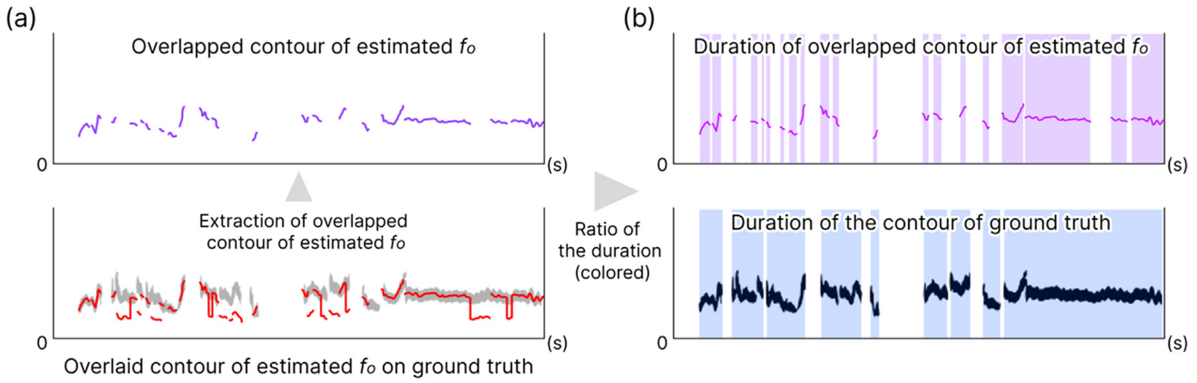
(a)
Overlapped contour of estimated fo
0
(s)
Extraction of overlapped
contour of estimated fo
0
(s)
Overlaid contour of estimated fo on ground truth
Ratio of
the duration
(colored)
(b)
Duration of overlapped contour of estimated fo
0
(s)
Duration of the contour of ground truth
0
(s)

FIG. 8. (Color online) Procedure to calculate the accuracy of the algorithms. (a) Extraction of the overlapped contour of the estimated f_o. (b) Calculation of the concordance rate of f_o.

TABLE I. f_o estimation algorithm examined in the analysis.

Algorithm	Reference	URL to download code or software
Filtered AC Raw CC	Boersma	https://www.fon.hum.uva.nl/praat/
SWIPE′	Camacho and Harris	https://github.com/SageBionetworks/PDScores/blob/master/bridge_ufb%20(for%20code%20generation)/swipe.m
BaNa	Ba and Yang	https://hajim.rochester.edu/ece/sites/wcng//project_bridge.html

B. Comparison of f_o concordance rates between SFEEDS and the other algorithms

Table II presents the percentiles of the f_o concordance rates for the examined algorithms across all 454 samples, both with and without perceptual hoarseness. The f_o concordance rate was significantly higher for SFEEDS than for the other algorithms, demonstrating the best performance. In particular, the effect size indicated a large difference in the f_o concordance rates between SFEEDS and filtered AC, raw

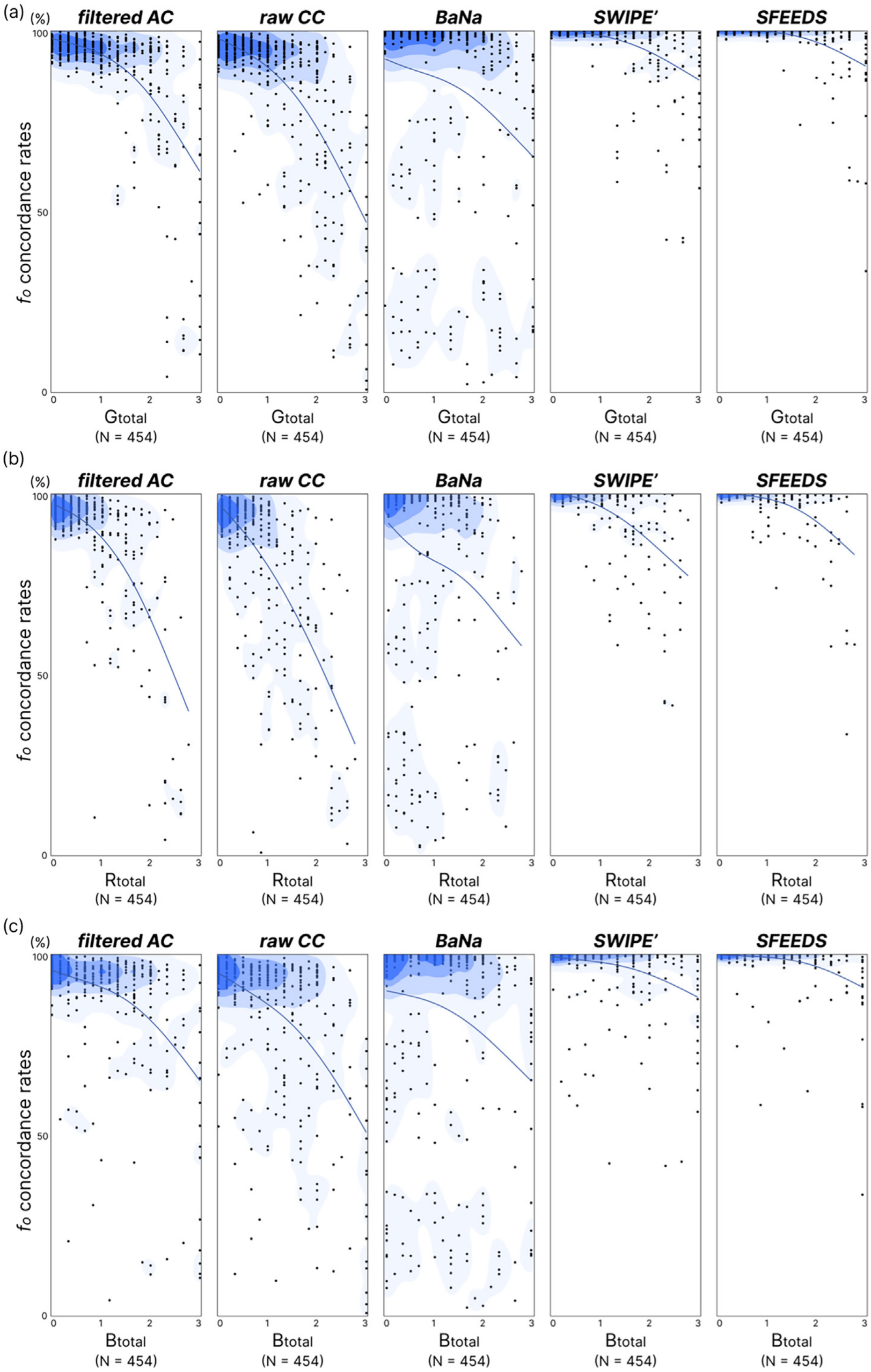

FIG. 9. (Color online) (a)–(c) Scatter plots illustrating the relationship between the f_o concordance rate and the degrees of G_{total}, $R_{total,}$, and B_{total}, respectively. Smoothed spline curves and distribution density are also shown in the plots.

TABLE II. Comparison of the accuracy of SFEEDS and the examined algorithms in all voice samples ($n = 454$).

	Filtered AC	Raw CC	BaNa	SWIPE′	SFEEDS
Median	94.8	93.9	98.8	99.7	99.9
Interquartile range	5.18	10.45	11.26	0.82	0.45
Wilcoxon’s signed-rank test, p value	<0.0001	<0.0001	<0.0001	<0.0001	—
Effect size	0.871 (large)	0.896 (large)	0.709 (large)	0.231 (small)	—

CC, or BaNa, whereas a small (but non-negligible) difference was observed between SFEEDS and SWIPE′. SFEEDS and SWIPE′ algorithms showed similar and exceptionally high median values. However, the interquartile range (IQR) of SFEEDS was approximately half that of SWIPE′, indicating greater accuracy.

Table III shows the results limited to 119 voice samples with $G_{total} < 0.5$, categorized as hoarseness-free. In this subset, the median f_o concordance rate for each algorithm was higher than that for the full set of 454 samples. The comparison trends between algorithms were consistent with the overall results, but SFEEDS again demonstrated an IQR less than half that of SWIPE′, with a small but significant difference between the two algorithms.

Tables IV, V, and VI present the results for voice samples rated as hoarse, with G_{total}, R_{total}, and $B_{total} > 0.5$, respectively. Across all types of hoarseness, SFEEDS exhibited the highest f_o concordance rates. As in previous comparisons, SFEEDS and SWIPE′ were very similar. However, for R_{total}, SFEEDS exhibited a markedly larger lower quartile (25th percentile), resulting in an IQR less than one third that of SWIPE′. Effect sizes confirmed that SFEEDS had large differences compared with filtered AC, raw CC, and BaNa, whereas the difference between SFEEDS and SWIPE′ remained small but noteworthy.

IV. DISCUSSION

The results of the degree of hoarseness and the distribution map of f_o concordance rate showed that the accuracy of f_o-estimation algorithms decreased with advanced hoarseness in the auditory-perceptual judgment for all algorithms examined (Fig. 9). In addition, the degree of roughness tended to make f_o estimation particularly difficult for hoarseness, suggesting that the presence of subharmonics affected the accuracy of estimation. A comparison of the f_o concordance rate showed that SFEEDS achieved the best accuracy in estimating f_o among all the other algorithms (Table II).

In the no-hoarseness group (Table III), SFEEDS had estimation accuracy superior to that of all other algorithms. In contrast, BaNa had the lowest estimation accuracy. BaNa was developed to improve the f_o concordance rate under environmental noise and enables good estimation even for voice samples with a low SNR. However, there were fewer than 10 speakers included in their study, suggesting that BaNa might be less robust to various pitches and voice-quality variations contaminated with roughness or breathiness (Ba *et al.*, 2012; Sukhostat and Imamverdiyev, 2015). Also, in the analysis using the samples with hoarseness (Table IV), the f_o concordance rate of SFEEDS was higher than that in all the other algorithms regardless of hoarseness type, and raw CC had the lowest f_o-estimation accuracy. The high f_o-estimation accuracy of SWIPE′, which reduces subharmonic errors using only first- and prime-order harmonics, was also shown to be accurate. However, in the analysis of distribution plots and effect-size comparisons for rough voices, the performance of SFEEDS was superior to that of SWIPE’. Consequently, SFEEDS demonstrated robustness superior to that of the other evaluated algorithms, not only in the analysis of non-dysphonic voices but also in the detection of hoarse voices, particularly in the case of rough voices.

A. Novelty of SFEEDS

Existing f_o-estimation methods have been developed with the intention of extracting speech segments when a recorded voice is contaminated with environmental noises. In other words, they were developed to separate periodic waveforms from nonperiodic environmental noise, which have achieved a high degree of accuracy. However, subharmonics are “sub-periodic noise,” as confirmed by spectrograms, and it remains difficult when using conventional methods to accurately distinguish dominant f_o from subharmonics. Therefore, a method similar to the way human visually distinguish between the f_o and subharmonics in a spectrogram should be incorporated into the SFEEDS algorithm. One such test is the Dominant Spectrum Test. In this test, the spectral peak within the 50–400 Hz range was used as a reference, and the lowest-frequency peak with

TABLE III. Comparison of the accuracy of SFEEDS and examined algorithms in hoarseness-free voice samples ($n = 119$).

	Filtered AC	Raw CC	BaNa	SWIPE′	SFEEDS
Median	96.4	95.8	99.3	99.9	99.9
Interquartile range	2.96	2.88	0.96	0.35	0.16
Wilcoxon’s signed-rank test, p value	<0.0001	<0.0001	<0.0001	0.0012	—
Effect size	0.982 (large)	0.994 (large)	0.799 (large)	0.254 (small)	—

TABLE IV. Comparison of the accuracy of SFEEDS and the examined algorithms in voice samples of $G_{total} > 0.5$ ($n = 335$).

	Filtered AC	Raw CC	BaNa	SWIPE′	SFEEDS
Median	94.2	92.2	98.3	99.7	99.9
Interquartile range	7.63	21.73	22.16	1.47	0.60
Wilcoxon's signed-rank test, p value	<0.0001	<0.0001	<0.0001	<0.0001	—
Effect size	0.848 (large)	0.878 (large)	0.694 (large)	0.231 (small)	—

sufficiently large spectral intensity compared to this reference was extracted as the f_o candidate.

The Dominant Spectrum Test focuses solely on spectral intensity without accounting for spectral periodicity, which can lead to subharmonic errors. Furthermore, the Sequential Spectrum Test assumes that f_o is more likely to maintain continuity than subharmonics. By combining these two tests and evaluating the possibility of subharmonic errors in each frame, f_o estimation can be performed with high accuracy, even in cases of complex spectral shapes that include subharmonics.

The Dominant Spectrum Test is based on the assumption that f_o in the low-frequency range has a sufficiently higher spectral intensity than that of subharmonics; however, if environmental noise equivalent to f_o is included in the low-frequency range, the estimation accuracy is expected to decline. Therefore, as mentioned in the Sec. II, this test should be used for the analysis of recorded speech in a sound-proof environment so that a sufficient SNR can be obtained.

Regarding algorithms that take into account the temporal continuity of f_o, there is a f_o-estimation method that implements the Viterbi algorithm (van Alphen and Van Bergem, 1989), which selects the best pitch candidate for each segment by finding the least-cost path through all segments (Boersma, 1993; Ba *et al.*, 2012). In contrast, the Sequential Spectrum Test in SFEEDS is an algorithm that considers the continuity of spectral intensity (dB) and spectral frequency (Hz) between frames, regardless of path length, and has a fundamentally different purpose. SFEEDS is a f_o-estimation method developed to separate subharmonics from f_o, which is linked to vocal vibration frequencies, and it does not guarantee accuracy in separating overtone structures from environmental noise that is not periodic.

As stated by Bechtold (2021), there is no f_o-estimation algorithm that can handle all types of signals and noises and satisfies the trade-off between arithmetic time and estimation performance; SFEEDS is also affected by that trade-off.

B. Ground truth of f_o definition

To verify the usefulness of the f_o-estimation method, it is necessary to define the ground truth of f_o for comparison. Commonly used methods include electroglottography (EGG) and those that use human pitch perception. However, there are various problems when they are used as the ground truth of f_o. Therefore, it was necessary to develop the new method to define the ground truth of f_o.

1. Problems with EGG

EGG is an excellent tool for non-invasive and indirect estimation of the regularity of vocal-fold vibration and the relative extent of the vocal-fold contact area including vertical direction. However, Bechtold (2021) concluded that EGG recordings showed pitch doubling, which was less pronounced in voice recordings, and that the EGG-based f_o ground truth was not suitable for the pitch detection algorithm (Bechtold, 2021).

2. Problems with the human sense of pitch

Human pitch perception only involves sensory pitch and not physical pitch. Human perception (Hess, 2012) is logarithmic, and lower pitches cannot be detected more accurately than higher pitches (Sukhostat and Imamverdiyev, 2015). As reported by Bechtold (2021), pitch sensation is not consistent enough to be used as the ground truth of f_o. In particular, it is impossible to accurately define f_o, which changes dynamically in a short period of time, in a sample of text read aloud with a fluctuating f_o based solely on a person's pitch perception. Indeed, studies using pitch sensation as the ground truth of f_o are limited to SV samples (Camacho and Harris, 2008; Anand *et al.*, 2021).

Therefore, we used the narrow-band spectrogram, a powerful tool for analyzing independent frequency bands. Even for voice samples containing multiple overtone structure complexes, such as those represented by subharmonics, it is easy to distinguish subharmonics from f_o on the

TABLE V. Comparison of the accuracy of SFEEDS and the examined algorithms in voice samples of $R_{total} > 0.5$ ($n = 218$).

	Filtered AC	Raw CC	BaNa	SWIPE′	SFEEDS
Median	92.8	86.2	96.3	99.5	99.8
Interquartile range	17.77	35.61	28.11	5.27	1.60
Wilcoxon's signed-rank test, p value	<0.0001	<0.0001	<0.0001	<0.0001	—
Effect size	0.795 (large)	0.847 (large)	0.643 (large)	0.193 (small)	—

TABLE VI. Comparison of the accuracy of SFEEDS and examined algorithms in voice samples of $B_{total} > 0.5$ ($n = 282$).

	Filtered AC	Raw CC	BaNa	SWIPE′	SFEEDS
Median	94.0	91.4	98.2	99.6	99.8
Interquartile range	7.92	26.62	20.73	1.93	0.75
Wilcoxon's signed-rank test, p value	<0.0001	<0.0001	<0.0001	<0.0001	—
Effect size	0.841 (large)	0.878 (large)	0.682 (large)	0.243 (small)	—

spectrogram. The reason for this is that subharmonics are unstable elements with a relatively low spectral intensity and a shorter duration than those of f_o, which can be clearly determined on the spectrogram. Therefore, in this study, we defined f_o on the spectrogram as the ground truth and performed a comparison test by plotting the estimated f_o on the spectrogram, which was calculated by using the f_o-estimation algorithm including SFEEDS.

C. Issues with the corpus for evaluation

To improve the measurement accuracy of the f_o-estimation algorithm, it is essential to have an appropriate voice-recording corpus and accurate ground truth; however, the conventional method is fraught with various problems. Among the publicly available voice-recording corpora, those that include the above-mentioned EGG and ground truth of f_o using human pitch sense are limited to nondysphonic speech (Bagshaw *et al.*, 1993; Garofolo *et al.*, 1993; Bagshaw, 1994; Plante *et al.*, 1995; Pirker *et al.*, 2011). In other words, there are no available voice samples that include a large number of hoarse voices, such as subharmonics, and that also include information on the ground truth of f_o. Furthermore, the corpus of voice recordings used in the previous validation of f_o-estimation algorithms mainly comprised nondysphonic voices, so it was impossible to evaluate the robustness of f_o estimation under various changes in hoarseness levels.

To create a corpus that clearly differentiates between complex f_o transitions and subharmonics during the oral reading of texts, we used a total of 454 recordings from a dataset used in the previous study (Hosokawa *et al.*, 2019b). As mentioned in Sec. II, this speech corpus consists of SV and CS samples and contains a large number of pathological as well as non-dysphonic voices. Therefore, the present research is the first f_o-estimation algorithm that enables robust f_o estimation regardless of the degree of hoarseness.

The development of SFEEDS has improved the f_o-estimation accuracy in rough voices, including subharmonics, which has been difficult to estimate accurately. Future studies will aim to use SFEEDS to quantify subharmonics and roughness, which has been considered difficult. Furthermore, SFEEDS is expected to achieve accurate f_o estimation for special singing voices, including subharmonics, such as death voices. SFEEDS is also expected to be implemented in acoustic software, such as filtering and pitch adjustment, for special singing methods, which have been considered difficult.

D. Study limitations

The analysis characteristics of Praat limited SFEEDS to f_o estimation in 6-Hz increments because a spectral analysis with a frame length of 0.1 s at the default setting of SFEEDS requires a frequency bandwidth of ≤ 5.38 Hz for the spectrum. In rough voices, estimation can be difficult in chaotic waveforms in which the vibration modes are not synchronized (Titze, 1995) and in subharmonics, which are not periodic and are completely independent of harmonic components, which is why it is difficult to separate them from environmental noises. It may also be impossible to estimate f_o accurately if the harmonic structure is not sufficiently large relative to the noise, such as in a highly atonic hoarse voice. Because the evaluation is based on samples of SV and text reading at rest, it may be necessary to adjust the parameters for voice samples with dynamic f_o transitions, such as singing or speech with emotion. The algorithm is currently limited to validation in the Japanese language, so validation in multiple languages will be required in the future.

V. CONCLUSION

The most significant outcomes of this study are as follows: We developed a novel algorithm for estimating f_o of speech. This algorithm, which includes a method for estimating the dominant harmonic structure and considers temporal variations in f_o, allows SFEEDS to substantially reduce subharmonic errors, achieving greater accuracy compared to conventional f_o-estimation methods. Additionally, by defining the ground truth of f_o on the spectrogram, we succeeded in establishing the ground truth of f_o, including subharmonics, which has been challenging to define using, e.g., human pitch perception or EGG. Future research should focus on developing acoustic analysis parameters that can accurately detect subharmonics.

SUPPLEMENTARY MATERIAL

See the supplementary material for disease breakdown of voice samples, specific details of SFEEDS and supplementary information about comparisons between the algorithms.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant No. JP21K16842.

AUTHOR DECLARATIONS

Conflict of Interest

All of the authors declare that they have no conflicts of interest in association with this study.

Ethics Approval

This study was conducted in accordance with the Helsinki Declaration of 1975 and its amendments as well as with the laws and regulations of Japan. This study was approved by the institutional review boards of Osaka University Hospital, Osaka International Medical & Science Center, and Kuma Hospital (Nos. 15497, 568, and 20120614-4).

DATA AVAILABILITY

The data table that supports the findings of this study is available on request from the corresponding author. The raw data are not publicly available due to privacy protection concerns.

Aichinger, P., Roesner, I., Schneider-Stickler, B., Leonhard, M., Denk-Linnert, D. M., Bigenzahn, W., Fuchs, A. K., Hagmuller, M., and Kubin, G. (**2017**). "Towards objective voice assessment: The diplophonia diagram," J. Voice **31**, 253.e17–253.e26.

Anand, S., Kopf, L. M., Shrivastav, R., and Eddins, D. A. (**2021**). "Using pitch height and pitch strength to characterize type 1, 2, and 3 voice signals," J. Voice **35**, 181–193.

Awan, S. N., and Awan, J. A. (**2020**). "A two-stage cepstral analysis procedure for the classification of rough voices," J. Voice **34**, 9–19.

Ba, H., Yang, N., Demirkol, I., and Heinzelman, W. (**2012**). "BaNa: A hybrid approach for noise resilient pitch detection," in *2012 IEEE Statistical Signal Processing Workshop*, pp. 369–372.

Bagshaw, P. C. (**1994**). "Automatic prosodic analysis for computer aided pronunciation teaching," Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.

Bagshaw, P. C., Hiller, S. M., and Jack, M. A. (**1993**). "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching," in *Proc. EUROSPEEECH' 93*, pp. 1003–1006.

Baken, R. J. (**1987**). *Clinical Measurement of Speech and Voice* (Taylor & Francis, London).

Barsties, B., and Maryn, Y. (**2016**). "External validation of the Acoustic Voice Quality Index version 03.01 with extended representativity," Ann. Otol. Rhinol. Laryngol. **125**, 571–583.

Bechtold, B. (**2021**). *Pitch of Voiced Speech in the Short-Time Fourier Transform: Algorithms, Ground Truths, and Evaluation Methods* (Universität Oldenburg, Oldenburg, Germany).

Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences (Amsterdam)*, pp. 97–110.

Boersma, P. (**2001**). "Praat, a system for doing phonetics by computer," Glot. Int. **5**, 341–345.

Camacho, A., and Harris, J. G. (**2008**). "A sawtooth waveform inspired pitch estimator for speech and music," J. Acoust. Soc. Am. **124**, 1638–1652.

Cavalli, L., and Hirson, A. (**1999**). "Diplophonia reappraised," J. Voice **13**, 542–556.

Cliff, N. (**1996**). *Ordinal Methods for Behavioral Data Analysis* (Psychology Press, London).

DeBodt, M. S., Wuyts, F. L., VandeHeyning, P. H., and Croux, C. (**1997**). "Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality," J. Voice **11**, 74–80.

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

Dejonckere, P., and Lebacq, J. (**1996**). "Acoustic, perceptual, aerodynamic and anatomical correlations in voice pathology," Otorhinolaryngol. Relat. Spec. **58**, 326–332.

Dejonckere, P., Schoentgen, J., Giordano, A., Fraj, S., Bocchi, L., and Manfredi, C. (**2011**). "Validity of jitter measures in non-quasi-periodic voices. Part I: Perceptual and computer performances in cycle pattern recognition," Logoped. Phoniatr. Vocol. **36**, 70–77.

Deliyski, D. D. (**1993**). "Acoustic model and evaluation of pathological voice production," in *Third European Conference on Speech Communication Technology*.

Deliyski, D. D., Shaw, H. S., and Evans, M. K. (**2005**). "Adverse effects of environmental noise on acoustic voice quality measurements," J. Voice **19**, 15–28.

Deliyski, D. D., Shaw, H. S., Evans, M. K., and Vesselinov, R. (**2006**). "Regression tree approach to studying factors influencing acoustic voice analysis," Folia Phoniatr. Logop. **58**, 274–288.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., and Pallett, D. (**1993**). "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," NIST Speech Disc. 1, Vol. 93. NASA STI/Recon Technical Report.

Hess, W. (**1983**). "Time-domain pitch determination," in *Pitch Determination Speech Signals: Algorithms Devices* (Springer-Verlag, Berlin), pp. 152–301.

Hess, W. (**2012**). *Pitch Determination of Speech Signals: Algorithms and Devices* (Springer Science & Business Media, New York).

Hosokawa, K., Barsties, B., Iwahashi, T., Iwahashi, M., Kato, C., Iwaki, S., Sasai, H., Miyauchi, A., Matsushiro, N., Inohara, H., Ogawa, M., and Maryn, Y. (**2017**). "Validation of the Acoustic Voice Quality Index in the Japanese language," J. Voice **31**, 260.e1–260.e9.

Hosokawa, K., von Latoszek, B. B., Ferrer-Riesgo, C. A., Iwahashi, T., Iwahashi, M., Iwaki, S., Kato, C., Yoshida, M., Umatani, M., and Miyauchi, A. (**2019b**). "Acoustic breathiness index for the Japanese-speaking population: Validation study and exploration of affecting factors," J. Speech Lang. Hear. Res. **62**, 2617–2631.

Hosokawa, K., von Latoszek, B., Iwahashi, T., Iwahashi, M., Iwaki, S., Kato, C., Yoshida, M., Sasai, H., Miyauchi, A., Matsushiro, N., Inohara, H., Ogawa, M., and Maryn, Y. (**2019a**). "The Acoustic Voice Quality Index version 03.01 for the Japanese-speaking population," J. Voice **33**, 125.e1–125.e12.

Kitayama, I. (**2024**). The scripts of SFEEDS, https://github.com/LarynxOsaka (Last viewed November 29, 2024).

Kitayama, I., Hosokawa, K., Iwaki, S., Yoshida, M., Miyauchi, A., Ogawa, M., and Inohara, H. (**2023**). "Validation of subharmonics quantification using two-stage cepstral analysis," J. Voice (published online).

Kwon, O.-W., Chan, K., Hao, J., and Lee, T.-W. (**2003**). "Emotion recognition by speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.

Latoszek, B. B. V., Maryn, Y., Gerrits, T., and De Bodt, M. (**2017**). "The Acoustic Breathiness Index (ABI): A multivariate acoustic model for breathiness," J. Voice **31**, 511.e11–511.e27.

Maryn, Y., Corthals, P., van Cauwenberge, P., Roy, N., and De Bodt, M. (**2010**). "Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels," J. Voice **24**, 540–555.

Noll, A. M. (**1964**). "Short-time spectrum and 'cepstrum' techniques for vocal-pitch detection," J. Acoust. Soc. Am. **36**, 296–302.

Noll, A. M. (**1967**). "Cepstrum pitch determination," J. Acoust. Soc. Am. **41**, 293–309.

Omori, K., Kojima, H., Kakani, R., Slavit, D. H., and Blaugrund, S. M. (**1997**). "Acoustic characteristics of rough voice: Subharmonics," J. Voice **11**, 40–47.

Pirker, G., Wohlmayr, M., Petrik, S., and Pernkopf, F. (**2011**). "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Interspeech*, pp. 1509–1512.

Plante, F., Meyer, G., and Ainsworth, W. (**1995**). "A pitch extraction reference database," Children **8**, 30–50.

Rabiner, L., Cheng, M., Rosenberg, A., and McGonegal, C. (**1976**). "A comparative performance study of several pitch detection algorithms," IEEE Trans. Acoust. Speech Signal Process. **24**, 399–418.

Ross, M., Shaffer, H., Cohen, A., Freudberg, R., and Manley, H. (**1974**). "Average magnitude difference function pitch extractor," IEEE Trans. Acoust. Speech Signal Process. **22**, 353–362.

Romano, J., Kromrey, J. D., Coraggio, J., and Skowronek, J. (**2006**). "Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys?," in *Annual Meeting of the Florida Association of Institutional Research*.

Sukhostat, L., and Imamverdiyev, Y. (**2015**). "A comparative analysis of pitch detection methods under the influence of different noise conditions," J. Voice **29**, 410–417.

Titze, I. (**1994**). "Fluctuations and perturbations in vocal output," *Principles of Voice Production* (Prentice Hall, Hoboken, NJ), pp. 209–306.

Titze, I. R. (**1995**). *Workshop on Acoustic Voice Analysis: Summary Statement* (National Center for Voice and Speech, Salt Lake City, UT).

Titze, I. R., and Liang, H. (**1993**). "Comparison of f_o extraction methods for high-precision voice perturbation measurements," J. Speech Lang. Hear. Res. **36**, 1120–1133.

van Alphen, P., and Van Bergem, D. (**1989**). "Markov models and their application in speech recognition," in *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, pp. 1–26.

van Latoszek, B., De Bodt, M., Gerrits, E., and Maryn, Y. (**2018**). "The exploration of an objective model for roughness with several acoustic markers," J. Voice **32**, 149–161.

Wang, C. (**2001**). *Prosodic Modeling for Improved Speech Recognition and Understanding* (Massachusetts Institute of Technology, Cambridge, MA).

Yamaguchi, H., Shrivastav, R., Andrews, M. L., and Niimi, S. (**2003**). "A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale," Folia Phoniatr. Logop. **55**, 147–157.

Zraick, R. I., Wendel, K., and Smith-Olinde, L. (**2005**). "The effect of speaking task on perceptual judgment of the severity of dysphonic voice," J. Voice **19**, 574–581.