A Falsetto Detection Algorithm for Enhancing Voice Gender Recognition

Ronald Mo
School of Computer Science and Engineering
University of Sunderland
Sunderland, United Kingdom
ronald.mo@sunderland.ac.uk

Calin Blendea
CASBI Limited
Cambridge, United Kingdom
calinblendea@casbi.co.uk

John Harper
Cambridge Machine Learning
Cambridge, United Kingdom
john.harper@cambridgemachinelearning.com

***Abstract*—This paper presents a novel falsetto detection algorithm designed to enhance the performance of Voice Gender Recognition (VGR). By incorporating Signal Processing techniques with insights from vocal pedagogy, our algorithm identifies falsetto in singing voice data to reduce gender identity ambiguity in vocal analysis. We used a pre-trained Deep Learning VGR model to assess the effectiveness of our algorithm. Experiments with various parameter settings demonstrate that the proposed algorithm reduced false positives in male voice detection and improved the VGR F1 score by a maximum of 5.3% for male voices and 2.6% for female voices. Our findings also highlight potential advancements in falsetto detection and provide insight for improving applications such as Voice Age Detection.**

***Index Terms*—signal processing, falsetto, falsetto detection, voice gender recognition, singing voice analysis**

I. Introduction

Falsetto is a unique vocal register commonly used in singing, theatrical performances, and opera [1], [2]. It is often employed to mimic female voices or create comedic effects due to its similarity to the female vocal range [3]–[5]. However, research on falsetto detection in singing is limited, likely due to the lack of annotated datasets and the complex spectral properties of this register. These challenges also impact voice analysis, as falsetto can blur the distinction between male and female voices, leading to potential errors in automated recognition systems. In light of this, we are interested in exploring the impact of falsetto on Voice Gender Recognition (VGR). That is, how the performance of VGR changes with the absence of falsetto for singing voice data.

Specifically, we propose a Signal Processing-based algorithm for detecting falsetto in singing voice data, building on existing research. Additionally, we implement a randomized approach for comparison. To evaluate its performance and its impact on VGR, we removed the falsetto from singing voice input using both our falsetto detection algorithm and the randomized method and fed the refined output into a VGR model. By evaluating classification accuracy, we aim to determine whether removing falsetto improves VGR performance and reduces misclassification. Additionally, we explore the best falsetto detection settings to optimize VGR accuracy. This study offers insights into the role of falsetto in voice perception and has broader applications in automated singing analysis, such as Voice Age Detection, as well as creative industries like singing voice editing and transformation.

The key contributions of this work are:

1) Proposing a Signal Processing-based falsetto detection algorithm for singing voice data.
2) Evaluating its performance on a Deep Learning-based VGR and suggesting an optimal setting.

The following sections provide a review of related literature, describe the adopted methodology, present experimental results, and conclude our work.

II. Background

A. Voices

The human voice is a powerful tool for expressing emotions. Changes in pitch, volume, and tone help convey different feelings. For example, singing is used in celebrations to express joy, while shouting or screaming can indicate anger or excitement. [6]–[9]. One vocal technique that significantly alters pitch is *falsetto*, which produces a light, airy sound. The term comes from the Italian *falso*, which means 'false' and refers to a vocal register beyond the natural range of the singer [10]. Traditionally, male singers used falsetto to reach higher pitches associated with female voices, but today, both men and women use it [5], [11]–[13].

Falsetto is produced differently from the natural, or *modal*, voice. In modal voice, the vocal folds fully contact each other during vibration, creating a rich, resonant tone. In contrast, falsetto—sometimes called head voice—is produced by stretching the vocal fold ligament, reducing contact between the vocal folds and increasing vibration frequency. This physiological difference gives falsetto its distinct tonal quality and ability to extend vocal range [3], [14]–[16].

B. Singing Voice Datasets

Several music-related datasets have been developed to support research in areas such as Audio Signal Processing. For instance, MedleyDB [17] provides royalty-free multitrack recordings that include both vocal and instrumental tracks, along with melody F0 annotations and instrument activity labels. These features make it particularly valuable for applications such as automatic instrument recognition and music source separation.

Despite this, datasets specifically tailored to singing voice research–particularly those addressing vocal techniques like falsetto–remain limited. Proutskova et al. [18] introduced a dataset featuring 10 vowel recordings in multiple languages, performed by a female vocalist across her full vocal range using phonation styles such as breathy and flow. VocalSet offers a more extensive collection, comprising 10 hours of recordings by professional singers employing a variety of vocal techniques (e.g., belt, breathy) in different musical contexts, including scales and arpeggios [19].

In the domains of speech therapy and voice science, the Perceptual Voice Qualities Database (PVQD) provides 296 clinical recordings of sustained vowels and spoken sentences, assessed by expert voice clinicians using standard evaluation protocols such as the CAPE-V and GRBAS scales [20].

C. Falsetto Analysis

Different methods have been used to study falsetto, including aerodynamic and acoustic analysis. Aerodynamic analysis examines airflow and air pressure during voice production, measuring factors like subglottal pressure, airflow, and phonation threshold pressure to assess vocal efficiency and quality [21], [22]. In contrast, acoustic analysis is a non-invasive method that evaluates voice characteristics such as pitch, loudness, and quality. The next section reviews related studies in acoustic analysis.

Previous studies have shown that the falsetto has unique acoustic features. For example, Hirano et al. [23] found that falsetto produced less noise energy and higher spectral decay compared to chest voice when singing high-pitched Japanese vowels. Li et al. [24] discovered that non-dysphonic female speakers had a higher pitch transition from modal to falsetto, while dysphonic speakers struggled to produce falsetto at certain loudness levels. Keating [15] suggested that falsetto can be distinguished from modal voice by a larger harmonic difference between the second and fourth harmonics ($H2*-H4*$), due to less energy in H4, and a lower Subharmonic-to-Harmonic Ratio (i.e., $SHR \in [0, 1]$), as falsetto lacks subharmonics. Lee et al. [16] also observed that the difference between the first and second harmonics is largest in falsetto, while modal voice shows the smallest difference.

D. Automatic Falsetto Detection

Although falsetto has been explored through various research methodologies, to the best of our knowledge, studies focused on automatic falsetto detection, such as those leveraging Machine Learning techniques, remain limited. It is likely due to technical challenges, particularly the time-consuming nature of collecting annotated falsetto data. The only notable previous work in this area is that of Mysore et al. [25], who proposed a method for automatic falsetto detection using a Support Vector Machine (SVM). They trained an SVM on voice data spanning 35 different pitches, with semitone increments, and utilizing 13 MFCCs as input features for classification. Their results demonstrated an accuracy of over 95% on their recorded dataset. However, there were several limitations in their work. Notably, the training data consisted of discrete semitone pitches, which may not represent the variability encountered in real-world singing. Real-world singing often does not adhere strictly to semitone intervals, as exemplified in techniques such as scatting. Therefore, their method may not generalize well to the more nuanced and varied pitch transitions found in everyday vocal performance.

E. Voice Gender Recognition (VGR)

Previous research on VGR has utilized Machine Learning and Deep Learning techniques. Harb and Chen [26] tested various VGR approaches, including Decision Trees, Gaussian Mixture Models, and Artificial Neural Networks (ANN) in multimedia data, revealing that ANN, when combined with acoustic and pitch characteristics, achieved a minimum classification accuracy of 90% for 1-second segments. This accuracy improved to 93% with further enhancements and reached 98.5% for 5-second segments.

Shue and Iseli [27] explored the integration of additional measures from the voice source, such as harmonic amplitude differences, as supplementary inputs for Machine Learning-based VGR. They trained an SVM using speech samples from speakers of different genders. Their findings indicated enhanced VGR performance across the majority of age groups. Expanding on this work, Chen et al. [28] incorporated three additional acoustic measures for VGR. Their study focused on gender classification of children's voices and demonstrated that these features enhanced VGR accuracy compared to the findings in Shue and Iseli. Bensoussan et al. [29] developed a Deep Neural Network for binary gender classification based on short audio samples of male and female voices. Their model achieved an overall accuracy of 92%, with a higher F1 score for female voices compared to male voices.

III. Methodology

A. Data Acquisition and Annotation

While, as discussed in Section II-B, several voice datasets exist, none include specific information on falsetto. To overcome this, we collected and annotated our dataset with a human expert. We recruited 11 male and 12 female indie singers from Europe, aged 20 to 32, with 2 to 15 years of vocal experience. Each singer performed their compositions in various languages and vocal registers.

	Male	Female
No. of Singers	11	12
No. of Songs	15	20
No. of Samples	674	856
Total Duration	1h 6m 42s	1h 38m 1s

TABLE I
The details of the created audio dataset.

All samples were recorded in a professional studio using studio-grade recording equipment. These recordings were unaccompanied, which means that only the vocal part was

Features	Description	Related Work
F_0	Fundamental Frequency	[27]
F_1, F_2, F_3,	First 3 Formant Frequencies	[27]
B_1, B_2	First 2 Formant Bandwidths	[27]
$H1*-H2*$	Corrected harmonic difference between 1^{st} and 2^{nd} harmonics	[15], [27]
HNR	Harmonic-to-Noise Ratio	[15], [28]
$H2*-H4*$	Corrected harmonic difference between 2^{nd} and 4^{th} harmonics	[15], [28]
SHR	Subharmonic-to-Harmonic Ratio	[15]

TABLE II
THE EXAMINED ACOUSTIC FEATURES.

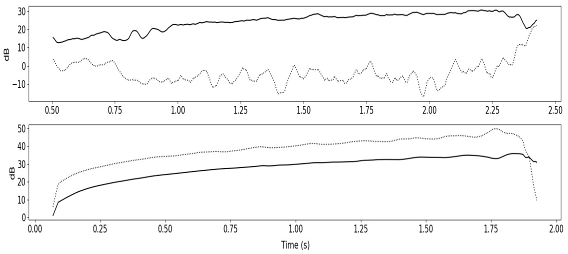

Fig. 1. Corrected harmonic amplitudes of falsetto voices for a professional singer (upper) and an average singer (lower). $H2*$ is shown as a solid line, while $H4*$ is represented by the dotted line.

present in the audio recording, and they were reverberation-free. The audio was recorded at a 44.1kHz sampling rate and a 24-bit bitrate. Loudness was normalized over all audio recordings. For better evaluation, the silence was removed and the recordings were split so that each sample contained a line of lyrics. Table I provides the details of the created dataset.

To annotate the recorded audio, two vocal coaches with seven and ten years of teaching experience and a Bachelor of Arts in Vocal Performance were engaged. Before the annotation process, a priming session was conducted to ensure consistency in the annotations. During this session, the vocal coaches were provided with examples of falsetto usage in pop music to ensure precision and consistency. To mitigate the risk of hearing fatigue and maintain consistency, they were instructed to take a 5-minute break every hour throughout the annotation process.

B. Falsetto Detection Algorithm

Given the low-resource constraints outlined in previous sections, we adopted a Signal Processing approach to develop our falsetto detection algorithm based on acoustic features reported in previous work [15], [27], [28]. To compute these acoustic features, we developed a Python implementation of VoiceSauce [30], as the latest MATLAB version no longer supports the original VoiceSauce. Table II summarizes the features we tested for this work.

Our initial analysis identified SHR and $H2*-H4*$ (discussed in Section II-C) as the most significant features to distinguish the modal voice from the falsetto. Consequently, we based our algorithm on these two features. However, our findings presented a somewhat different perspective compared to previous research.

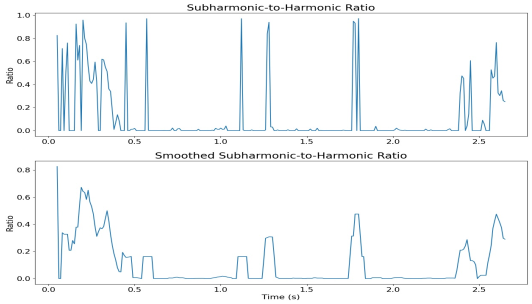

Fig. 2. SHR of a falsetto voice (upper) and after smoothing (lower).

Specifically, for $H2*-H4*$, our results diverged from prior studies. While [15] suggested that $H2*-H4*$ is generally large in falsetto due to the relatively low amplitude of $H4*$, our analysis indicated that this difference is influenced by the singer's level of expertise. Professional singers exhibited patterns aligning with previous findings, whereas average or amateur singers tended to demonstrate a high **negative** $H2*-H4*$ value. Fig. 1 shows an example of the corrected harmonic amplitudes of falsetto voices performed by professional and average singers.

In light of this, we propose the absolute difference between $H2*$ and $H4*$ (i.e., $abs(H2*-H4*)$) as a new measurement for our falsetto detection. This approach takes into account the varying differences observed between professional and amateur singers, with the positive difference seen in professional singers and the negative difference in amateurs, providing a more robust and consistent measure for distinguishing falsetto usage across different skill levels.

Our SHR observations were generally consistent with the results reported in [15]. However, we found that SHR was not stable for falsetto detection as it usually fluctuated in falsetto. To address this, we applied a moving average to smooth the SHR values. Fig. 2 shows the SHR in a falsetto voice before and after smoothing.

Our proposed falsetto detection algorithm incorporates both features using a rule-based approach. First, these two features were preprocessed. The $abs(H2*-H4*)$ values were resampled to match the length of SHR normalized by min-max normalization to ensure they fall within the range $[0,1]$:

$$abs_{norm} = \frac{abs - min(abs)}{max(abs) - mix(abs)} \quad (1)$$

where abs denotes $abs(H2*-H4*)$. For SHR, it was first smoothed as discussed above. Since SHR values are typically lower in falsetto and higher in modal voice, we applied the following inversion transformation for mathematical convenience:

$$SHR_{inverted} = -SHR + 1 \quad (2)$$

In other words, $SHR_{inverted}$ will tend to 1 if the input singing voice is falsetto, and approach 0 if it is not. The average of these two features was used as an indication of falsetto. Our falsetto detection is defined as follows:

	Male			Female		
	Precision	Recall	F1	Precision	Recall	F1
Baseline	**1.000**	0.605	0.754	0.709	**1.000**	0.830
Random	0.993	0.563	0.718	0.687	0.996	0.813
$Model_{f=0.6}$	0.994	0.664	0.796	0.740	0.996	0.849
$Model_{f=0.62}$	**1.000**	**0.676**	**0.807**	**0.748**	**1.000**	**0.856**
$Model_{f=0.64}$	0.994	0.647	0.784	0.731	0.996	0.843
$Model_{f=0.66}$	0.993	0.626	0.768	0.719	0.996	0.835
$Model_{f=0.68}$	0.993	0.626	0.768	0.719	0.996	0.835
$Model_{f=0.7}$	0.993	0.613	0.758	0.713	0.996	0.831
$Model_{f=0.75}$	**1.000**	0.613	0.760	0.713	**1.000**	0.833

TABLE III

EVALUATION METRICS FOR VGR. THE BEST RESULTS ARE IN BOLD. *Random* REFERS TO THE RANDOM REMOVAL APPROACH. f INDICATES THE FALSETTO THRESHOLD.

$$I(t) = \begin{cases} 1 & \text{if } average_t >= f \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $I(t)$ is the falsetto indicator function at time t, $average_t$ is the average of the features at time t, and f is the threshold. $I(t) = 1$ indicates a possible falsetto at time t in the singing voice, and $I(t) = 0$ indicates a modal voice. Finally, a segment is considered falsetto if its duration is at least 4096 samples long (roughly 92.9 ms for a 44.1 kHz sampling rate) [31].

To further examine the effectiveness of our algorithm, we implemented a randomized approach that labels audio segments as falsetto 50% of the time. Since the unintended removal of non-falsetto segments can negatively impact VGR performance by eliminating gender identity information from the singing voice data, any inaccurate falsetto detection method would degrade VGR performance.

C. Voice Gender Recognition

To investigate how falsetto impacts VGR, we used a pre-trained Deep Learning VGR.[1]This is a Wav2vec 2.0 model [32] which was fine-tuned on the Librispeech-clean-100 dataset [33] for the gender recognition task.[2]

We evaluated VGR using the dataset outlined in Section III-A. Initially, the input audio was processed using our falsetto detection algorithm, applying various falsetto thresholds, as well as the previously described randomized approach, to remove segments identified as falsetto. The remaining audio was then analyzed by the VGR model for gender classification. To facilitate this process, we developed a Python-based pipeline for detecting and eliminating falsetto segments, ensuring seamless integration with the VGR system. Notably, falsetto detection was applied to both male and female recordings to enable a comprehensive evaluation. This approach allowed us to determine the optimal threshold value for maximizing VGR performance.

IV. RESULTS

Table III presents the evaluation metrics for gender detection across different settings. The baseline model refers to the original Wav2vec 2.0 gender detection model (i.e., without falsetto detection), which demonstrated better performance in detecting female voices compared to male voices. However, the baseline model tended to misclassify male voices as female, as indicated by the relatively low recall and precision for male and female voices, respectively.

To assess our approach, we first evaluated the randomized method discussed in Section III-B. Our results indicated that this approach led to a performance degradation that adversely affected all evaluation metrics. In other words, the inaccurate removal of audio segments had a significant *negative* impact on VGR accuracy, highlighting the importance of preserving gender identity information in singing voice data.

To further investigate the impact of our algorithm on VGR performance, we experimented with various threshold values for falsetto detection. This analysis was conducted to validate our hypothesis that removing falsetto enhances VGR accuracy. We initially tested threshold values ranging from 0.5 to 0.75 in increments of 0.05 and observed that VGR performance began to degrade when the threshold dropped below 0.6, likely due to excessively strict falsetto detection. As a result, we concentrated our analysis on threshold values of 0.6 and above. As shown in Table III, a threshold of 0.62 yielded the best performance across all evaluation metrics, resulting in improved F1 scores of 0.807 for male voices and 0.856 for female voices–corresponding to gains of 5.3% and 2.6%, respectively.

Our results also offer insight into the effects of varying threshold values. Lower thresholds improve recall for male voices by correctly identifying more male voices, while higher thresholds slightly enhance recall for female voices. This aligns with expectations, as a higher threshold retains more falsetto segments, preserving gender ambiguity and making VGR more challenging. Conversely, a lower threshold enforces stricter falsetto detection, which may misclassify modal sounds as falsetto and remove key gender identity cues, such as timbral differences. Nevertheless, all tested thresholds outperformed both the baseline model and the randomized approach, demonstrating that our falsetto detection algorithm effectively

[1] https://huggingface.co/alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech

[2] https://ai.meta.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

identifies falsetto segments and enhances VGR performance.

V. Conclusions

In this paper, we introduce a novel falsetto detection algorithm and evaluate its performance within a Deep Learning-based VGR system. Our results demonstrate the effectiveness of the proposed algorithm, leading to a significant reduction in false positives and an overall improvement in gender classification accuracy, with an F1 score improvement by a maximum of 5.3% for males and 2.6% for females. Additionally, we identify an optimal threshold value of 0.62 for falsetto detection to enhance classification accuracy.

Based on our findings, we suggest areas for future research. First, using an adaptive threshold for falsetto detection could further improve accuracy. Since previous studies such as [5] have linked age to falsetto usage, another direction could be integrating our falsetto detection into systems that recognize age [34]. Our algorithm may also help with dysphonia detection, as individuals with dysphonia often struggle with falsetto production [24]. Finally, our work could have creative applications in music, such as enhancing vocal separation or informing music composition.

VI. Acknowledgment

We would like to express our gratitude to Crystal Yuen from the Voice Research Lab at the University of Hong Kong for her valuable advice on falsetto.

References

[1] Frederick Swanson, “The falsetto voice: its legitimacy and its usefulness,” *Choral Journal*, vol. 15, no. 9, pp. 13, 1975.

[2] Andrew Parrott, “Falsetto beliefs: the ‘countertenor’cross-examined,” *Early Music*, vol. 43, no. 1, pp. 79–110, 2015.

[3] Garyth Nair, “Voice pedagogy: The term” falsetto”: Navigating through the semantic minefield,” *Journal of Singing-The Official Journal of the National Association of Teachers of Singing*, vol. 60, no. 1, pp. 53–60, 2003.

[4] Robert J Podesva, “Phonation type as a stylistic variable: The use of falsetto in constructing a persona 1,” *Journal of sociolinguistics*, vol. 11, no. 4, pp. 478–504, 2007.

[5] Brian Stross, “Falsetto voice and observational logic: Motivated meanings,” *Language in society*, vol. 42, no. 2, pp. 139–162, 2013.

[6] Robert T Sataloff, “The human voice,” *Scientific American*, vol. 267, no. 6, pp. 108–115, 1992.

[7] Klaus R Scherer, “Expression of emotion in voice and music,” *Journal of voice*, vol. 9, no. 3, pp. 235–248, 1995.

[8] Christer Gobl and Ailbhe Nı Chasaide, “The role of voice quality in communicating emotion, mood and attitude,” *Speech communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[9] Eduardo Coutinho, Klaus R Scherer, and Nicola Dibben, “Singing and emotion,” *The Oxford handbook of singing*, pp. 297–314, 2014.

[10] Heidi Koelz, “Falsetto,” *The Antioch Review*, vol. 71, no. 2, pp. 223–232, 2013.

[11] Bradley K Fugate, *More than men in drag: Gender, sexuality, and the falsettist in musical comedy of Western civilization*, Ph.D. thesis, The University of North Carolina at Greensboro, 2006.

[12] Robert Crowe, ““he was unable to set aside the effeminate, and so was forgotten”: Masculinity, its fears, and the uses of falsetto in the early nineteenth century,” *19th-Century Music*, vol. 43, no. 1, pp. 17–37, 2019.

[13] Robert L Garretson, “The falsettists,” *Choral Journal*, vol. 24, no. 1, pp. 5, 1983.

[14] Malte Kob, Nathalie Henrich, Hanspeter Herzel, David Howard, Isao Tokuda, and Joe Wolfe, “Analysing and understanding the singing voice: recent progress and open questions,” *Current bioinformatics*, vol. 6, no. 3, pp. 362–374, 2011.

[15] Patricia A Keating, “Acoustic measures of falsetto voice,” in *annual meeting of the Acoustical Society of America. Providence, RI*, 2014.

[16] Yogaku Lee, Mitsuru Oya, Tokihiko Kaburagi, Shunsuke Hidaka, and Takashi Nakagawa, “Differences among mixed, chest, and falsetto registers: a multiparametric study,” *Journal of voice*, vol. 37, no. 2, pp. 298–e11, 2023.

[17] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello, “Medleydb: A multitrack dataset for annotation-intensive mir research.,” in *Ismir*, 2014, vol. 14, pp. 155–160.

[18] Polina Proutskova, Christophe Rhodes, Tim Crawford, and Geraint Wiggins, “Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing,” *Journal of New Music Research*, vol. 42, no. 2, pp. 171–186, 2013.

[19] Julia Wilkins, Prem Seetharaman, Alison Wahl, and Bryan Pardo, “Vocalset: A singing voice dataset.,” in *ISMIR*, 2018, pp. 468–474.

[20] Patrick R Walden, “Perceptual voice qualities database (pvqd): database characteristics,” *Journal of Voice*, vol. 36, no. 6, pp. 875–e15, 2022.

[21] Fariborz Alipour, Eileen M Finnegan, and Ronald C Scherer, “Aerodynamic and acoustic effects of abrupt frequency changes in excised larynges,” *Journal of Speech, Language, and Hearing Research*, 2009.

[22] Marco Guzman, Karol Acevedo, Fernando Leiva, Vasti Ortiz, Nicolas Hormazabal, and Camilo Quezada, “Aerodynamic characteristics of growl voice and reinforced falsetto in metal singing,” *Journal of Voice*, vol. 33, no. 5, pp. 803–e7, 2019.

[23] Minoru Hirano, Seishi Hibi, and Tomoaki Sanada, “Falsetto, head/chest, and speech mode: An acoustic study with three tenors,” *Journal of Voice*, vol. 3, no. 2, pp. 99–103, 1989.

[24] Nicole YK Li and Edwin M-L Yiu, “Acoustic and perceptual analysis of modal and falsetto registers in females with dysphonia,” *Clinical linguistics & phonetics*, vol. 20, no. 6, pp. 463–481, 2006.

[25] Gautham J Mysore, Ryan J Cassidy, and Julius O Smith, “Singer-dependent falsetto detection for live vocal processing based on support vector classification,” in *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*. IEEE, 2006, pp. 1139–1142.

[26] Hadi Harb and Liming Chen, “Voice-based gender identification in multimedia applications,” *Journal of intelligent information systems*, vol. 24, pp. 179–198, 2005.

[27] Yen-Liang Shue and Markus Iseli, “The role of voice source measures on automatic gender classification,” in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4493–4496.

[28] Gang Chen, Xue Feng, Yen-Liang Shue, and Abeer Alwan, “On using voice source measures in automatic gender classification of children’s speech,” in *Eleventh Annual Conference of the International Speech Communication Association*. Citeseer, 2010.

[29] Yael Bensoussan, Jeremy Pinto, Matthew Crowson, Patrick R Walden, Frank Rudzicz, and Michael Johns III, “Deep learning for voice gender identification: proof-of-concept for gender-affirming voice care,” *The Laryngoscope*, vol. 131, no. 5, pp. E1611–E1615, 2021.

[30] Yen-Liang Shue, Patricia Keating, Chad Vicenik, and Kristine Yu, “Voicesauce,” *p. Program available online at http://www. seas. ucla. edu/spapl/voicesauce/. UCLA*, 2009.

[31] Yukara Ikemiya, Kazuyoshi Yoshii, and Katsutoshi Itoyama, “Singing voice analysis and editing based on mutually dependent f0 estimation and source separation,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 574–578.

[32] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[33] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[34] Syed Rohit Zaman, Dipan Sadekeen, M Aqib Alfaz, and Rifat Shahriyar, “One source to detect them all: gender, age, and emotion detection from voice,” in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2021, pp. 338–343.