DOI: 10.17469/O2112AISV000013

MATTEO GAY, CLAUDIA ROBERTA COMBEI

Whose Voice Speaks Volumes? The Problem with Gender Identification from Speech[1]

This study investigates the topic of voice-based gender identification, focusing on aspects that concern accuracy and ethical implications. One aim is to review the literature available in order to understand what shapes the expectations of gender as revealed through voice. Then, an experimental investigation is conducted to examine the impact of age on voice-based gender classification. Our evaluation of a Multilayer Perceptron neural network model using MFCC features indicates 0.90 overall accuracy in binary gender classification. This seemingly high accuracy score would mask a substantial 10% misclassification rate in real-world applications. When considering age, accuracy drops further, with varying scores across the seven groups (0.64-0.88). Generally, our results indicate age-related variability in model performance, highlighting limitations and ethical concerns in generalizing across age groups. As concerns societal implications, our literature review and experiments suggest that greater awareness and diversity-informed approaches are needed in the design, development, and marketing of speech technologies.

Keywords: speech technology, voice, gender, bias, ethics.

1. *Introduction*

The recent advancements in voice-based technologies have been possible due to significant progress in the fields of speech and natural language processing (NLP), natural language understanding (NLU), and natural language generation (NLG). It has been reported that more complex deep learning architectures, such as recurrent neural networks (RNNs) and end-to-end training techniques, can improve the accuracy of automatic speech recognition (ASR), particularly as regards adverse conditions (Baby et al., 2022; Chen et al., 2023a).

Nevertheless, even advanced ASR systems and well-trained acoustic models can still struggle under challenging conditions (e.g., noise, non-native speech, etc.). Inspired by the fact that human listeners use contextual cues to infer meaning and reduce reliance on auditory input, Chen et al. (2023b) propose an open-source benchmark that uses large language models (LLMs) for ASR error correction, leveraging N-best decoding hypotheses to improve speech recognition accuracy.

[1] This paper is the result of close and continuous collaboration between the two authors, who are equally responsible for its content. For the purposes of Italian academia, C.R. Combei wrote § 1 (Introduction), § 2 (The topic), and § 5 (Conclusions), while M. Gay wrote § 3 (Data and methods) and § 4 (Results).

The study of Yu et al. (2024) goes in a similar direction, as it compares different ways of connecting LLMs to speech encoders for ASR. The authors claim that LLMs that use the Q-Former method may achieve better accuracy, even for out-of-domain speech. Additionally, they propose a segment-level Q-Former for handling longer speech to further improve accuracy, even though it appears that excessively long inputs increase the likelihood of hallucinating LLMs.

Other areas of research, such as intent recognition for conversational Artificial Intelligence (AI) chatbots (Chandrakala et al., 2024) or dialogue modelling and management (Fernández-Rodicio et al., 2020), have also contributed to the advancement of speech technologies, allowing, for instance, voice assistants to engage in conversations that are more natural and aware of the context. The combination of the aforementioned techniques has facilitated the development of commercial applications for everyday usage, such as real-time speech translation (Bentivogli et al., 2021), accessibility tools for visually impaired users (Norkhalid et al., 2020), and other voice-controlled interfaces.

These recent advancements in speech technologies, along with the possibility to use voice commands to interact with electronic devices (e.g., smartphones, computers, virtual assistants, in-vehicle infotainment, etc.) have profoundly transformed the way we communicate and access information. In line with what has been said above, voice-based applications have evolved well beyond basic commands (e.g., recognizing dictated numbers) to now being able to perform complex and sophisticated tasks (e.g., robust speech recognition under adverse conditions).

Voice has long been used also in biometric technologies, particularly within security systems designed for access control and user authentication (Fong, 2012). Since biometric voice authentication relies on the speaker's unique characteristics, it is featured in commercial applications for speaker verification that allow to unlock personal devices and access home banking, or smart home systems through vocal commands (Meng et al., 2020). Moreover, phone surveillance operations, border security devices, and forensic applications have been known to make use of voice classification techniques to categorize speakers based on attributes like gender[2], ethnicity, and age (Jain, Ross, 2015; Dumbrava, 2021).

Furthermore, customer relationship management systems have recently evolved to incorporate voice-based demographic classification for highly targeted advertising and marketing strategies (Aravinda et al., 2019). Consider, for instance, a scenario where a person is either scrolling through a social media application on their phone or using voice commands in a search engine application. Both applications could employ voice classification technology through the user's microphone (assuming the user has given prior permission for microphone access).

[2] To align with the terminology used in the corpus documentation used in our study (Burkhardt et al., 2010, 2023), we consistently refer to this variable as 'gender'. Besides ensuring coherence with the literature on the topic at hand, this terminological choice facilitates consistency throughout our analyses. The meaning of the term 'gender' is detailed in § 2.1.

The speech classification technology could analyse the voice interacting with the device, and based on that evaluation, assume the gender of the user. Then, this information could be used to customize the advertisements displayed to the user. For example, a person whose voice is classified by the system as male could receive advertisements for beard grooming products, while a person classified as female could receive advertisements for lipstick. Finally, it has been suggested that voice assistants on commercial devices are now able to tailor their responses based on the perceived gender of the user, inferred from voice characteristics and other cues derived from their behaviour (Alnuaim et al., 2022).

Similar to automated systems, humans also engage in the practice of inferring gender from the vocal characteristics of the speakers they hear and/or talk to. According to Dunkerson and Weiler (2023), listeners tend to assign a (binary) gender identity to a speaker exclusively on the basis of voice traits (e.g., when engaging in a telephone conversation with an unfamiliar interlocutor, one usually decides whether to address them as "Mr." or "Ms.", solely based on auditory cues). Nonetheless, Sutton (2020) warns that gender is not inherent in voice (listeners assign gender to voice) and that it is actually constructed through a number of variables (see also § 2.1).

As a matter of fact, the issue of assuming a person's gender, whether by humans or algorithms, has been a topic of intense debate in recent years. One significant issue reported in the literature is misgendering, which according to Fosch-Villaronga et al. (2021: 1) "reinforces gender stereotypes, accentuates gender binarism, undermines privacy and autonomy, and may cause feelings of rejection, impacting people's self-esteem, confidence, and authenticity". Their research examines the consequences of misgendering and highlights how algorithmic bias[3] can inadvertently violate privacy and perpetuate social biases and prejudices regarding gender identity.

Our study goes in a similar direction, and as anticipated throughout this section, it aims to critically address the concerns regarding automatic systems tasked with classifying speakers into gender categories based on their voice. To this end, we undertake a two-pronged approach. First, looking at the literature on the topic at hand, we examine the constructs that inform our inferences about gender as revealed through voice; in our exploration of the scholarly debate, we also highlight the potential issues of these assumptions in real-world situations. Second, we complement our theoretical arguments with an empirical investigation on gender recognition from speech by exploring the impact of the age variable on the detection process.

In addition to this introduction, the paper contains four other sections organized as follows: § 2 details the relationship between gender and voice, as

[3] We would like to stress the fact that bias in technology should not be relegated only to the issue of algorithmic bias. In fact, a significant portion of bias originates from naturally occurring data produced by human beings that are used to train and develop NLP technologies (Schramowski et al., 2022). These data inherently reflect the stereotypes, prejudices, and discriminatory practices that exist in the real world.

well as the motivations and objectives of our study; § 3 describes the data and the methods; § 4 presents the results of our analysis; § 5 concludes the paper.

2. *The topic*

This work deals with a topic that resonates with the growing attention towards ethical considerations in language technologies. A recent position paper by Sarker (2024) emphasizes the potential of AI and LLMs, while at the same time stressing the need to promote awareness and responsible practices in their integration in society. The paper highlights the importance of ethics, trustworthiness, and fairness in the development and use of these technologies, given their increasing presence in real-world applications. A relevant issue addressed is algorithmic bias, which can result in unfair or discriminatory outcomes, as well as in the spread and consolidation of harmful stereotypes and prejudices (Blodgett et al., 2020). Moreover, as highlighted by a large body of research (see, inter alia, Wu et al., 2022; Singh et al., 2023), the opaque nature of these systems makes it difficult to understand the mechanisms behind their outputs. In order to address these challenges, Sarker (2024) argues for the promotion of accountability, transparency, and interdisciplinary collaboration among stakeholders. Our study aligns with the existing literature on the ethical implications of developing algorithms for everyday tasks, as we advocate for continued debate and research on the topic to ensure a responsible deployment of these technologies.

The timeliness of this type of research is further demonstrated by the recent legislative developments regarding AI in the European Union. In fact, the European Parliament and Council negotiators reached a preliminary agreement on the Artificial Intelligence Act (AI Act)[4] on December 8, 2023, that was formally adopted on March 13th, 2024. The AI Act establishes a regulatory framework that assigns responsibilities to both developers and users of AI systems, with the level of accountability contingent upon the perceived risk associated with the technology. These regulations outline that some AI technologies and applications are entirely banned. They include biometric categorization systems that rely on sensitive data, unauthorized large-scale collection of facial recognition data from public sources (e.g., internet, CCTV footage, etc.), the use of AI for emotion recognition in workplaces and schools, social scoring systems, predictive policing based on user profiling, and AI designed to employ cognitive behavioral manipulation techniques targeting individuals or vulnerable groups (e.g., voice-activated toys that encourage dangerous behavior in children).

Voice technology encounters high expectations from users, whether they are individuals, private companies, or public institutions and organizations. However, what users might not know is that each decision of these automated

[4] The text adopted as the AI Act is available here: https://www.europarl.europa.eu/doceo/document/TA-9-2024-03-13-TOC_EN.html (accessed on May 30th, 2024).

systems consists of various intermediate steps that bring in elements of statistical probability and chance. The widespread idea that AI technologies, including voice-based systems, are infallible presents some problems, as it undermines critical scrutiny of their outputs in real-world scenarios and perpetuates undue trust. In fact, Kidd and Birhane (2023: 1222) warn that "[o]verhyped, unrealistic, and exaggerated capabilities permeate how generative AI models are presented" and this, of course, contributes to the "misconception that these models exceed human-level reasoning".

Our paper is motivated by the need to address the gap between the perceived infallibility of these technologies and the reality of their limitations, inaccuracies, biases, and ethical concerns.

2.1 Gender and voice

Commercial applications provide the statistically most probable outcome for speech recognition, verification, and classification tasks, but the factors influencing their outputs remain unknow to the public, due to the lack of transparency in the (commercial) model development and to the complexities associated with its interpretability. Moreover, the results depend on the quality and quantity of the training data employed for the construction of the model.

At the same time, the outputs of voice-based technologies are influenced by the developer's preconceived expectations regarding human voices and their interaction with demographic factors, such as gender and age (e.g., how an adult female voice is expected to differ from an adolescent male voice). As shown by previous research (see, inter alia, Fant, 1966; Nordström, 1977), these assumptions are typically supported by anatomical differences (e.g., size and shape of the vocal tract and vocal folds).

Traditionally, the phonetic differences between female and male speakers have been investigated in terms of the static acoustic and perceptual consequences of various articulatory dimensions, indicating, among others, an average difference of approximately 20% in vocal tract length between females and males (Simpson, 2001). Specifically, the pioneering work by Fant (1966, 1975) identifies larger laryngeal cavities and proportionally longer pharynges in male speakers as compared to female speakers.

In more recent times, the study of Smith et al. (2007) synthesises a range of vowels from recordings of four different speakers (an adult female, an adult male, a young boy, and a young girl) to assess whether gender and age influence the listeners' perception, after the voices are equated for glottal-pulse rate and vocal-tract length. The results show that listeners manage to distinguish children from adults by their voices, but they are unable to differentiate between male and female voices. A recent work by Zhang (2021) investigates the contribution of vocal fold length, thickness, and depth to differentiate between male and female voice production. The results indicate that the differences in voice production between

adult females, adult males, and children can be explained by differences in length and thickness, while the effect of vocal fold depth is small.

That being said, speakers and listeners use their voice to shape and contest the societal understandings of masculinity, femininity, and sexuality (Meyerhoff, Ehrlich, 2019). In fact, according to the social constructionist theories, gender emerges not as a fixed, inherent, or biological quality, but as a dynamic construct that is perpetually shaped and enacted in our daily interactions, both with ourselves and with others (Butler, 1999). Previous research has shown that the standards for effectively performing gender differs across historical and cultural contexts (Schilt, Westbrook, 2009). According to Dunkerson and Weiler (2023: 2), the Western paradigm fundamentally associates the 'correct' gender performance with binary genitalia distinctions (for female and male) and heterosexuality, constituting the so called 'cisheteronormativity'. In ordinary social interactions, however, gender attribution is more likely to be an inference based on a culturally constructed set of cues and behaviours associated with masculinity and femininity (West, Zimmerman, 1987).

The developers of speech technology likely base their projections regarding gender and voice on the two factors discussed so far. First, the culturally ingrained perceptions of masculinity and femininity, encompassing notions of gender-specific behaviours. Second, acoustic and auditory traits originating, for instance, from biological differences in the vocal tract and vocal folds. Besides these two elements, the corpora used for training purposes in the development of voice-based applications play a crucial role, as gender biases or misrepresentations in the training data have a significant impact on the performance of the systems developed. All things considered, insufficient and poorly representative data together with the developers' preconceived expectations have problematic consequences on the functioning of voice applications, resulting in various issues, such as misidentification and misgendering, as well as in the reinforcement of gender biases.

Various external variables can, in fact, alter the characteristics of one's voice. Smoking, for instance, may render the voice hoarse (Tafiadis et al., 2017), while alcohol consumption may result in slurred speech (Pisoni, Martin, 2023). Altered voice productions are also caused by dysphonia in individuals experiencing conditions that affect the vocal tract, such as the cold, the flu, the Parkinson's and the Alzheimer's disease (Contreras et al., 2023).

Recent research has also demonstrated that acoustic features, particularly fundamental frequency (F0), which is commonly used in voice-based gender classification tasks, are often influenced by conversational, contextual, and psycho-social factors. For example, a paper by Hughes and Puts (2021: 1) shows that speakers often lower their voice pitch in order "to exert dominance, display status, and compete with rivals". Then, Bradshaw et al.'s (2022) study indicates that F0 variability is heightened during initial conversational turns, such as greetings, likely serving to signal identity, mark boundaries, or attract attention.

The assumption that algorithms can accurately identify a person's gender based on their physical characteristics, such as the voice, is fraught with significant risks. As highlighted by Hamidi et al. (2018), diverse gender identities, including transgender and nonbinary individuals, have expressed profound apprehensions regarding the use of automated recognition and verification systems. The primary concern of these individuals regards the possibility that language technologies misgender or misunderstand them. In the online article "Queer Eye for AI: Risks and limitations of artificial intelligence for the sexual and gender diverse community" published on May 26th, 2023, for Open Global Rights (hosted by the Center for Human Rights and Global Justice and the Future of Rights Program at New York University School of Law)[5], Ilia Savelev illustrates the following problematic scenario:

> Imagine calling your bank to resolve an issue, but the customer service representative suddenly accuses you of fraud and threatens to suspend your bank account if you don't invite the "real" you. Why? Because their artificial intelligence (AI) voice recognition system concluded that your voice was not "male" enough to match their records. This is a universal experience among transgender individuals, including myself, and is just one of the significant risks that AI poses to the LGBTQI+ community.

In fact, the research of Ovalle et al. (2023) reveals a dominance of binary gender norms reflected by the AI models and stresses the need for further investigation into the manifestation of transgender and non-binary invisibility and bias within language technology.

A different example of misgendering based on physical traits is provided by Peynircioğlu et al. (2017) in their study regarding the McGurk effect in gender identification. Their results show that while participants accurately identified gender 100% of the time in conditions where video and audio were matched, the accuracy dropped to 31% in conditions where video and audio were mismatched. This finding indicates that visual gender cues can override auditory gender identification, revealing a strong non-speech McGurk effect.

2.2 From theory to practice

The elements discussed in the previous sections highlight the need for scholarly research that can guide the development of more inclusive speech technologies able to represent the complexity and diversity of human voices. In particular, interdisciplinary studies – with insights from gender studies, computer science, sociology, psychology, and linguistics – are needed to address the concerns surrounding speech classification systems.

To contribute to this debate, we conducted an empirical investigation on gender recognition from speech by exploring the effect of age on the identification process. Our hypothesis is that the age variable has a negative effect on the performance

[5] The article is available here: https://www.openglobalrights.org/risks-limitations-artificial-intelligence-sexual-gender-diverse-community/ (accessed on May 30th, 2024).

of the gender classification task, thereby diminishing the expected reliability and efficacy of this type of application.

This paper uses a practical real-world example where the factors within the data can significantly impact the performance of a speech classification system. Thus, we seek to demonstrate how the presence of an experimental variable (i.e., age) can influence the accuracy in identifying the classes of the target variable (i.e., gender). By doing so, our goal is to raise awareness of the potential for errors in these systems, discussing, at the same time the need for, a more prudent, inclusive, and informed approach to their design, development, and use.

3. *Data and methods*

Our study evaluates one of the prevailing baseline technologies in automatic speech classification, namely the application of neural network architectures based on Mel Frequency Cepstral Coefficients features. The corpus under examination is derived from the Agender corpus (Burkhard et al., 2010) for the German language, a well-established and extensive telephonic speech dataset built for age and gender classification, originally released for the INTERSPEECH 2010 Paralinguistic Challenge. The Agender corpus comprises 65,241 individual utterances from 945 native German speakers, categorized into four age classes. These age classes, except for children, are further divided based on the gender variable, resulting in seven groups: CHILD (7-14 years old, one class), YOUNG (15-24 years old, two classes), ADULT (25-54 years, two classes), and SENIOR (55-80 years, two classes). Each speaker delivered 18 speech items in up to six different sessions.

Since we are not engaging in the replication of the challenge, neither we aim to achieve high performance scores and compete with previous results, we decided to randomly split the train/development corpus to create the train/test sets (85% and 15%). Special attention was given to ensuring that the same speaker did not appear in both sets to eliminate any possibility of voice-specific pre-knowledge bias. As a result, our corpus includes the seven classes displayed in Tab. 1: CHILD (106 speakers, 6,802 instances, not divided by gender), YOUTH (99 females, 7,360 instances; 88 males, 6,189 instances), ADULT (113 females, 7,934 instances; 107 males, 6,929 instances), and SENIOR (133 females, 8,485 instances; 134 males, 9,375 instances). Thus, the final dataset for our experiment included a total of 53,074 instances from 770 speakers.

Table 1 - *Dataset description: f and m abbreviate female and male, x represents children w/o gender discrimination. The last two columns represent the number of speakers/instances per partition (Train and Test).*

Class	*Group*	*Age*	*Gender*	*#Train*	*#Test*
1	**C**hild	7-14	*x*	415/5,965	57/839
2	**Y**outh	15-24	*f*	393/6,151	79/1,209

3	**Y**outh	15-24	*m*	343/5,130	73/1,059
4	**A**dult	25-54	*f*	472/6,857	77/1,077
5	**A**dult	25-54	*m*	403/5,869	73/1,033
6	**S**enior	55-80	*f*	494/7,105	96/1,380
7	**S**enior	55-80	*m*	551/7,986	87/1,227

The distinctiveness of this corpus lies in the fact that the original audio files were recorded using standard cell phones (GSM) and landline connections at a sampling rate of 8,000 Hz in 8-bit ALAW format, introducing substantial disturbances and noise into the acoustic spectrum of the voice recordings. The training and development sets of the Agender corpus are publicly accessible and can be downloaded from the BAS CLARIN repository[6] of the Bavarian Archive for Speech Signals at Ludwig-Maximilians-Universität Munich. The files are provided in .raw format, so for the purposes of our analysis we converted them to .wav.

As mentioned in § 1, in the domain of speech processing, there exists a significant area of research known as speech classification. Unlike ASR, the primary objective of speech classification is not to obtain textual transcriptions of utterances but rather to assign human speech production to predefined classes related primarily to the speaker. These classes encompass attributes such as age range, gender, presence of clinical disorders, emotional states, speaker intentions (e.g., asking a question, making a statement, expressing uncertainty), and geographical provenance.

When approaching a classification problem from a computational point of view, typically, the primary focus is to optimize the classifier algorithm. However, before reaching the mapping function, an essential and non-negligible step is the selection and extraction of the most informative features from the input data, in order to reduce the space of the problem. These features (numerically encoded to feed the classification function) capture different aspects of the speech signal, containing valuable information for distinguishing between different classes. While the importance of specific features may vary depending on the classification task and the nature of the data, several key features have been widely recognized as crucial in speech classification, such as Mel Frequency Cepstral Coefficients (MFCCs), Shifted Delta Cepstral Coefficients (SDCC), Linear Prediction Cepstral Coefficients (LPCC), among low-level spectral features, spectral centroid, pitch-related features (mainly F0), as well as high-level features such as jitter and shimmer, commonly used in speech disorders and emotion detection (Teixeira et al., 2013).

In this study, we trained a vanilla Multilayer Perceptron (two hidden layers, 80 and 40 nodes) with a binary classifier head using the Scikit-learn MLPClassifier implementation on Python 3.9, employing only MFCC features. The model was trained on pre-processed audio data, which included noise reduction, silence trimming, and pre-emphasis techniques. Since our goal was to identify a possible

[6] The Agender corpus is available here: https://clarin.phonetik.uni-muenchen.de/BASRepository/index.php?target=Public/Corpora/aGender/aGender.1.php (accessed on May 30th, 2024).

source of misclassification (i.e., misgendering), not to build a robust and high-performing model[7], we opted for this simple yet fundamental architecture to maintain low complexity and higher interpretability.

4. *Results*

In this section, we present the evaluation results for the Multilayer Perceptron classifier in two specific tasks: overall gender prediction (G) and gender prediction within specific age classes (G-in-A). We report as measures F1 metric and accuracy.

The best score, overall accuracy and F1 of 0.90 (Tab. 2), was achieved in the binary G classification (excluding the first class, consisting of children aged 7 to 14 years old, for which gender information is not provided in the dataset). Table 2 shows the relevant confusion matrix for the results of this task.

Table 2 - *Confusion Matrix, gender binary classification task*

	Male	*Female*
Male	3187	291
Female	394	3006

While these metrics might suggest that the model performs well in distinguishing between male and female speakers, the 10% of the samples that were incorrectly classified represent a substantial number of individuals being misgendered in such a simple task, highlighting a critical issue that pure numerical performance metrics may overlook.

Notably, we observed an improvement (+0.05) in classification accuracy after applying the previously described preprocessing techniques (see § 3), which appears to be a crucial initial step when dealing with a noisy corpus like a telephone recorded speech dataset. It is also worth mentioning that the model exhibited a slight preference for predicting male gender. The performance on pre-processed audio file confirms the fact that MFCC features alone have, indeed, a relevant role in gender detection.

However, the most interesting results achieved are those produced by the experiment on G-in-A classification (Tab. 3 and Tab. 4).

Table 3 - *Results for Gender in selected Age Class experiment (Y = Young, 15-24 yrs; A= Adult, 25-54 yrs; S = Senior, 55-80 yrs; M = Male, F = Female.)*

	YF + YM	*AF + AM*	*SF + SM*
F1 score	0.92	0.84	0.80

[7] For more complex models addressing the gender classification task, including experiments on the Agender corpus, see Qawaqneh et al. (2017) and Burkhardt et al. (2023).

In particular, we attempted to train the model on gender recognition within specific age ranges, and we discovered that our SK-Model yielded better results in correctly detecting gender in younger speakers compared to seniors, with demonstrating median performance in the adults' class (Tab. 3).

As a demonstration of the utility and precision of MFCCs in suggesting possible correlations between speech, vocal tract characteristics, gender, and age, we also attempted to predict gender within specific age class pairings that were constructed for this purpose (Tab. 4). In this case, we included children as a separate class, since the Agender Corpus does not provide gender information for them. We observed that discrimination was generally easier between children and male speakers than between children and female speakers.

Table 4 - *Results for Child (C) vs Male/Female across different age classes.*

	C + YF	*C + YM*	*C + AF*	*C + AM*	*C + SF*	*C + SM*
F1 score	**0.64**	0.86	**0.76**	0.87	**0.79**	0.88

Specifically, the accuracy in predicting the values Child or Male gender remained consistently around 0.86-0.88 across all age classes (Young, Adult, Senior), whereas the model performed better in predicting Child or Female gender with older age classes than with younger (voice characteristics more similar to those of children), ranging from 0.64 to 0.79.

To substantiate these observations, we conducted bootstrap resampling analyses to evaluate the statistical significance of the differences in macro-averaged F1 scores between binary classifiers. When comparing C + YF to C + SF, the 95% Confidence Interval (CI) for the mean difference in F1 scores, derived from 10,000 bootstrap samples, was [0.071933, 0.126724]. This result allows us to reject the null hypothesis of no difference with high confidence. Similarly, significant differences were observed when comparing C + YF to C + AF (95% CI: [0.114848, 0.167898]) and C + AF to C + SF (95% CI: [-0.065162, -0.017584]), as detailed in Fig. 1, 2, and 3, in the Appendix.

Additionally, similar significant results were obtained when analyzing differences across gender groups within same age classes (e.g., Fig. 4, in the Appendix). Moreover, the slight improvement in distinguishing between males and children with increasing age classes did not reach statistical significance (95% CI: [-0.0028, 0.0355]), as shown in Fig. 5 and 6, in the Appendix. As reported by the literature cited in § 2.1, these findings may be linked to the physiological evolution of the vocal tract and vocal folds across different ages and the distinct physical characteristics of females and males.

Our findings reveal that the accuracy of the model varies greatly across different age groups, indicating that age acts as a confounding variable in voice-based gender classification. This variability in performance suggests a possible limitation in the capability of the model to generalize across the lifespan of an individual. We would like to stress that this issue extends beyond the influence of age on model

performance; it also raises significant ethical concerns related to misclassification in systems using architectures similar to the one we tested. This is particularly problematic for under-represented groups that may be more susceptible to the negative effects of bias and misgendering.

5. *Conclusions*

This paper addressed the topic of voice-based gender identification that, in very recent times, has started to receive attention in the scholarly debate, particularly as regards the issue of misgendering. First, we examined the topic, by reviewing the existing literature, to outline the elements that determine the way we shape our expectations about gender and voice. Second, we performed an experiment to investigate the impact of age on the task of gender classification from speech. We hypothesized that the age variable would affect the performance of gender identification, thereby reducing the accuracy of the model.

To this end, we tested one of the prevailing baseline technologies in automatic speech classification, namely the deployment of Multilayer Perceptron neural network architectures based on MFCC features. The model achieved an overall accuracy and F1 score of 0.90 in binary gender classification (the class of children aged 7 to 14 were excluded from this analysis). Even though these metrics may suggest robust performance in recognizing male and female speakers, the 10% error rate would represent a substantial number of misgendering instances, in a real-world situation.

As hypothesized, when predicting all seven classes, thus co-varied by the age factor, the accuracy decreased. Male speakers and children were identified with an accuracy from 0.86 to 0.88 across all age groups (young, adult, senior), while female speakers and children displayed a wider range of accuracy, namely from 0.64 to 0.79, with older age groups being more accurately classified than younger ones.

In general, our results showed that the accuracy of the model varied across different age groups, highlighting the role of age in gender classification from speech. This variability not only affects model performance but also reflects the limitations of the systems that are based on models similar to the one we tested. At the same time, it suggests that the topic needs attention from both researchers and the industry, for a more cautious, diversity-aware approach to the design, development, and marketing of speech technology. Switching to the societal perspective, we believe that studies like the one presented here are needed to raise awareness about the potential errors, inaccuracies, and biases in voice-based applications.

Appendix

Figure 1

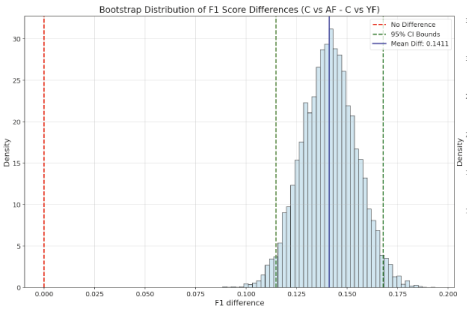


Figure 2

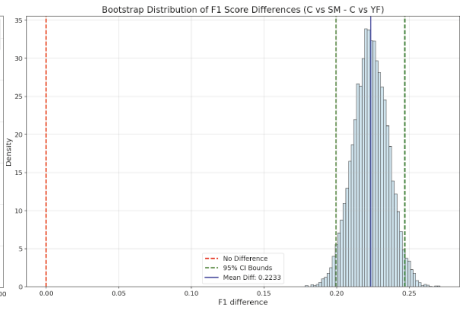


Figure 3

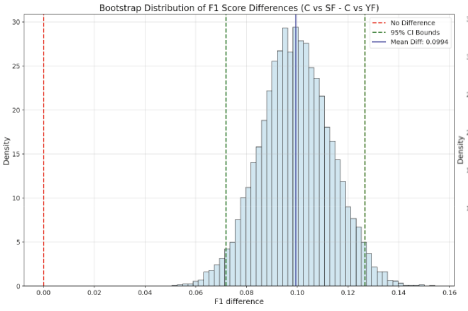


Figure 4

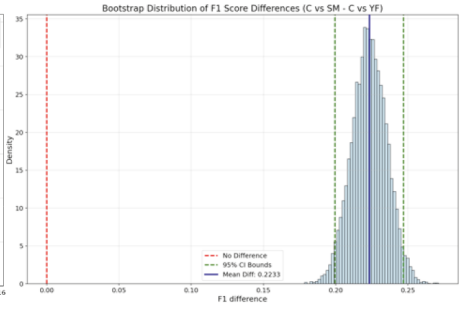


Figure 5

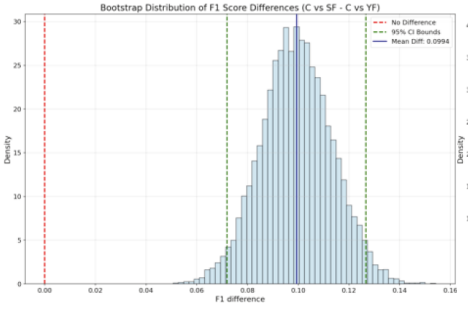
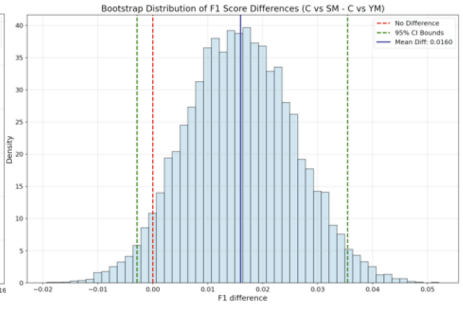


Figure 6



Acknowledgements

The first draft of this paper was written while C.R. Combei was engaged as a researcher under the initiative PON Ricerca e Innovazione 2014-2020 – Linea Innovazione (D.M. 1062/2021) at the University of Pavia.

References

Alnuaim, A.A., Zakariah, M., Shashidhar, C., Hatamleh, W.A., Tarazi, H., Shukla, P.K. & Ratna, R. (2022). Speaker gender recognition based on deep neural networks and ResNet50. In *Wireless Communications and Mobile Computing*, 2022(Special Issue), 4444388. https://doi.org/10.1155/2022/4444388.

Aravinda, D., Arpitha, N. & Mamatha, M. (2019). Gender Voice Classification using Deep Learning Convolutional Neural Networks. In *Journal of Critical Reviews*, 6(6), 2595-2601.

Baby, D., D'Alterio, P. & Mendelev, V. (2022). Incremental learning for RNN-Transducer based speech recognition models. In *Proceedings of Interspeech 2022*, Incheon, Korea, 18-22 September 2022, 71-75. https://doi.org/10.21437/Interspeech.2022-10795.

Bentivogli, L., Cettolo, M., Gaido, M., Karakanta, A., Martinelli, A., Negri, M. & Turchi, M. (2021). Cascade versus direct speech translation: Do the differences still make a difference? In Zong, C., Xia, F., Li, W. & Navigli, R. (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, Bangkok, Thailand, 1-6 August 2021, 2873-2887. https://doi:10.18653/v1/2021.acl-long.224.

Blodgett, S.L., Barocas, S., Daumé III, H. & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In Jurafsky, D., Chai, J., Schluter, N. & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online, 5-10 July 2020, 5454-5476. https://doi.org/10.18653/v1/2020.acl-main.485.

Bradshaw, L., Chodroff, E., Jaeger, L. & Dellwo, V. (2022). Fundamental frequency variability over time in telephone interactions. In *Proceedings of Interspeech 2022*, Incheon, Korea, 18-22 September 2022, 101-105. https://doi.org/10.21437/Interspeech.2022-10669.

Burkhardt, F., Eckert, M., Johannsen, W. & Stegmann, J. (2010). A Database of Age and Gender Annotated Telephone Speech. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*, La Valetta, Malta, 19-21 May 2010, 1562-1565.

Burkhardt, F., Wagner, J., Wierstorf, H., Eyben, F. & Schuller, B. (2023). Speech-based age and gender prediction with transformers. In *Proceedings of the 15th ITG Conference on Speech Communication*, Aachen, Germany, 20-22 September 2023, 1-5. https://doi.org/10.30420/456164008.

Butler, J. (1999). *Gender Trouble: Feminism and the Subversion of Identity*. Tenth Anniversary Edition. London/ New York: Routledge.

Chandrakala, C.B., Bhardwaj, R. & Pujari, C. (2024). An intent recognition pipeline for conversational AI. In *International Journal of Information Technology*, 16, 731-743. https://doi.org/10.1007/s41870-023-01642-8.

Chen, Z.-C., Yang, C.-H.H., Li, B., Zhang, Y., Chen, N., Chang, S.-Y., Prabhavalkar, R., Lee, H.-Y. & Sainath, T. (2023a). How to estimate model transferability of pre-trained speech models? In *Proceedings of Interspeech 2023*, Dublin, Ireland, 20-24 August 2023, 456-460. https://doi.org/10.21437/Interspeech.2023-1079.

Chen, C., Hu, Y., Yang, C.-H.H., Siniscalchi, S.M., Chen, P.-Y. & Chng, E.-S. (2023b). HyPoradise: An open baseline for generative speech recognition with large

language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M. & Levine, S. (Eds.), *Advances in Neural Information Processing Systems*, New Orleans, USA, 10-16 December 2023, Vol. 36, 31665-31688.

Contreras, R.C., Viana, M.S., Fonseca, E.S., Dos Santos, F.L., Zanin, R.B. & Guido, R.C. (2023). An Experimental Analysis on Multicepstral Projection Representation Strategies for Dysphonia Detection. In *Sensors*, 23(11), 5196. https://doi.org/10.3390/s23115196.

Dumbrava, C. (2021). Artificial intelligence at EU borders: Overview of applications and key issues. In *European Parliamentary Research Service*, 21(12), 690706.

Dunkerson, A., Weiler, A. (2023). "Oh! Based on voice, assigned female at birth": Transmasculine voices and gender construction. In *Canadian Graduate Journal of Sociology and Criminology*, 6(1), 1-18. https://doi.org/10.15353/cgjsc-rcessc.v6i1.5012.

Fant, G. (1966). A note on vocal tract size factors and non-uniform F-pattern scalings. In *STL-QPSR*, 4, 22–30.

Fant, G. (1975). Non-uniform vowel normalization. In *STL-QPSR*, 2(3), 1-19.

Fernández-Rodicio, E., Castro-González, Á., Alonso-Martín, F., Maroto-Gómez, M. & Salichs, M.Á. (2020). Modelling multimodal dialogues for social robots using communicative acts. In *Sensors*, 20(12), 3440. https://doi.org/10.3390/s20123440.

Fong, S. (2012). Using hierarchical time series clustering algorithm and wavelet classifier for biometric voice classification. In *Journal of biomedicine & biotechnology*, 2012(1), 215019. https://doi:10.1155/2012/215019.

Fosch-Villaronga, E., Poulsen, A., Søraa, R.A. & Custers, B.H.M. (2021). A little bird told me your gender: Gender inferences in social media. In *Information Processing & Management*, 58(3), 102541. https://doi.org/10.1016/j.ipm.2021.102541.

Hamidi, F., Scheuerman, M.K. & Branham, S.M. (2018). Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In Mandryk, R., Hancock, M. (Eds.), *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, Montreal, QC, Canada, 21-26 April 2018, Article No. 8, 1-13.

Hughes, S.M., Puts, D.A. (2021). Vocal modulation in human mating and competition. In *Philosophical Transactions of the Royal Society B: Biological Sciences,* 376(1840), 1-10, 20200388. https://doi.org/10.1098/rstb.2020.0388

Jain, A.K., Ross, A. (2015). Bridging the gap: from biometrics to forensics. In *Philosophical Transactions of the Royal Society of London*. Series B, Biological Sciences, 370(1674), 20140254.

Kidd, C., Birhane, A. (2023). How AI can distort human beliefs. In *Science*, 380(6650), 1222–1223. https://doi.org/10.1126/science.adi0248.

Meng, Z., Bin Altaf, M.U. & Juang, B.-H. (2020). Active voice authentication. In *Digital Signal Processing*, 101, 102672. https://doi.org/10.1016/j.dsp.2020.102672.

Meyerhoff, M., Ehrlich, S. (2019). Language, gender, and sexuality. In *Annual Review of Linguistics*, 5, 455-475. https://doi.org/10.1146/annurev-linguistics-052418-094326.

Nordström, P.-E. (1977). Female and infant vocal tracts simulated from male area functions. In *Journal of Phonetics*, 5(1), 81-92. https://doi.org/10.1016/S0095-4470(19)31115-5.

Norkhalid, A.M., Faudzi, M.A., Ghapar, A.A. & Rahim, F.A. (2020). Mobile application: Mobile assistance for visually impaired people - Speech Interface System (SIS). In *Proceeding of the 8th International Conference on Information Technology and Multimedia (ICIMU)*, Selangor, Malaysia, 24-26 August 2020, 329-333. https://doi:10.1109/ICIMU49871.2020.9243450.

Ovalle, A., Goyal, P., Dhamala, J., Jaggers, Z., Chang, K.-W., Galstyan, A., Zemel, R. & Gupta, R. (2023). "I'm fully who I am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, Chicago, USA, 12-15 June 2023, 1246–1266. https://doi.org/10.1145/3593013.3594078.

Peynircioğlu, Z.F., Brent, W., Tatz, J.R. & Wyatt, J. (2017). McGurk effect in gender identification: Vision trumps audition in voice judgments. In *The Journal of General Psychology*, 144(1), 59-68. https://doi.org/10.1080/00221309.2016.1258388.

Pisoni, D.B., Martin, C.S. (1989). Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses. In *Alcoholism: Clinical and Experimental Research*, 13(4), 577-587. https://doi.org/10.1111/j.1530-0277.1989.tb00381.x.

Qawaqneh, Z., Abu Mallouh, A. & Barkana, B.D. (2017). Deep neural network framework and transformed MFCCs for speaker's age and gender classification. In *Knowledge based system*, 115, 5-14. https://doi.org/10.1016/j.knosys.2016.10.008.

Sarker, I.H. (2024). LLM potentiality and awareness: A position paper from the perspective of trustworthy and responsible AI modeling. In *Discover Artificial Intelligence*, 4(40). https://doi.org/10.1007/s44163-024-00129-0.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C.A. & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. In *Nature Machine Intelligence*, 4(4), 258–268. https://doi.org/10.1038/s42256-022-00458-8.

Schilt, K., Westbrook, L. (2009). Doing Gender, Doing Heteronormativity: "Gender Normals," Transgender People, and the Social Maintenance of Heterosexuality. In *Gender & Society*, 23(4), 440-464. https://doi.org/10.1177/0891243209340034.

Simpson, A.P. (2001). Dynamic consequences of differences in male and female vocal tract dimensions. In *Journal of the Acoustical Society of America*, 109(5), 2153–2164. https://doi.org/10.1121/1.1356020.

Singh, C., Askari, A., Caruana, R. & Gao, J. (2023). Augmenting interpretable models with large language models during training. In *Nature Communications*, 14, 7913, 1-11. https://doi.org/10.1038/s41467-023-43713-1.

Smith, D.R.R., Walters, T.C. & Patterson, R.D. (2007). Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. In *Journal of the Acoustical Society of America*, 122(6), 3628–3639. https://doi.org/10.1121/1.2799507.

Sutton, S.J. (2020). Gender ambiguous, not genderless: Designing gender in voice user interfaces (VUIs) with sensitivity. *In Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*, Bilbao, Spain, 22-24 July 2020, Article No. 11, 1–8. https://doi.org/10.1145/3405755.3406123.

Tafiadis, D., Toki, E.I., Miller, K.J. & Ziavra, N. (2017). Effects of early smoking habits on young adult female voices in Greece. In *Journal of Voice*, 31(6), 728-732. https://doi.org/10.1016/j.jvoice.2017.03.012.

Teixeira, J.P., Oliveira, C. & Lopes, C. (2013). Vocal acoustic analysis – Jitter, shimmer and HNR parameters. In *Procedia Technology*, 9, 1112-1122. https://doi.org/10.1016/j.protcy.2013.12.124.

West, C., Zimmerman, D.H. (1987). Doing gender. In *Gender & Society*, 1(2), 125-151. https://doi.org/10.1177/0891243287001002002.

Wu, T., Terry, M. & Cai, C.J. (2022). AI Chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '22)*, New Orleans, USA, 29 April-5 May 2022, Article No. 385, 1-22. https://doi.org/10.1145/3491102.3517582.

Yu, W., Tang, C., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z. & Zhang, C. (2024). Connecting speech encoder and large language model for ASR. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024),* Seoul, Korea, 14-19 April 2024, 12637-12641. https://doi.org/10.1109/ICASSP48485.2024.10445874.

Zhang, Z. (2021). Contribution of laryngeal size to differences between male and female voice production. In *Journal of the Acoustical Society of America*, 150(6), 4511–4521. https://doi.org/10.1121/10.0009033.