2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Speech Separation via Harmonic Suppression in Multi-Speaker Conversations to Assist Individuals with Hearing Loss

Kai Ito[1], Yasuaki Ishikawa[2], and Taku Itami[3]

***Abstract*— It is difficult for deaf and hard-of-hearing people to obtain information from their hearing, particularly in group conversations where multiple speakers overlap. In this study, we propose a speech separation and recognition system that does not rely on a deep neural network but instead focuses on the removal of harmonic components. Specifically, we propose a method to extract the frequency components of one of the sounds from a mixed-gender audio signal by removing the harmonics of the other. The effectiveness of this system is evaluated by separating each individual voice from the mixed signal and measuring the recognition accuracy using an automatic speech recognition (ASR) system. We discuss the proposed method and validation results in terms of speech separation and recognition accuracy.**

I. INTRODUCTION

Approximately 430 million people worldwide currently suffer from disabling hearing loss, and this number is expected to exceed 700 million — 1 in every 10 people — by 2050 [1]. The percentage of people with hearing loss increases with age, and it is estimated that more than 25 % of people over the age of 60 have a disabling hearing loss [1]. In particular, individuals with sensorineural hearing loss face difficulty in correctly interpreting speech information as ”words” and communicating orally even if they can perceive it as ”sound” [2]. As a result, they experience significant barriers in verbal communication, especially in complex auditory environments such as group discussions and classroom interactions [2].

[1]Graduate School of Science and Technology, Aoyama Gakuin University, Kanagawa, Japan k.ito.sce@gmail.com
[2]Professor Department of Electrical Engineering and Electronics, School of Science and Technology, Aoyamagakuin University, Kanagawa, Japan yishikawa@ee.aoyama.ac.jp
[3]Professor Department of Electronics and Bioinfomatics School of Science and Technology, Meiji University, Kanagawa, Japan itami@meiji.ac.jp

Automatic speech recognition (ASR) systems have been proposed as a means to help people with hearing loss obtain auditory information in visual form [3]. Such systems can be beneficial for students in educational settings, enabling them to obtain useful lecture content visually [3]. However, the performance of conventional ASR systems tends to deteriorate in situations where multiple speakers are talking simultaneously, which remains a critical issue [4]. Against this background, visualizing multi-speaker conversations is of critical importance for people with hearing loss.

II. RELATED WORK

Various studies have been conducted on speech separation to address the problem of overlapping speech in multi-speaker environments. In the field of speech separation, Lutati et al. proposed a deep neural network (DNN) that trained with over 100 hours of speech data to separate overlapping speech signals [5]. However, real-time speech separation and recognition require rapid processing, making it essential to reduce computational cost and latency [5] [6]. Yabate et al. proposed a non-DNN approach that performs speech separation between musical instruments and vocals by masking harmonic components [7]. However, the computational cost remains an issue, since even this method requires training for 200 epochs [7]. These studies primarily evaluated separation performance using metrics such as Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Artifacts Ratio (SAR), and Signal-to-Residual Ratio (SRR), without assessing speech recognition accuracy [5] [7]. In contrast, Masuyama et al. developed a system that jointly trains speech separation and recognition models using DNNs, aiming to improve recognition accuracy in multi-speaker scenarios [8]. In this study, their evaluation using both SDR and Word Error Rate (WER) revealed that these

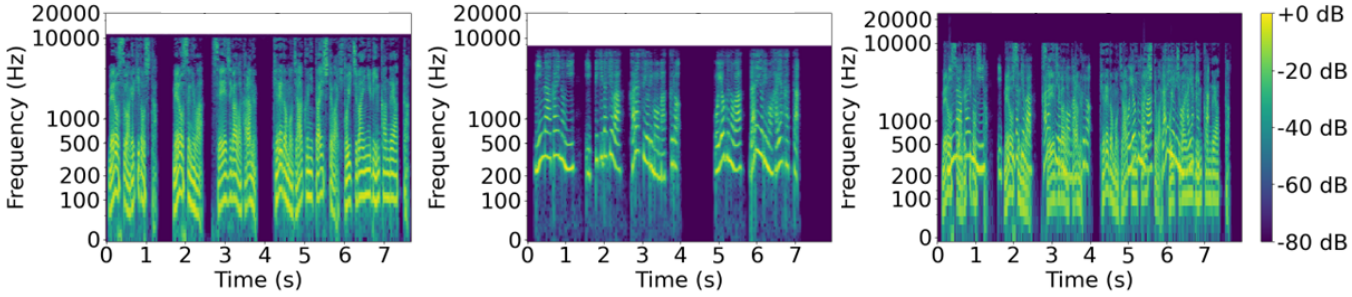

Fig. 1. Spectrogram of the sound source used. From left to right: male speech, female speech, and mixed audio

metrics do not always correlate positively, indicating that separation does not necessarily guarantee high recognition accuracy [8].

These findings highlight the need for separation approaches that directly consider ASR performance rather than relying solely on signal-level metrics.

III. Proposed Method

A. Harmonic-Based Speech Separation without a DNN

Fig. 1 shows the spectrograms of male speech, female speech, and their mixed audio signal, respectively. In these figures, the vertical axis represents frequency, the horizontal axis represents time, and the color indicates the intensity of each frequency component. When male and female voices are mixed, their respective frequency components interfere with one another, leading to a degradation in speech recognition accuracy. To address this issue, we developed a harmonic-based speech separation method that does not rely on DNN. In the proposed method, the mixed signal is first divided into 50 ms segments to capture local periodicity. For each segment, f_0 of the target speaker ' s clean speech is estimated. Based on the estimated f_0, frequency bands corresponding to integer multiples of f_0—i.e., "harmonics" are identified [9]. Then, a band-stop filter is applied to attenuate these harmonics in the mixed signal within the range of 0–10,000 Hz [10]. By reducing the harmonic components of the interfering speaker, the harmonic structure of the target speaker is relatively enhanced, thereby achieving speech separation.

B. Fundamental Frequency Estimation

The fundamental frequency (f_0) of each speaker was estimated from isolated speech audio using the "Autocorrelation Method" [11]. This approach was selected for its simplicity and robustness against noise, making it suitable for non-DNN-based signal processing. The estimation procedure is summarized as follows. First, for each segment, the direct current (DC) component was removed by subtracting the mean value, preventing bias in the correlation computation. Then, the autocorrelation function $R(\tau)$ was computed according to Equation (1):

$$R(\tau) = \sum_{n=0}^{N-1-\tau} x[n]x[n+\tau] \tag{1}$$

where $x[n]$ represents the speech signal, τ is the lag, and N is the number of samples in a segment. Next, to identify the fundamental period, the algorithm searches for the lag $\tau_{\max}$ that yields the highest peak of $R(\tau)$ within a predefined range corresponding to plausible human voice frequencies (typically 10–1000 Hz). This restriction helps avoid false detections due to subharmonics or high-frequency noise. Then, the fundamental frequency f_0 is then computed according to Equation (2):

$$f_0 = \frac{f_s}{\tau_{max}} \tag{2}$$

where f_s is the sampling frequency. In the implementation, the autocorrelation is calculated using NumPy ' s `correlate()` function, and the lag range is determined from the sampling rate (f_s) and the specified frequency limits. The resulting f_0 values are then used to determine the center frequencies of band-stop filters applied to the other speaker ' s signal for harmonic suppression.

C. Selective Attenuation for Female Voice Extraction

In general, the fundamental frequency (f_0) of adult male speakers ranges from 80 to 175 Hz, while that of adult female speakers ranges from 160 to 270 Hz [12]. Therefore, the harmonics of male speech often overlap with many of the harmonics of female speech. When applying the method described in Section III-A to attenuate male harmonics, this overlap may cause the unintended suppression of female harmonics, thereby leading to a loss of the target speech. To address this issue, we propose a selective harmonic attenuation method that limits the upper frequency bound of male harmonics to be attenuated during female voice extraction. Specifically, we set the upper limit of male harmonics to be attenuated from 0 Hz to f_0, $2f_0$, and $3f_0$ of the corresponding female speaker's estimated harmonics in each segment and the attenuation process was performed.

D. Data Processing

All data processing in this study was implemented using Python. The processing flow of the female speech extraction method proposed in Section III-C is illustrated in Fig. 2. This figure visualizes the entire processing pipeline, including segmentation, fundamental frequency estimation, and selective band-stop filtering for each segment. Since this method must adapt to time-varying fundamental frequencies, the audio signal is segmented into short time frames, and the target frequency bands are attenuated for each segment. Previous studies have suggested that a window size between 20 and 50 ms is appropriate for pitch analysis [13]. To ensure sufficient information for accurate estimation of the fundamental frequency f_0, this study adopted the upper limit of this range, i.e., a window size of 50 ms. For each segment, if the male single-source audio that was used as the target for harmonic removal was silent, no processing was performed. If speech was present, the male fundamental frequency f_{0B} was estimated using the autocorrelation method. If the female single-source audio was silent, the upper limit of harmonics to be removed (denoted as freq_limit) was set to 10,000 Hz. Otherwise, the female f_{0C} was estimated, and freq_limit was set to either f_{0C}, $2f_{0C}$, or $3f_{0C}$. After that, all harmonic components of the male speech, represented as $if_{0B}(i = 1, 2, 3, \dots)$, were attenuated using a band-stop filter in the range of $if_{0B}\pm$ the specified attenuation bandwidth, up to the freq_limit. After processing all segments, they were concatenated sequentially to reconstruct the audio file.

IV. Verification of Effectiveness

A. Experimental Environment

The experiment was conducted in a quiet indoor environment (ambient noise level: approximately 10.5 dB SPL)

using pre-recorded human speech to simulate a typical conversation with minimal background noise.

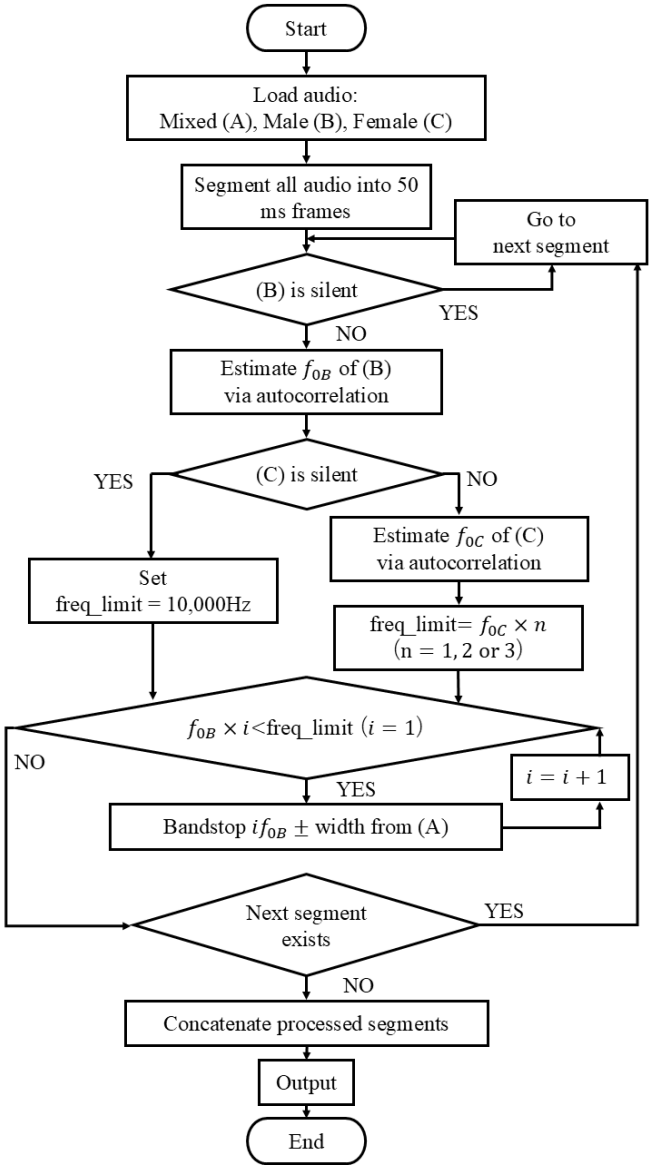

Fig. 2. Flowchart of the proposed harmonic suppression process for female speech extraction. The method segments the audio into 50 ms frames and adaptively attenuates male harmonics based on estimated fundamental frequencies. The frequency range for suppression is dynamically adjusted according to the presence of female speech and its f_0.

B. Experimental Setup and Materials

To evaluate the effectiveness of each method, the audio signals were played on a computer (X4-i5CMLABW11, MouseComputer Co., Ltd) and recorded using a smartphone microphone (Xperia 5 IV XQ-CQ44, SONY).

The audio sources used for evaluation consisted of male speech (50 characters, with an average fundamental frequency of 105.5 Hz) and female speech (39 characters, with an average fundamental frequency of 280.7 Hz). Spectrograms of the male speech, female speech, and their mixed audio are shown in Fig. 1.

Three ASR systems were used for evaluation:

- YYProbe (AISIN CORPORATION): designed for noisy environments
- Live Transcribe (Google): a cloud-based ASR by Google
- UD Talk (Shamrock Records, Inc.): designed for classroom and meeting support

This experimental setup is illustrated in Fig. 3.

Fig. 3. Experimental setup showing the PC playback and smartphone recording arrangement.

C. Evaluation Procedure and Parameters

The recognition performance was assessed by measuring the number of correctly transcribed characters using various ASR systems.Three types of audio were evaluated: single-source audio from male and female speakers, mixed audio combining both audios, and the speech extracted from mixed signal.

Harmonic components are known to have a certain frequency bandwidth [9]. Therefore, when attenuating each harmonic using a band-stop filter, we defined three levels of attenuation bandwidths: $if_0 \pm 25$ Hz, $if_0 \pm 50$ Hz, and $if_0 \pm 100$ Hz ($i = 1, 2, 3, \ldots$). This allowed us to examine how different attenuation bandwidths affect ASR performance.

In addition, when extracting female speech, we evaluated the effect of varying the upper limit of male harmonics to be attenuated. Specifically, we defined three levels based on the female speaker's estimated f_0: from 0 Hz to f_0, $2f_0$, and $3f_0$. Recognition accuracy was compared across these conditions.

V. Results

A. ASR Performance on Original and Mixed Audio

For each audio source and ASR system, speech recognition was conducted three times, and the average percentage of correctly recognized characters relative to the total number of characters was calculated as the recognition rate. The recognition rate r was computed using Equation (3), where

W denotes the total number of characters in the original transcription, and w_1, w_2, and w_3 represent the number of correctly recognized characters in each of three trials:

$$r = \frac{w_1 + w_2 + w_3}{3W} \times 100 \tag{3}$$

Table I and II show the recognition rates for single-speaker audio and mixed audio respectively. As shown in Table I, all ASR systems achieved a perfect recognition rate of 100% for single-speaker audio. In contrast, Table II indicates a significant degradation in recognition performance for mixed audio due to speech interference, with the lowest recognition rate for male speech was 9.33%, and for female speech, it was 8.55%, clearly demonstrating a substantial decline compared to single-speaker conditions.

TABLE I
RECOGNITION RATES FOR SINGLE-SPEAKER AUDIO[%]

Audio		Male Speech	Female Speech
ASR	YYProbe	100	100
	Live Transcribe	100	100
	UD Talk	100	100

TABLE II
RECOGNITION RATES FOR MIXED AUDIO[%]

Recognition Target		Male Speech	Female Speech
ASR	YYProbe	34.7	69.2
	Live Transcribe	9.33	8.55
	UD Talk	48.0	30.8

B. ASR Performance After Harmonic Removal

The ASR results after removing all harmonic components of either male or female speech from the mixed audio (within the 0–10,000 Hz frequency range) are shown in Tables III and IV, respectively. From Tables II and III, recognition rates improved when extracting male speech compared to the original mixed audio. Specifically, when using a band-stop width of $if_0 \pm 25$ Hz($i = 1, 2, 3, \dots$), YYProbe and UD Talk showed improvements of 34.0% and 36.0%, respectively. When the band-stop width was $if_0 \pm 50$ Hz($i = 1, 2, 3, \dots$), Live Transcribe showed an improvement of 12.7%. The spectrograms for the stopband width of ± 25 Hz and ± 50 Hz are shown in Fig. 4 which aso indicates that the fundamental frequency components of the male speech remained after applying the band-stop filter. On the other hand, as seen in Tables IV and II, the recognition rates for female speech after harmonic attenuation were lower than those for the original mixed audio. The spectrogram for the stopband width of ± 100 Hz is shown in Fig. 5. Fig. 5 reveals that, due to the low fundamental frequency of the male voice and the wide stopband width setting, the band-stop filtering suppressed not only the male harmonics but also a broad range of frequency components, including those associated with the female speech.

TABLE III
RECOGNITION RATES FOR EXTRACTED MALE SPEECH[%]

Stopband Width		± 25 Hz	± 50 Hz	± 100 Hz
ASR	YYProbe	68.7	63.3	35.3
	Live Transcribe	17.3	22.0	16.0
	UD Talk	84.0	79.3	38.7

TABLE IV
RECOGNITION RATES FOR EXTRACTED FEMALE SPEECH[%]

Stopband Width		± 25 Hz	± 50 Hz	± 100 Hz
ASR	YYProbe	7.33	8.00	5.33
	Live Transcribe	0.00	6.00	0.00
	UD Talk	0.67	6.00	6.00

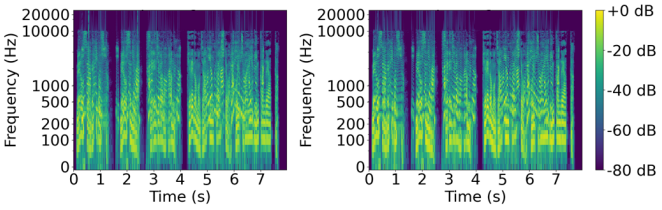

Fig. 4. Spectrograms of male speech extraction: from left to right, stopband width of ± 25 Hz and ± 50 Hz

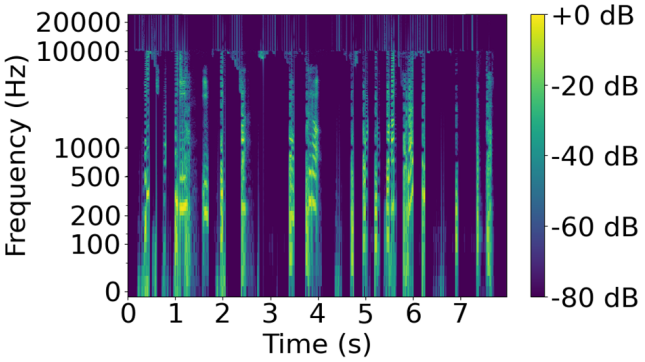

Fig. 5. Spectrograms of female speech extraction: stopband width of ± 100 Hz

C. Recognition Accuracy after Selective Attenuation of Male Harmonics Based on Female Harmonic Limits

Based on the proposed method described in Section III-C, we evaluated the recognition performance when the attenuation of male harmonics was limited up to the first, second, and third harmonics of the female speech. The results are presented in Tables V through VII respectively. From the comparison of Tables II and V through VII, it can be observed that the proposed selective attenuation approach improved recognition accuracy over the original mixed audio. Maximum improvements of 26.5%, 33.4%, and 35.0% were observed for YYProbe, Live Transcribe, and UD Talk, respectively. When the stopband width was set to ± 25 Hz around the fundamental and harmonic frequencies,

and the male harmonics were attenuated up to the second harmonic of the female speech, YYProbe achieved a high recognition rate of 95.7%. The spectrogram at that time is shown on the left side of Fig. 6. In contrast, as shown on the right side of Fig. 6, when the stopband width was widened to ± 100 Hz and male harmonics were attenuated up to the third harmonic, narrow spacing between harmonics caused adjacent band-stop filters overlap. This led to the suppression of a continuous frequency range rather than selective harmonic attenuation.

TABLE V
ASR (YYPROBE) RECOGNITION RATES WITH UPPER LIMIT ON REMOVED HARMONICS[%]

Stopband Width	± 25 Hz	± 50 Hz	± 100 Hz
Up to f_0	76.9	80.3	38.1
Up to $2f_0$	95.7	46.2	67.6
Up to $3f_0$	80.3	35.9	9.40

TABLE VI
ASR (LIVE TRANSCRIBE) RECOGNITION RATES WITH UPPER LIMIT ON REMOVED HARMONICS[%]

Stopband Width	± 25 Hz	± 50 Hz	± 100 Hz
Up to f_0	41.0	41.9	32.5
Up to $2f_0$	36.8	13.7	5.98
Up to $3f_0$	1.71	3.42	2.56

TABLE VII
ASR (UD TALK) RECOGNITION RATES WITH UPPER LIMIT ON REMOVED HARMONICS[%]

Stopband Width	± 25 Hz	± 50 Hz	± 100 Hz
Up to f_0	65.8	59.8	54.7
Up to $2f_0$	61.5	48.7	21.4
Up to $3f_0$	27.4	11.1	13.7

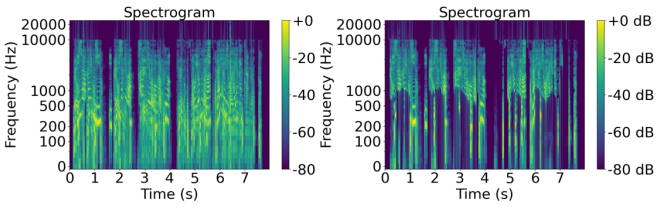

Fig. 6. Spectrograms of male speech after harmonic attenuation of female speech. From left to right: attenuation up to the second harmonic with a stopband width of ± 25 Hz, and attenuation up to the third harmonic with a stopband width of ± 100 Hz.

VI. DISCUSSION

In the case of male speech extraction, as seen in Fig. 1, the fundamental frequency (f_0) of male speech is generally lower than that of female speech. This implies that the mixed audio contains male harmonics that do not overlap with those of the female speech. As shown in Table III, when female harmonics were attenuated, the non-overlapping male harmonics became more prominent, which likely contributed to the improved recognition accuracy after applying the band-stop processing. Furthermore, recognition performance improved in all cases with the exception of the UD Talk result with a ± 100 Hz stopband width. Fig. 4 confirms that the male fundamental frequency components were retained after band-stop processing, suggesting that preserving f_0 is critical for ASR performance.

In contrast, Fig. 1 indicates that the female harmonics tend to overlap with the male ones in the mixed audio because the female f_0 is higher. This overlap may have resulted in the loss of essential frequency components from the female voice, which could explain the lower recognition rates shown in Table IV. Moreover, Fig. 5 suggests that when the stopband width is wide and the f_0 of the attenuated voice is low, adjacent band-stop filters may overlap. This can lead to the removal of a continuous range of frequencies, not only the harmonics but also broader spectral components, making accurate extraction of the target speech difficult. Tables V—VII demonstrate that limiting the upper bound of the male harmonics for attenuation resulted in improved recognition accuracy. This can be attributed to the fact that, as the harmonic order (i) increases, the intensity of the harmonic components decreases (as observed in Fig. 1). This suggests that attenuating the strong lower-order male harmonics allows the female spectral features to stand out relatively more, making recognition easier even if weaker male harmonics are preserved. Furthermore, as shown in Table V–VII, when the stopband width was set to ± 100 Hz and harmonics up to the third order were attenuated, the recognition rate significantly decreased for all ASR systems. As shown on the right side of Fig. 6, the frequency components within the specified range were completely eliminated, indicating that both the fundamental and the second harmonic components play a crucial role in speech recognition.

Based on these results, the cutoff frequency for attenuation could be automatically determined by analyzing the harmonic strength distribution of the mixed signal. Specifically, by analyzing the relative strength of each harmonic and automatically selecting the attenuation upper limit (either f_0 or $2f_0$) accordingly, the practicality and robustness of the proposed method can be improved.

Additionally, the difference in optimal stopband widths among ASR systems can be attributed to the variation in their acoustic models and training datasets. Each ASR may have different sensitivities to specific frequency bands or varying levels of noise robustness, leading to differences in performance under the same processing conditions. As shown in Table III, UD Talk achieved the highest recognition rate of 84.0% in the case of male speech extraction. This suggests that when the separation condition allows relatively clear speech components to remain, using an ASR system optimized for clear and intelligible speech suitable for educational settings can lead to higher recognition accuracy. In contrast, Table V indicates that YYProbe achieved a recognition rate of 95.7% for female speech extraction. This implies

that even when the residual harmonics remain in the extracted audio as noise-like components, ASR systems with high noise robustness can still achieve high recognition accuracy. Furthermore, to approach the ideal recognition rate of 100%, training ASR systems on two types of speech datasets—one containing only fundamental frequency components, and the other containing both fundamental and second harmonic components—may prove effective.

This study used isolated male and female tracks under controlled conditions to examine the effect of harmonic attenuation on ASR performance. However, we acknowledge that the generalization of the results is limited. However, for practical applications, it is necessary to estimate each speaker's fundamental frequency directly from mixed speech without using isolated tracks. Various methods have been proposed for fundamental frequency estimation under such conditions. For example, Cuesta et al. proposed a multi-pitch estimation method combining harmonic constant-Q transform and convolutional neural networks (CNNs) [14]. Other approaches include speaker separation based on f_0 tracking using Kalman filters, and end-to-end automatic speech recognition models that simultaneously recognize speaker attributes, as proposed by Kanda et al. [15] [16]. In addition, it is essential to consider more complex scenarios such as same-gender speaker separation. Advanced band-stop filtering methods that take into account the amplitude of each harmonic or features other than harmonics may offer further improvements. Although this experiment was conducted in a quiet environment, real-world settings often include multiple noise sources. To enable practical deployment, incorporating additional techniques such as noise reduction using low-pass filtering is considered effective. However, we acknowledge that the generalization of the results is limited. Future work will expand the dataset to include more speakers and utterances, and evaluate performance under realistic acoustic environments with noise, reverberation, and overlapping speech. These extensions will enhance the practical applicability, robustness, and generalizability of the proposed method.

While DNN-based methods have achieved high performance in signal-level metrics (e.g., SDR, SIR, SAR), these measures mainly reflect perceptual quality rather than ASR accuracy [5] [7]. In contrast, this study directly targets ASR accuracy, showing that even without DNNs, selective attenuation of harmonic components based on f_0 differences can improve recognition performance in mixed-gender audio. Therefore, the proposed method offers a lightweight, interpretable, and computationally efficient alternative that contributes to understanding the relationship between harmonic structures and ASR performance.

VII. Conclusions

In this study, we aimed to support individuals with hearing loss in understanding and participating in conversations involving multiple speakers. To this end, we proposed a method for speech enhancement that applies band-stop filters to selectively attenuate individual harmonic components, and evaluated its effectiveness in improving ASR accuracy. The results demonstrated that recognition performance can be improved by tuning processing parameters according to the characteristics of each ASR system. In future work, we aim to extend this approach to more practical applications, including speech separation from same-gender mixed audio and noise reduction in real-world environments, in order to develop a system suitable for real-world deployment.

This study provides a foundation for developing speech recognition systems that are more accessible and inclusive for people with hearing loss.

References

[1] World Health Organization (WHO), "Deafness and hearing loss" https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, Feb. 2025 (accessed Jul. 2025).

[2] The University of Tokyo Office for Disability Equality, "What you should know about hearing impairment and deafness", https://ds.adm.u-tokyo.ac.jp/en/receive-support/hearing.html, Mar. 2025 (accessed Jul. 2025)

[3] Alan Chern, Ying-Hui Lai, Yi-ping Chang, Yu Tsao, Ronald Y. Chang, and Hsiu- Wen Chang, "A Smartphone-Based Multi-Functional Hearing Assistive System to Facilitate Speech Recognition in the Classroom", in Article in IEEE Access June 2017.

[4] Minsoo Kim, and Gil-Jin-Jang, "Speaker-Attributed Training for Multi-Speaker Speech Recognition Using Multi-Stage Encoders and Attention-Weighted Speaker Embedding", in Appl. Sci., 2024, 14 (18) .

[5] Shahar Lutati, Eliya Nachmani, and Lior Wolf "Seplt: Approaching a Single Channel Speech Separation Bound", in Interspeech, 2022.

[6] W3C, "Synchronization Accessibility User Requirements", https://www.w3.org/TR/saur/, Jun. 2023 (accessed Jul. 2025).

[7] Kohei Yatabe, and Daichi Kitamura, "Determined BSS Based on Time-frequency Masking and Its Application to Harmonic Vector Analysis", in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vlo. 29, 2021.

[8] Yoshiki Masuyama, Xuankai Chang, Wangyou Zhang, Samuele Cornell, Zhong-Qiu Wang, Nobutaka Ono, Yanmin Qian, and Shinji Watanabe "EXPLODING THE INTEGRATION OF SPEECH SEPARATION AND RECOGNITION WITH SELF-SUPERVISED LEARNING REPRESENTATION" in IEEE Workshop on Application of Signal Processing to Audio and Acoustics, 2023.

[9] Ingo R. Titze, "How Are Harmonics Produced at the Voice Source?" in Journal of Singing, May/June 2009 Volume 65, No 5, pp.575-576.

[10] SciPy, "butter - SciPy v1.16.0 Manual" https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.butter.html, Jun. 2025 (accessed Jul. 2025).

[11] Zoran Milibojevic, Bajan Prlinčević, and Dijana Kostić, "Estimation the Fundamental Frequency of the Speech Signal Using Autocorrelation Algorithm", in UniTech, 2021.

[12] Deividas Eringis, and Gintautas Tamulevičius, "Improving Speech Recognition Rate through Analysis Parameters", in Electrical Control and Communication Engineering Volume 5(2014), Issue 1 pp. 61-66.

[13] Meddy Fouquet, Katarzyna Pisanski, Nicolas Mathevon, and David Reby, "Seven and up: individual differences in male voice fundamental frequency emerge before puberty and remain stable throughout adulthood", in The Royal Society Open Science, 2016.

[14] Helena Cuesta, Brian McFee, and Emilia Gómez, "Multiple F_0 Estimation in Vocal Ensembles Using Convolutional Neural Networks", in 21st International Society for Music Information Retrieval (ISMIR) Coference, 2020.

[15] Aidan O. T. Hogg, Christine Evers, and Alastair Howe Moore, "Overlapping Speaker Segmentation Using Multiple Hypothesis Tracking of Fundamental Frequency", in IEEE/ACM Transactions on Audio Speech and Language Processing, 2021.

[16] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, Tianyan Zhou, and Takuya Yoshioka, "Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers" in INTERSPEECH, 2020.