Frequency-based Pre-processing for Reducing the Effect of Sex in Machine-Learning Voice Analysis for Alcohol Intoxication Detection

CESARINI VALERIO, COSTANTINI GIOVANNI
Department of Electronic Engineering,
University of Rome Tor Vergata,
Via del Politecnico, 1, 00100 Rome,
ITALY

Abstract: - In machine learning-based voice analysis the sex of the subjects is the most important covariate, as vocal differences between sexes lead to significant discrepancies in acoustic features and model generalization. We propose a signal processing pipeline that normalizes audio recordings of female subjects to exhibit characteristics more similar to those of males, or vice versa. We computed ratios of fundamental frequency, formants, breathiness, and pitch variability from a baseline dataset and used them to build a pipeline based on the decomposition of a signal into fundamental frequency contour, spectral envelope and aperiodicity. We tested on a classification task for the detection of intoxicated versus sober subjects and observed a systematic increase in accuracy when using our processing on the dataset. Accuracies increase up to 5.5% across all three considered algorithms: Support Vector Machine, Random Forest, and Neural Network, with the latter model applied on processed data achieving a state-of-the-art accuracy of 81.79%. *Summary:* This paper presents a frequency-based audio preprocessing method to reduce gender-related acoustic variability in machine learning-based voice analysis, particularly for detecting alcohol intoxication. The authors developed a pipeline to normalize female voice characteristics toward male standards, addressing differences in fundamental frequency, formants, breathiness, and pitch variability. Applied to a dataset of sober and intoxicated speech samples, this preprocessing significantly improved classification accuracy for intoxication detection, achieving up to a 5.5% accuracy increase, with a neural network reaching a state-of-the-art accuracy of 81.79%. The pipeline effectively minimized sex-based acoustic variability without introducing notable perceptual artifacts, thus enhancing generalization and model reliability across genders in voice-based machine learning tasks.

Key-Words: - voice, gender, alcohol, frequency, speaker recognition, audio.

Received: March 12, 2025. Revised: June 14, 2025. Accepted: July 17, 2025. Published: November 10, 2025.

1 Introduction

With the rapid ascent of Artificial Intelligence-based techniques and digitalization, automated voice analysis plays an increasingly crucial role in applications like emotion detection [1], automatic speech recognition [2], speaker identification [3], and clinical voice assessment [4], due to its capacity to quantify voice characteristics related to a certain speaker condition.

Common methodologies in automated voice analysis typically involve machine learning (ML) algorithms such as Support Vector Machines (SVM) [5] or Random Forest applied to acoustic features that include fundamental frequency (F0), formant frequencies [6], spectral characteristics, voice quality parameters (e.g., jitter, shimmer [7]), and cepstral coefficients like MFCCs [8]. Additionally, high-level Deep Learning methodologies are rapidly evolving thanks to increased processing power, delivering promising performance but often sacrificing interpretability, [9], [10].

The main premise in voice analysis is that a certain speaker state can be detected in voice and is reflected in a systematic and measurable change in its characteristics and features. However, there are intrinsic factors influencing voice characteristics independently of the primary analysed condition, especially the biological sex of the speaker and his/her age. These parameters inherently affect vocal anatomy and physiology, leading to pronounced variations in acoustic features.

Differences between male and female voices are especially important due to significant and consistent divergences, which affect the accuracy and reliability of automated systems by shifting the acoustic features and also creating inherent clusters within the data. For instance, gender differences may lead algorithms to confuse gender-specific

acoustic traits with pathology-related characteristics, causing misclassification.

The primary differences between male and female voices involve pitch, vocal tract length, breathiness and pitch variability. These differences translate to acoustically measurable features, which could be summarised, as many studies already proved, with the following differences exhibited by males with respect to female subjects:

- Lower pitch, or F0, [6];
- Longer vocal tract, which translates to slightly lower formant frequencies (F1, F2, F3), [11];
- Less presence of breath in voice, resulting in less high-frequency components, [12];
- More pitch variability, [13].

Most modern voice analysis applications still do not employ any kind of sex normalization, which results in the risk of models being biased. This effect is not mitigated by using balanced dataset, since the inherent internal clustering and the strong differences in feature distributions between sexes still effectively hinder ML training.

As suggested by [14] or [15], human listeners tend to naturally "normalize" gender characteristics to isolate other relevant parameters. Most attempts at normalizing voice analysis by gender were derived from speaker normalization techniques used in speaker recognition, which are optimized for isolating personal peculiarities of each subject. These techniques either focus on normalizing feature values after-extraction [16], [17], employing gender-normalized or independent features such as the system proposed by [18] employing MLP "bottleneck" features [19], or using pre-processing techniques to modify the audio signal. Another solution is the usage of independent ML models for each sex, or sex-specific optimized models such as [20], where the authors analyze the influence of multiple speech emotion features in male and female speech, and establish optimal feature sets for male and female emotion recognition, increasing speech emotion recognition accuracy. One of the latest promising results is represented by [21], where the authors explore the subject by normalizing female voice to male-like counterparts by using pitch and formant shifting through the commonly employed software Praat, [22].

The purpose of this study is the proposal of a pre-processing pipeline to modify female audio towards a male-like counterpart, or vice versa. In comparison to [21], our approach takes into account other factors besides frequency shifting of F0 and formants and uses more processing techniques that retain an almost perfect perceptual quality, which is a necessity considering that processing inherently creates artifacts which may worsen the audio signal in other ways.

This study uses a baseline dataset to determine male-to-female ratios for F0, formants, breathiness and pitch variability, which are then used to build a pre-processing pipeline that transforms a female audio recording into a male-like counterpart. The methodology is then applied to a relevant voice analysis problem in the form of binary classification for alcohol inebriation detection, implementing ML classifiers on unprocessed and processed versions of the dataset to assess the benefits of our sex normalization. The analysis is accompanied by an observation of the most relevant acoustic features used to discriminate between male and female subjects within the same dataset, confirming that all F0 and similarly related features are not detected anymore.

2 Materials and Methods

2.1 Baseline Dataset

We used a baseline dataset of male and female voice for statistical purposes to derive the parameters for the pre-processing pipeline. We employed the Gender Recognition by Voice (original) dataset by Murthada Najim [23], under the free-to-use license Apache 2.0, containing 5768 audio files by female speakers and 10400 files by male speakers uttering more than 1130 English sentences. The dataset contains lossless .wav files with a depth of 16 bits and a sampling rate of 16000 Hz.

2.2 Dataset for Classification: ALC by BAS

The pipeline built using the parameter values computed on the baseline dataset was used to pre-process recordings of female subjects in a binary classification task on the ALC dataset by the Bavarian Archive of Speech Signals (BAS), [24].

The dataset contains vocal samples of 162 German native people (85 males, 77 females), both in a sober and in an alcohol-induced inebriation state (Blood Alcohol Level > 0.05%), recorded in car cabins. For ease of interpretation and in line with many studies employing the dataset for automatic voice-based drunkenness detection, we only used a subset of the dataset containing the tongue twister "*Die Köchin mit dem Tupfenkopftuch kocht Karpfen in dem Kupferkochtopf*" which was demonstrated as being the most effective vocal task for the comparison, recorded with a Beyerdynamic

Opus54.16/3 condenser headset microphone in a 16-bit .wav format, down-sampled to a rate of 16000 Hz. The final dataset thus consisted of 324 audio samples by the same set of speakers in sober and drunk state.

2.3 Voice Conversion Parameters

With the aim of building a pre-processing pipeline that transforms audio recordings by female speakers into files that have acoustic characteristics more like male speakers or vice versa, we derived numeric parameters as male-to-female ratios, which we used to feed advanced signal processing methods to "normalize" selected acoustic features in the original recordings. The most relevant features that differentiate male from female speech have been thoroughly described in phoniatric and speech production literature and can be summarized as F0, formants, pitch variability, and breathiness. For each of these features, we computed male-to-female ratios by averaging the feature across all male recordings and dividing it by the average across all female recordings. The ratios are *r0* (F0), *rf* (formants), *rpv* (pitch variability) and *rb* (breathiness). The formula for r0 is given as an example:

$$r0 = \frac{\overline{F0_M}}{\overline{F0_F}} \quad (1)$$

where $\overline{F0_M}$ stands for the average of the F0 for all male recordings, the same goes for female recordings. The features, which were then averaged and divided, were computed as such:

Average F0 was computed using the YIN algorithm [25], with a minimum candidate frequency of 40 Hz and a maximum of 300 Hz. The algorithm outputs a tracking vector for each subject, which was averaged.

The "formant" ratio was the result of the average of the values of the three formants F1, F2 and F3, found using the Burg algorithm [26] based on Linear Predictive Coefficients (LPC) applied to a resampled version of the data, as implemented in Praat. The resulting formant tracking vectors endured median filtering (window = 3 samples) to remove artifacts and ensure frequency localization of the formant, before being averaged for each subject. Female and male-corpus averages for F1, F2 and F3 were computed and then ratioed. Sinche the three ratios were similar (range 88-92%), they were then averaged into a single formant ratio, especially considering that formant shifting is a complex procedure and doing it once brings great benefits to the quality of the reconstructed audio.

Pitch variation was computed by dividing the standard deviation of the F0 tracking vector by its mean, thus obtaining a coefficient of variation invariant of the pitch center.

Breathiness was defined as the average spectral energy above 5KHz, and computed from the FFT of each audio file, [12]. Since female speakers have a higher pitch contour which leads to higher formants and harmonics, their spectral energy is naturally distributed more towards higher frequency. On the other hand, breathiness should only reflect non-phonatory acoustic components related to the presence of air in speech. For these reasons, breathiness values for female speakers were computed after shifting down pitch and formants to more male-like ranges, with the methodologies that will be described shortly.

2.4 Pre-processing Pipeline

The pre-processing pipeline used to transform female-recorded audios into male-like recordings consisted of the following steps:

- Pitch shifting (downwards);
- Formant shifting of the whole formant structure (downwards);
- Increase of pitch variability;
- Mitigation of breathiness through shelf equalization.

In order to minimize the artifacts and errors of shifting procedures, we opted to minimize the number of spectral decompositions and subsequent re-syntheses. Therefore, we performed a spectral decomposition employing the WORLD vocoder, [27]. The WORLD vocoder is a high-performance speech synthesis technique that decomposes the spectrum into three components: F0 contour extracted using the Harvest algorithm [28], spectral envelope extracted using the Cheaptrick algorithm [29], and aperiodicity extracted using the D4C algorithm [30].

Following that decomposition, we performed the following operations:

- Pitch shifting: re-weighting of the F0 contour by multiplying by *r0*;
- Pitch variability control: expansion of the shifted F0 contour:

$$f0' = \overline{f0} + rpv \cdot (f0 - \overline{f0}) \quad (2)$$

where f0 is the original contour, from which the mean value $\overline{f0}$ is subtracted allowing re-weighing of the variation components (expansion if *rpv* > 0,

smoothing/compression if *rpv* < 0) before adding back the original mean value.

- Formant shift: the whole formant structure is shifted by warping the spectral envelope. This is achieved by creating a new frequency axis warped by the *rf* rate and redistributing spectral content by interpolating the original spectral envelope.
- Re-synthesis: the new f0 contour, warped spectral envelope and aperiodicity (untouched) components are finally used to re-synthesize audio: periodic components are constructed from the spectral envelope and F0 contour, while noise components are shaped by the aperiodicity map.

The pitched audios then endure the last step of "breathiness control" by the usage of a shelf equalizer implemented with an FFT whose values above the cutoff value (5KHz) are re-weighted by the *rb* ratio by constructing an attenuation curve with a short, smooth transition at the cutoff.

Figure 1 details the whole pipeline that transform a female audio to a male-like audio.

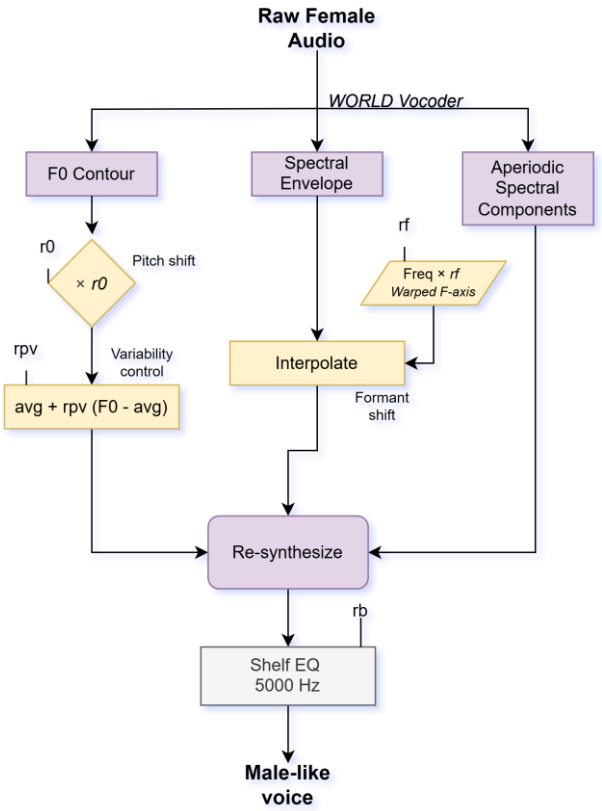

Fig. 1: Pre-processing pipeline, starting from a raw female-voice audio that endures pitch shift, pitch variability expansion, formant shift (all using WORLD vocoder) and then shelf EQ (cutoff 5KHz) to transform into a male-like voice

2.5 Machine Learning Tasks

The ALC tongue twister dataset was prepared for binary classification tasks (sober vs. non-sober), which were deployed in two versions: raw and with female audios being processed with our proposed methodology. In accordance with many studies regarding medium-to-small datasets for voice analysis, we employed a feature-based Machine Learning (ML) pipeline.

Raw audios were trimmed to remove leading and trailing silences using Praat, then normalized to a maximum of 1.

Afterwards, 6373 acoustic features were extracted using OpenSMILE [31] and the Interspeech Compare 2016 feature set [32], which contains features from all of the most relevant domains in voice analysis, including energy, perceptual measurements, frequency, Cepstrum [34], RASTA [33], F0, prosody. The features were reduced with a supervised feature selection procedure performed using a custom implementation based on Mark Hall's Correlation-Based Feature Selector (CFS) [34], which works by identifying an optimal subset of features with a merit-based search method, where the merit is computed as:

$$Merit = \frac{k \cdot \bar{r}_{fc}}{\sqrt{k + k(k-1) \cdot \bar{r}_{ff}}} \quad (3)$$

where $\bar{r}_{fc}$ and $\bar{r}_{ff}$ are the average correlation of the k features in the subset respectively with the class label and the other features. The CFS is paired with a search method to select the best feature subset in a minimum redundancy, maximum relevance fashion; the subset size is unpredictable and exclusive to each comparison. In order to avoid bias, two different feature selection procedures were employed: one for the raw dataset and another for the processed version. The selection was implemented by hand with custom routines employing a forward greedy search algorithm [35], gradually increasing the subset size and stopping after 3 continuous iterations without Merit increase.

Finally, selected features were used to feed three different ML algorithms chosen amongst the best-performing in voice analysis, namely:

Support Vector Machine (SVM) [5] with soft-margins and a linear kernel, trained on scaled features and a Complexity value of 1;

Random Forest (RF), trained for 500 iterations on bags as big as 80% of the original dataset, sampled with repetition, [36];

Neural network (NN), comprised of one input layer, one ReLu hidden layer of 5 neurons and one ReLu

output layer (2 neurons), trained for 500 epochs with an Adam optimizer, [37].

Performances were evaluated by 5-fold cross-validation in order to avoid test set-induced bias, and are reported in terms of accuracy since the dataset is balanced.

The CFS was also used on a different split of the same dataset, dividing female subjects from male regardless of the intoxication status. This was done in order to evaluate the most relevant features that distinguish between the two classes, in order to assess the changes brought by the pre-processing.

All of the methodologies were implemented on Python using the following libraries: librosa (YIN algorithm), parselmouth (Burg algorithm), pyworld (WORLD vocoder), scipy (FFT), opensmile and scikit-learn (all ML models). Computational resources consisted of a Windows-based system equipped with an 11th-generation Intel i9 processor, 128 GB of RAM, and an NVIDIA RTX 4000 GPU.

3 Results

3.1 Voice Conversion Parameters Values

The values for the voice conversion parameters (female to male) are reported in Table 1 alongside a brief explanation. These parameters are used as multiplicative factors, e.g., *r0* = 0.72 means that the F0 of a female voice is lowered to F0*0.72. On the contrary, *rpv* = 1.81 means that male voices exhibits a higher pitch variation (confirmed by literature, [13]).

Figure 2, Figure 3 and Figure 4 display the spectrogram of a female voice before and after conversion, alongside the spectrogram of their difference.

Table 1. Voice Conversion Parameters (Female to Male)

Parameter	Explanation	Value
r0	Fundamental Frequency ratio	0.72
rf	Formant (averaged for F1, F2, F3) ratio	0.91
rpv	Pitch variation (std. of F0 vector divided by mean F0)	1.81
rb	Breathiness ratio (spectral energy above 5KHz)	0.81

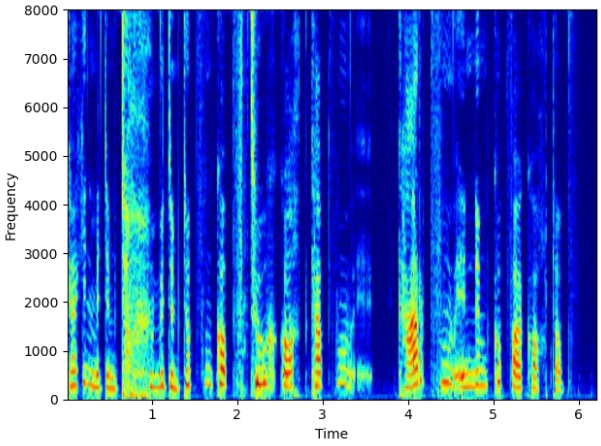

Fig. 2: Raw spectrogram of a female speaker

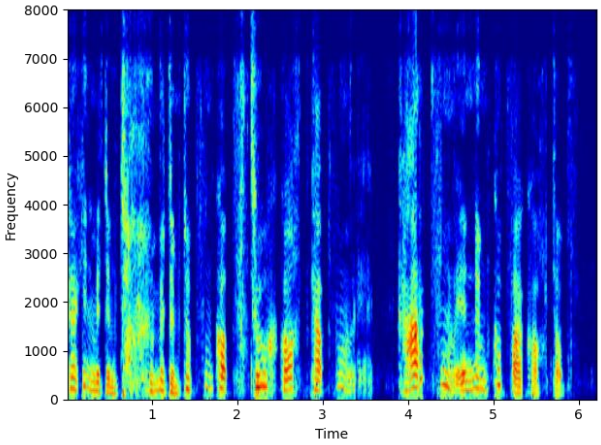

Fig. 3: Spectrogram of the processed counterpart of Figure 2

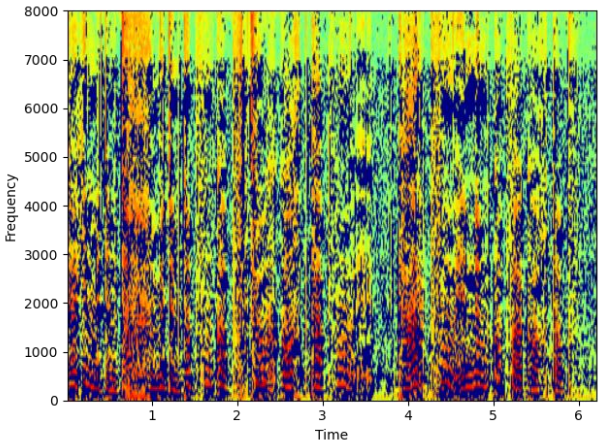

Fig. 4: "Delta" spectrogram displaying the differences between the pre and post-processing spectrograms of the same voice (Figure 2 and Figure 3)

3.2 Classification Results

Table 2 reports the classification results on the raw dataset and on its processed counterpart, in terms of

balanced cross-validated Accuracy for each ML model.

Table 2. Classification Results

Dataset	Accuracy (%)		
	SVM	**RF**	**NN**
Raw (no processing)	74.69	74.70	76.25
Processed (female audios endured the proposed pipeline)	77.15	75.62	81.79
Δ (processed – raw accuracy)	2.46	0.92	5.54

3.3 Male-Female Distinction: Pre and Post-processing

Table 3 displays the top 15 (for ease of reading) most relevant acoustic features with their OpenSMILE nomenclature, as identified by the CFS for the male-to-female distinction of the ALC dataset. It is evident that the processed subset does not rely anymore on features related to F0 or formants which are not present in the whole subset.

Table 3. Top 15 Acoustic Features for Male vs. Female Discrimination, Pre and Post-Processing

Raw (unprocessed) dataset	**Processed dataset**
F0final_sma_quartile1	pcm_fftMag_spectralRollOff75.0_sma_percentile1.0
mfcc_sma[11]_amean	mfcc_sma[8]_quartile2
mfcc_sma[14]_iqr1-2	mfcc_sma[12]_quartile3
mfcc_sma[11]_skewness	mfcc_sma[4]_quartile2
mfcc_sma[3]_amean	mfcc_sma[6]_quartile2
audSpec_Rfilt_sma[0]_quartile1	pcm_fftMag_spectralEntropy_sma_linregc1
mfcc_sma[4]_qregc3	mfcc_sma[14]_quartile3
mfcc_sma[14]_amean	mfcc_sma[10]_quartile2
F0final_sma_quartile2	pcm_fftMag_spectralEntropy_sma_range
audSpec_Rfilt_sma[4]_qregc2	mfcc_sma[8]_quartile1
mfcc_sma[13]_amean	mfcc_sma[8]_flatness
F0final_sma_amean	mfcc_sma[6]_peakMeanAbs
mfcc_sma[11]_upleveltime50	audSpec_Rfilt_sma[0]_iqr1-3
mfcc_sma_de[11]_kurtosis	mfcc_sma[4]_peakMeanAbs
mfcc_sma[14]_rqmean	mfcc_sma[8]_lpgain

4 Discussion and Conclusion

The premise of this study was to propose a pre-processing pipeline that could normalize audio recordings to mitigate the effect of sex as a voice covariate. The pipeline we built is based on the most relevant differences between male and female voices and was implemented with advanced signal processing methodologies that ensured minimal phase modifications and maximum signal integrity. The pipeline was implemented to transform female recordings into male-like audio.

We then used the pipeline to pre-process all recordings by female speakers in the ALC dataset (tongue twister task) and set up binary classifications.

From the results, it is evident that the pre-processing brings a definite increase in classification performance regardless of the considered classifier, the best result being achieved by the neural network, which gained 5.54% accuracy, resulting in a state-of-the-art performance of 81.79%, [38], [39].

Although the three chosen classifiers are based on different inference logics, they all improve after our pre-processing, with Random Forest gaining the least.

The usage of CFS as a feature selection mechanism applied independently on raw and processed datasets ensures that both comparisons are unbiased and implemented within the best possible feature set.

These results confirm the potential of our proposed pipeline to reduce the effects of sex in voice analysis: with sex being the most crucial covariate, small-to-medium datasets such as the one we employed benefit the most from its normalization. In fact, reducing the effects of sex by normalizing F0, formants, breathiness, and pitch variability with pre-fixed ratios does not hinder the capability of the models to detect variations in these domains imposed by other conditions.

The processed audio recordings of female speakers from the ALC, as well as the baseline Gender Recognition dataset, were evaluated by ear by audio engineers and audio experts and were confirmed as being absolutely intelligible, with no noticeable noisy artifacts, but not indistinguishable from naturally male recordings. Thus, it is important to state that, although the pre-processing appears to improve classification performance in voice analysis, it is not meant to be a way to convert female-spoken audio to sound exactly like male audio from a perceptual point of view.

The delta spectrogram in Figure 4, obtained by subtracting spectrogram values from raw audio and its processed counterpart, shows how the majority of

the modifications applied by our pipeline are concentrated towards the lower frequencies (F0 and formants), as well as an obvious reduction above 5 kHz due to the shelf equalization. The presence of definite vertical lines corresponding with the onset of speech confirms the fact that our pipeline is specific to the processing of speech in our predefined regions and leaves other characteristics of the signal relatively unaltered.

As a final confirmation, we set up a male vs. female classification task over the same dataset in order to observe the most relevant acoustic features, as extracted by the CFS. The results confirm that all features related to F0 or formants, characterized by the "F0" prefix in the table, are not used anymore.

One limitation to this study is the fact that male and female voices are not an absolute, and all people exhibit various degrees of "masculinity" or "femininity" in their voice, due to their anatomy (vocal tract length), cultural habits or language. Besides, people who identify in genders different from cisgender heterosexual males or females or who are undergoing hormonal therapy, exhibit proven differences in voice that cannot be categorized as male or female-like, [40]. From a technical standpoint, the crucial characteristics that dictate the kind of audio pre-processing to use are strictly related to the anatomy of a person's vocal tract: it is crucial to underline that this aspect is not necessarily related to a person's sexual or gender identification, also considering the fact that the human voice is a spectrum.

The usage of signal processing for the modification of an audio file is a procedure that may inherently introduce artifacts in the data, which, despite not being linked to the speakers' sex anymore, may still hinder ML training.

The advantages of a sex-independent pre-processing is crucial throughout all scenarios of voice analysis: first of all, it allows better generalization for classification tasks by mitigating critical inter-class differences. Secondly, a bottom-down approach can be viable, where research outputs can be compared to their sex-independent variants to study underlying mechanisms of the condition under exam.

Our future experimentations will expand the present study by re-applying the procedure to other datasets and classification tasks, as well as exploring the potential of the algorithm as a data augmentation solution. The overarching objective of this methodology remains the enhancement of all AI-based voice analyses across diverse scenarios by effectively neutralizing sex-dependent characteristics in voice signals.

In conclusion, the proposed pre-processing pipeline effectively reduces sex-based acoustic variability in voice analysis tasks. Its implementation enhances classification accuracy significantly without compromising signal intelligibility or integrity, highlighting its potential utility for studies involving voice analysis across genders.

Acknowledgement:
The authors would like to thank Voicewise S.r.l. for purchasing and providing access to the ALC dataset, and Murthada Najim for providing the Gender Recognition voice dataset under a free Apache 2.0 License.

References:

[1] J. L. Bautista, Y. K. Lee, H. S. Shin, «Speech Emotion Recognition Based on Parallel CNN-Attention Networks with Multi-Fold Data Augmentation», *Electronics*, vol. 11, fasc. 23, Art. fasc. 23, Jan 2022, doi: 10.3390/electronics11233935.

[2] G. Costantini, V. Cesarini, E. Brenna, «High-Level CNN and Machine Learning Methods for Speaker Recognition», *Sensors*, vol. 23, fasc. 7, Art. fasc. 7, Jan 2023, doi: 10.3390/s23073461.

[3] Cesarini, V.; Costantini, G. Reverb and Noise as Real-World Effects in Speech Recognition Models: A Study and a Proposal of a Feature Set. *Appl. Sci*. 2024, *14*, 11446. https://doi.org/10.3390/app142311446.

[4] M. Alves, G. Silva, B. C. Bispo, M. E. Dajer, P. M. Rodrigues, «Voice Disorders Detection Through Multiband Cepstral Features of Sustained Vowel», *J. Voice*, vol. 37, fasc. 3, pp. 322–331, May 2023, doi: 10.1016/j.jvoice.2021.01.018.

[5] C. Cortes, V. Vapnik, «Support-vector networks», Mach. Learn., vol. 20, fasc. 3, pp. 273–297, Sept. 1995, doi: 10.1007/BF00994018.

[6] G. Fant, "Acoustic Theory of Speech Production". Book, Ed. *Walter de Gruyter*, 1970.

[7] J. P. Teixeira, C. Oliveira, C. Lopes, «Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters», *Procedia Technol.*, vol. 9, pp. 1112–1122, 2013, doi: 10.1016/j.protcy.2013.12.124.

[8] B. P. Bogert, «The quefrency alanysis of time series for echoes ; Cepstrum, pseudo-

autocovariance, cross-cepstrum and saphe cracking», *Time Ser. Anal.*, pp. 209–243, 1963.

[9] J. Kaur e A. Kumar, «Speech Emotion Recognition Using CNN, k-NN, MLP and Random Forest», 2021, pp. 499–509. doi: 10.1007/978-981-15-9647-6_39.

[10] Cesarini, V.; Saggio, G.; Suppa, A.; Asci, F.; Pisani, A.; Calculli, A.; Fayad, R.; Hajj-Hassan, M.; Costantini, G. Voice Disorder Multi-Class Classification for the Distinction of Parkinson's Disease and Adductor Spasmodic Dysphonia. *Appl. Sci.* **2023**, *13*, 8562. https://doi.org/10.3390/app13158562.

[11] S. Umesh, S. V. Bharath Kumar, M. K. Vinay, R. Sharma and R. Sinha, "A simple approach to non-uniform vowel normalization," *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, 2002, pp. I-517-I-520, doi: 10.1109/ICASSP.2002.5743768.

[12] J. Hillenbrand, R. A. Cleveland, e R. L. Erickson, «Acoustic Correlates of Breathy Vocal Quality*», J. Speech Lang. Hear. Res.*, vol. 37, fasc. 4, pp. 769–778, Aug. 1994, doi: 10.1044/jshr.3704.769.

[13] C. Henton, «Pitch dynamism in female and male speech», *Lang. Commun.*, vol. 15, fasc. 1, pp. 43–61, Jan 1995, doi: 10.1016/0271-5309(94)00011-Z.

[14] G. Pino Escobar, J. Terry, B. P. Kriengwatana, P. Escudero, «Speech normalization across speaker, sex and accent variation is handled similarly by listeners of different language backgrounds: Australasian International Conference on Speech Science and Technology», *Proc. Sixt. Australas. Int. Conf. Speech Sci. Technol*. 6-9 Dec. 2016 Parramatta Aust., pp. 161–164, 2016.

[15] K. Johnson, M. J. Sjerps, «Speaker Normalization in Speech Perception», in *The Handbook of Speech Perception*, John Wiley & Sons, Ltd, 2021, pp. 145–176. doi: 10.1002/9781119184096.ch6.

[16] J. D. Miller, «Auditory-perceptual interpretation of the vowel», *J. Acoust. Soc. Am.*, vol. 85, fasc. 5, pp. 2114–2134, 1989, doi: 10.1121/1.397862.

[17] A. H. Fabricius, D. Watt, e D. E. Johnson, «A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics*», Lang. Var. Change*, vol. 21, fasc. 3, pp. 413–435, Oct. 2009, doi: 10.1017/S0954394509990160.

[18] T. Schaaf, F. Metze, «Analysis of gender normalization using MLP and VTLN features», presented at Proc. Interspeech 2010, pp. 306–309. doi: 10.21437/Interspeech.2010-117.

[19] P. Fousek, L. Lamel, J.-L. Gauvain, «Transcribing broadcast data using MLP features», presented at Proc. Interspeech 2008, pp. 1433–1436. doi: 10.21437/Interspeech.2008-414.

[20] L.-M. Zhang, Y. Li, Y.-T. Zhang, G. W. Ng, Y.-B. Leau, H. Yan, «A Deep Learning Method Using Gender-Specific Features for Emotion Recognition», *Sensors*, vol. 23, fasc. 3, Art. fasc. 3, Jan 2023, doi: 10.3390/s23031355.

[21] D. Rizhinashvili, A. H. Sham, G. Anbarjafari, «Gender Neutralisation for Unbiased Speech Synthesising», *Electronics*, vol. 11, fasc. 10, Art. fasc. 10, Jan 2022, doi: 10.3390/electronics11101594.

[22] P. Boersma, D. Weenink, «PRAAT, a system for doing phonetics by computer», *Glot Int.*, vol. 5, pp. 341–345, gen. 2001.

[23] «Gender Recognition by Voice(original)». Accessed on: March 15, 2025. [Online]. https://www.kaggle.com/datasets/murtadhanajim/gender-recognition-by-voiceoriginal

[24] F. Schiel, C. Heinrich, S. Barfüßer, T. Gilg, «ALC: Alcohol Language Corpus», in Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco: European Language Resources Association (ELRA), May 2008.

[25] A. de Cheveigne, H. Kawahara, «YIN, a fundamental frequency estimator for speech and musica)», *J Acoust Soc Am*, vol. 111, fasc. 4, 2002.

[26] A. Gray, D. Wong, «The Burg algorithm for LPC speech analysis/Synthesis», *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, fasc. 6, pp. 609–615, dic. 1980, doi: 10.1109/TASSP.1980.1163489.

[27] M. Morise, F. Yokomori, K. Ozawa, «WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications*», IEICE Trans. Inf. Syst.*, vol. E99.D, fasc. 7, pp. 1877–1884, 2016, doi: 10.1587/transinf.2015EDP7457.

[28] M. Morise, «Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals», in *Interspeech 2017*, ISCA, ago. 2017, pp. 2321–2325. doi: 10.21437/Interspeech.2017-68.

[29] «CheapTrick, a spectral envelope estimator for high-quality speech synthesis, Speech Communication, Vol. 67, March 2015, pp. 1-7. https://doi.org/10.1016/j.specom.2014.09.003.

[30] M. Morise, «D4C, a band-aperiodicity estimator for high-quality speech synthesis», *Speech Commun.*, vol. 84, pp. 57–65, nov. 2016, doi: 10.1016/j.specom.2016.09.001.

[31] F. Eyben, B. Schuller, «openSMILE:): the Munich open-source large-scale multimedia feature extractor», *ACM SIGMultimedia Rec.*, vol. 6, fasc. 4, pp. 4–13, gen. 2015, doi: 10.1145/2729095.2729097.

[32] B. Schuller et al., «The INTERSPEECH 2016 computational paralinguistics challenge: 17th Annual Conference of the International Speech Communication Association, INTERSPEECH 2016», *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-September-2016, pp. 2001–2005, 2016, doi: 10.21437/Interspeech.2016-129.

[33] H. Hermansky, N. Morgan, «RASTA processing of speech», *IEEE Trans. Speech Audio Process.*, vol. 2, fasc. 4, pp. 578–589, ott. 1994, doi: 10.1109/89.326616.

[34] M. A. Hall, «Correlation-based Feature Selection for Machine Learning», PhD Thesis, University of Waikato, New Zealand, 1999.

[35] M. K. and K. Johnson, "Stepwise Selection", from *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Accessed on: February 19, 2023, [Online]. https://bookdown.org/max/FES/greedy-stepwise-selection.html (Accessed Date: October 10, 2024).

[36] J. R. Quinlan, «Induction of decision trees», Mach. Learn., vol. 1, fasc. 1, pp. 81–106, mar. 1986, doi: 10.1007/BF00116251.

[37] D. P. Kingma e J. Ba, «Adam: A Method for Stochastic Optimization», January 30, 2017, arXiv: arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980.

[38] F. Amato, V. Cesarini, G. Olmo, G. Saggio, G. Costantini, «Beyond breathalyzers: AI-powered speech analysis for alcohol intoxication detection», Expert Syst. Appl., vol. 262, p. 125656, March 2025, doi: 10.1016/j.eswa.2024.125656.

[39] D. Bone, M. Black, M. Li, A. Metallinou, S. Lee, e S. Narayanan, "Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors.", *Interspeech 2011*, p. 3220. doi: 10.21437/Interspeech.2011-805.

[40] VL. Holmes, G. Rieger, e S. Paulmann, «The effect of sexual orientation on voice acoustic properties», *Front. Psychol.*, vol. 15, ago. 2024, doi: 10.3389/fpsyg.2024.1412372.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.