Informative Speech Features based on Emotion Classes and Gender in Explainable Speech Emotion Recognition

Huseyin Ediz Yildirim
Information Science
Open University of the Netherlands
Heerlen, the Netherlands
ediz.yildirim@ou.nl

Deniz Iren
Information Science
Open University of the Netherlands
Heerlen, the Netherlands
deniz.iren@ou.nl

***Abstract*—Emotions manifest in various aspects of human speech. While the tonality of the speech is a crucial indicator of emotions, other aspects such as word selection, pronunciation, and other paralinguistic features also provide valuable insights. Some of these aspects are considered universal, others are influenced by cultural and personal aspects, with gender being one of the most significant factors affecting emotional expressions. In this study, we aimed at investigating the effect of gender on emotional descriptors in speech. Specifically, we used intelligible paralinguistic speech features in Speech Emotion Recognition and employed Shapley values to measure the effect of gender on speech features. Furthermore, we empirically evaluated whether a reduced set of informative features could provide sufficient information for emotion recognition. Additionally, we investigated how gender influences auditory expressions of emotions.**

Our experiments show that besides the physical impact on fundamental speech frequencies, gender also affects how emotional phrases are spoken, and how prosody and phonology change. In addition to that, reducing the input size using the feature informativeness does not have a significant effect on the model accuracy whereas it shrinks the input size drastically by 98% on average. Finally, our comparative experiments on genders show that some speech features are more informative for capturing particular emotions exhibited by different genders. Therefore, we report that with a multi-layer feature set that consists of obscure and interpretable paralinguistic features, a novel data fusion approach could yield an explainable speech emotion recognition model. Furthermore, it is possible to reduce the input size and computational requirements by implementing feature reduction and gender information for speech emotion recognition tasks.

***Index Terms*—speech emotion recognition, affective computing, explainable machine learning, feature selection**

I. Introduction

Emotions are an essential part of human communication. They influence decision-making and social interactions and even serve as a mechanism for survival. Humans have a natural way of displaying and perceiving emotions. However, experts have conflicting views on whether emotional expressions are universal or not [1]. This makes affective computing tasks challenging for researchers and practitioners. Contemporary affective computing applications use auditory, visual, physiological, biological, and behavioural modalities. Speech emotion recognition (SER) is a sub-field of affective computing that aims at inferring emotions reflected in speech. Speech is a fundamental form of human communication, and emotions play a crucial role in conveying meaning through speech. A speech is characterized by *linguistic* and *paralinguistic* features. In a speech, linguistic features determine what message is delivered, while paralinguistic features define how that message is conveyed. Some paralinguistic features are *intelligible* because they represent concepts that are easily understood such as the duration of pauses in speech. Others are *obscure* as they represent statistical characteristics of underlying audio signals such as zero-crossing rate, thus, they are not outright intelligible.

The core assumption of SER is that the affective states of an individual are reflected in the speech features [2].These features are dependent on the speaker's individual physical, cultural [3], lingual [4], and acoustic [5] characteristics. Therefore, features extracted from the speech can be used for inferring the affective state of the speaker.

General approaches that tackle SER use machine learning. In recent years, advances in machine learning and audio signal processing have enabled the development of SER systems with increasing performance. Typical approaches incorporate obscure paralinguistic features that are extracted similarly for any kind of audio data [6]. There are also studies in the literature that employ intelligible paralinguistic [7] and linguistic speech features [8].

SER performance might be improved when speaker characteristics such as gender are taken into consideration [9]. Speech features have a varying degree of informativeness for certain emotion classes. A systematic analysis of speech features, their relationship with the speaker's gender, and their impact on conveyed emotions remain understudied. This paper is purposed to address this gap by providing an analysis of gender effects on paralinguistic speech features and their role in the model explanation while investigating the ways to reduce the input data size. We aim to add to the discussion on the summarisation of emotional activity along with model explainability and efficiency.

In this study, we address the following research questions:

979-8-3503-2745-8/23/$31.00 ©2023 IEEE

RQ1. Which speech features are more informative for different speaker genders and emotion classes in SER?

RQ2. Can a subset of speech features that are selected by their informativeness be as useful as the full feature set in SER?

RQ3. Which speech features are more prone/robust to gender bias in SER?

The rest of the paper is organized as follows. Section II lays down the background information that is referred to in this paper, section III reviews the literature on speech feature analysis associated with emotion labels and speech characteristics. Section IV describes our research method. Section V reports the results of our study. In Section VI, we discuss our findings. Finally, in Section VII, we conclude the paper.

II. Background

Audio is the digitally recorded representation of sound waves of particle vibrations. Speech sound is generated by vibrating the air using the vocal folds (cords) in the larynx (also known as the voice box) and shaping the vibrated air using articulators such as the pharynx, and nasal and oral organs [10]. Vocal cords produce irregular sine waves that congregate on top of each other constituting raw speech. To be able to further understand the speech, features of different speech aspects have been defined [11] [12] [13](Figure 1) . A variety of feature sets and signal-processing methods have been developed [14] to extract meaningful information from speech. These methods aim at performing on general audio data and they usually employ statistical calculations, preceding signal processing practices [15], acoustic feature extraction [16], and deep learning techniques [17]. When it comes to speaking, the human brain processes other cues along with the sound and interprets them along with paralinguistic speech features that imply how the speech is done. Pronunciation of phonemes [18], speed of speech [19], pauses [20], and the complexity of the speech [21] influence how the speaker is emotionally perceived by the listener. The way the speech is conducted should be taken as important as the content of the speech for perceiving the emotions in a human-like fashion.

A. Explanations of Features

For extracting intelligible paralinguistic speech features from audio data, we used DisVoice [12] paralinguistic speech feature extraction library. Specifically, Disvoice's static feature extraction methods from Glottal [22], Phonological [23] and Prosodic [24] categories were used for this study (Figure 2). The explanations of the descriptors and features are as follows.

- Glottal Features: Glottal speech features are computed from sustained vowels and continuous speech.
  - Glottal Closure Instant (GCI): Defines the moment where the vocal tract system is significantly excited during the production of the speech.
  - Opening Quotient (OQ): The ratio of time that vocal cords are open within a glottal cycle. The Quasi Opening Quotient (QOQ) is a regularly used derivative of which is based on the amplitude of the glottal pulse [25].
  - Normalized Amplitude Quotient (NAQ): A time-domain measure derivative of the glottal pulses. Correlated to the energy and defines how 'pressured' a vowel or phoneme is voiced [26].
  - Difference Between the First Two Harmonics (H1H2): Measurements of difference in glottal harmonic amplitude. This is related to the use of different pitches and harmonics in speech. This feature is used in studies on voice and language quality.
  - Harmonic Richness Factor (HRF): Computed via the ratio of the sum of glottal harmonics amplitude and fundamental frequency amplitude. This feature is usually used in studies on vocal quality and speaking disorders. HRF in modal speech is higher than HRF in stressed speech.
- Phonological Features: There are 18 log-likelihood ratio descriptors computed, corresponding to the phonological classes from the Phonet toolkit [23]. These features present how single phonemes were pronounced Table I.

- Prosody Features: Prosodic features are computed from continuous speech based on duration, fundamental frequency (F0), and energy. Average (avg), min, max, tilt, contour, skewness (skw), mean-square-error(mse), and, kurtosis (kurt) are computed for F0 and energy of voiced and unvoiced segments of the audio. The oscillation state of the vocal cords determines if the phoneme is voiced or unvoiced [27]. Unvoiced segments are prosodical phases where the sound is still produced in the mouth but not using the lung-full breath. The energy of the first unvoiced

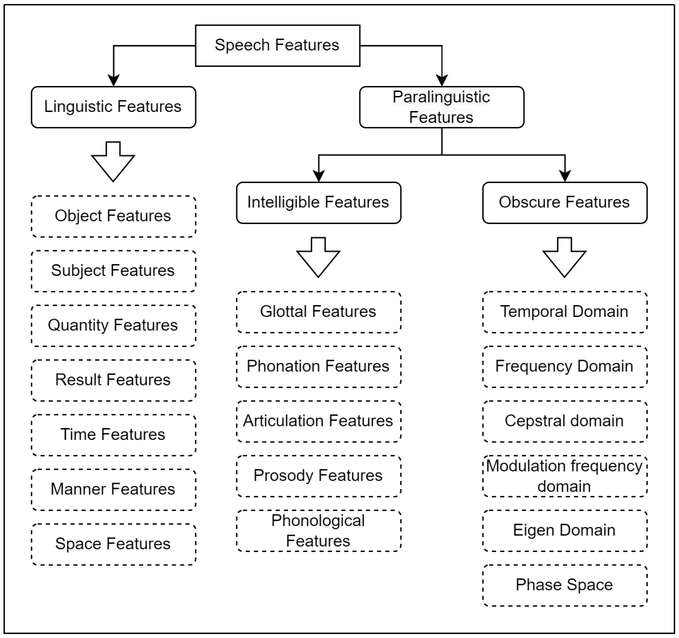

Fig. 1: Lingual and paralingual speech features

vocalic	/a/, /e/, /i/, /o/, /u/
consonantal	/b/, /tS/, /d/, /f/, /g/, /x/, /k/, /l/, /λ/, /m/, /n/, /p/, /r/, /r/, /s/, /t/
back	/a/, /o/, /u/
anterior	/e/, /i/
open	/a/, /e/, /o/
close	/i/, /u/
nasal	/m/, /n/
stop	/p/, /b/, /t/, /k/, /g/, /tS/, /d/
continuant	/f/, /b/, /tS/, /d/, /s/, /g/, /λ/, /x/
lateral	/l/
flap	/r/
trill	/r/
voiced	/a/, /e/, /i/, /o/, /u/, /b/, /d/, /l/, /m/, /n/, /r/, /g/, /λ/
strident	/f/, /s/, /tS/
labial	/m/, /p/, /b/, /f/
dental	/t/, /d/
velar	/k/, /g/, /x/
pause	/sil/

TABLE I: Phonological Classes in Phonet Toolkit

segments shows how 'strong' the first voiced phoneme is ended. Tilt is calculated from a linear estimation of F0 in a segment.

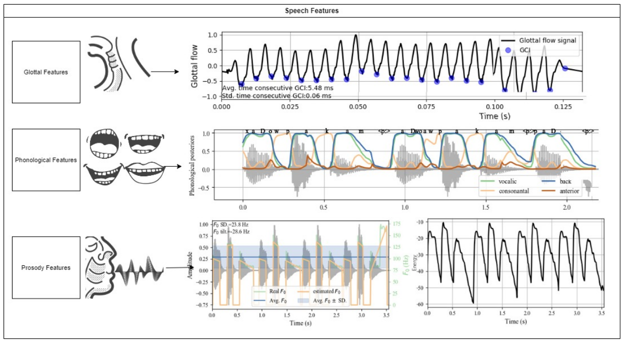

Fig. 2: Diagram of Extracted Features[1]

III. Related Work

Voice quality and prosody features were reviewed in different emotional speech conditions [28]. The authors concluded that speech features indicate the emotional state of the speaker but related features should be extracted and selected carefully. In another study with multiple corpora [29], researchers studied acoustic features in detail and found that the shape of the pitch is not as emotionally informative as the contour of the pitch in speech. They concluded that although these features are not the most informative among obscure paralinguistic features, they are more emotionally prominent. Despite the initial studies using similar datasets and obscure paralinguistic features that were extracted to pursue speech analysis, there have been recent studies with varying extensions, too. CK et al. [30] investigated SER using obscure bispectral features from speech and glottal waveforms on different classifiers to analyse the influence of different feature sets on identifying levels of stress. Their research showed that although bispectral features are more convenient for certain stress levels, employing also glottal features increases the performance of most feature sets.

Input dimensionality and the speech duration for an efficient SER were examined [31]. The authors used only two sets of obscure paralinguistic features, namely Subharmonic-to-Harmonic Ratio (SHR) and Wavelet Packet Transform to compose a feature vector of size 384 for each speech utterance. They were able to reduce the input without compromising accuracy and reported that it is possible to recognize emotions from a one-second-long speech by using sufficient feature sets.

Understanding the influence of individual features is a cumbersome job since the number of available obscure paralinguistic features goes up to over 1500 including the low-level descriptors for some feature extraction libraries [32]. Further elements of SER on the obscure paralinguistic feature level include signal processing steps, normalization, the level of segmentation and windowing, decoding, and implementation of deep networks for AI-driven feature extraction.

IV. Research Method

In this study, we used Berlin Database of Emotional Speech (EmoDB) which is a public, acted speech corpus in which 10 actors speak out 10 sentences in seven emotional classes (i.e., happy, angry, anxious, fearful, bored, disgusted, and neutral) [33]. Although the dataset contains a limited number of speech utterances, the acted nature and labeled genders of the dataset make it suitable for our study. To make intuitive explanations, we followed the activation and pleasantness axes of Russel's circumplex emotion model for negative/positive relationships between emotions [34]. Intelligible paralinguistic speech features were computed using DisVoice feature extraction library. Glottal [22], prosodic [24] and phonological [23] static features were computed. The feature set consists of 247 features that were derived from the descriptors and their functions.

We used several classification algorithms with different characteristics; Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K Nearest Neighbors (KNN), and Multilayer Perceptron (MLP). A train-test split was applied with an 8:2 ratio. The random seed is set to 0 to ensure repeatability. First, we trained all the models using the full dataset with the full set of intelligible paralinguistic features. Then, we utilized Shapley Additive explanations (SHAP) to determine the importance of each feature for a specific prediction. We applied independent t-tests to statistically assess the level of difference in features between genders and emotional classes. After including only a subset of the most informative features, we re-trained the models and reported the metrics against the full set. Secondly, we divided the dataset to achieve a binary classification problem to calculate emotion-specific results. To do that, we isolated an emotion class and randomly selected the same size of data points from the rest of the dataset and repeated this for each class. For each binary classification problem, we trained each model to get SHAP values for each feature and re-trained the models with a subset of the most informative features, and reported the metrics. We repeated this task with the full dataset versus with only one gender. In the end, as the combination of eight emotion items (i.e., sets

[1] Visuals from the DisVoice website: https://disvoice.readthedocs.io/

of each emotion, neutral, and all emotions), three gender items (i.e., sets of each gender and both genders together), and five classifiers, we executed 120 reports Figure 3.

In summary, our research design for each research question is as follows:

- RQ1 / EXP1 : Create a balanced subset for each emotional class by randomly sampling from the excluded classes against the focused class. Train 5 different models and calculate SHAP values for each feature.
- RQ2 / EXP2 : For each model and SHAP output from EXP1, train the same algorithm with a subset[2] of features. For each instance in EXP1, another training with the subset was executed.
- RQ3 / EXP3: Repeat EXP1 for each gender, and compare feature distributions for genders. Disclose the most robust and prone features for emotion classes and genders.

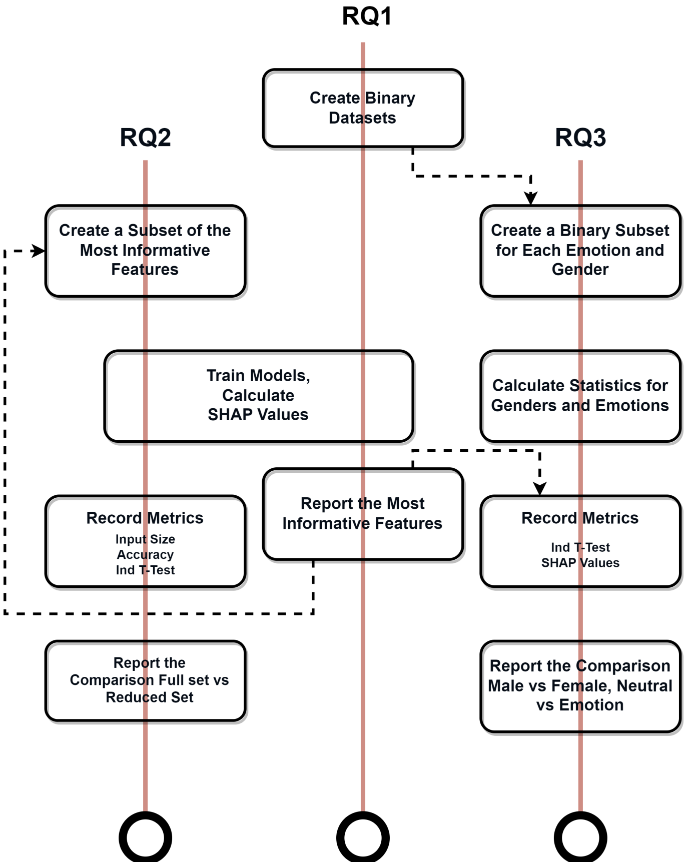
RQ1
RQ2
RQ3
Create Binary Datasets
Create a Subset of the Most Informative Features
Create a Binary Subset for Each Emotion and Gender
Train Models, Calculate SHAP Values
Calculate Statistics for Genders and Emotions
Report the Most Informative Features
Record Metrics
Input Size
Accuracy
Ind T-Test
Record Metrics
Ind T-Test
SHAP Values
Report the Comparison Full set vs Reduced Set
Report the Comparison Male vs Female, Neutral vs Emotion

Fig. 3: Experiment Flow Diagram

V. Results

Our experiments showed that changes in emotional state of the speaker are traceable by analyzing the paralinguistic speech features. These corresponding features differ based on gender, and also possibly other characteristics of speakers.

Most of the information of speech audio is carried by the fundamental frequency (F0) and its descriptors. Still, the emotional state influences different paralinguistic speech features such as HRF, voiced segment, and statistical functions of features. By investigating this affinity, we can enhance model explainability, enable feature-level fusion for multimodal applications, and create personalized models that can be deployed on personal devices.

A. Emotion Classes and Gender

1) Anger: The most informative feature for anger in 'male' speakers are NAQ, HRF, and descriptors of the F0. For 'female' speakers, the most informative features are the amount of energy of unvoiced segments, MSE of F0 (Table II). Having a low average in energy of the unvoiced segments along with a high MSE of F0 means the speaker has peaks both energy-wise and pitch-wise.

Results show that having pressured and breathy phonemes are likely to be a sign of anger in male speakers. Furthermore, having peaked fundamental frequency and poorly vocalized phonemes are likely to be a sign of anger in female speakers.

Feature	**Angry Mean**	**Neutral Mean**	**p-value**	**M Mean**	**FM Mean**	**p-val M-FM**
avg NAQ	0.01702	0.01110	**p < .05**	0.01510	0.01873	**p < .05**
F0avg	219.6217	147.2545	**p < .05**	203.5409	234.0224	**p < .05**
F0msemax	1663.6183	293.1819	**p < .05**	1136.4265	2135.7305	**p < .05**
avg1Eunvoiced	-56.4941	-42.1615	**p < .05**	-54.8625	-57.9551	p > .05
avg HRF	316.4295	164.1660	p > .05	6135.2749	11537.5397	p > .05

TABLE II: Statistics of the Most Informative Features, Class = "Anger", M=Male, FM= Female

2) Boredom: The most informative features for boredom in male speakers are the measurements of the tilt of F0, the energy and duration of the voiced and unvoiced segments, and F0 of the last voiced segment. The most informative features for boredom in female speakers are F0 of the last voiced segments, H1H2, the tilt of F0, and HRF.

Having a higher standard deviation and lower tilt in the last voiced segment means that male speakers finish their bored speech by more gradually lowering the amplitude of the F0 compared to females. For female speakers, the difference between the fundamental and the second frequency gets lower and the speech gets prosaic. As boredom occurs in the speech, HRF gets higher for female speakers than it gets for male speakers (Table III).

Feature	**Bored Mean**	**Neutral Mean**	**p-value**	**M Mean**	**FM Mean**	**p-val M-FM**
F0 tilt max	133.5792	302.5252	**p < .05**	76.1814	177.2515	p > .05
F0 tilt std	208.3606	283.6525	**p < .05**	162.3647	243.3574	**p < .05**
Vrate	2.1566	2.8953	**p < .05**	2.2233	2.10594	p > .05
Last F0	121.1462	116.2330	p > .05	90.0769	144.7859	**p < .05**
Last Energy	-19.1551	-18.0254	p > .05	-17.8513	-20.1472	**p < .05**

TABLE III: Statistics of the Most Informative Features, Class = "Boredom", M=Male, FM= Female

3) Disgust: The most important features for disgust in male speakers are descriptors of F0, maximum of tilt, HRF, and descriptors of phonological stop and nasal phonemes. For female speakers, descriptors of phonological flaps, and

[2] To keep the variance while reducing the input size, the subset size is fixed to the five most informative features for each experiment.

descriptors of HRF and F0 are the most informative features according to SHAP values.

Results show that male speakers have a closer to normal distribution of F0 in disgusted speech, pronunciation of nasal phonemes gets weaker, and stop phonemes get more pressurized compared to female speakers. An increase in the standard deviation of HRF and log-likelihood of flap phonemes are important features for classifying disgust in female speakers (Table IV). This shows that female speakers tend to change the breathiness of their voices frequently. Female speakers also put a varying emphasis on their flap phonemes when they are disgusted in contrast to males. *(p-value disgust vs neutral* $M > 0.05, FM < 0.05$*)*

Feature	Disgust Mean	Neutral Mean	p-value	M Mean	FM Mean	p-val M-FM
F0 kurtosis	0.0713	0.5580	p > .05	1.0381	-0.2324	**p < .05**
stop mean	0.2318	0.2952	**p < .05**	0.3883	0.1826	**p < .05**
nasal mean	-0.4059	-0.2722	**p < .05**	-0.5168	-0.3710	**p < .05**
flap mean	0.0889	0.0281	**p < .05**	0.0703	0.0947	p > .05
std HRF	3059.8902	2915.2193	p > .05	405.3850	3894.1630	**p < .05**

TABLE IV: Statistics of the Most Informative Features, Class = "Disgust", M=Male, FM= Female

4) Anxiety / Fear: The most informative features for male speakers and class fear are the average F0 of the last voiced segment, QOQ, HRF and descriptors of log-likelihood of the back phonemes. For female speakers, the most informative features are the log-likelihood of stop phonemes and the tilt of the F0.

Results show that male speakers increase the number of quasi-glottal openings and speak the phonemes in rush. For female speakers, on the other hand, the deviation from F0 and the speed of this deviation increases significantly compared to male speakers (Table V).

Feature	Fear Mean	Neutral Mean	p-value	M Mean	FM Mean	p-val M-FM
Last F0	188.4097	116.2330	**p < .05**	152.8503	227.2017	**p < .05**
avg QOQ	0.5035	0.4659	**p < .05**	0.5002	0.5071	p > .05
stop mean	0.2585	0.2952	p > .05	0.3475	0.1615	**p < .05**
std F0	32.7045	25.3972	**p < .05**	25.9408	40.0830	**p < .05**
tilt F0	-211.0739	-132.4231	**p < .05**	-218.2681	-203.2257	p > .05

TABLE V: Statistics of the Most Informative Features, Class = "Anxiety / Fear", M=Male, FM= Female

5) Happiness: For the emotion class happiness, the most impactful features for male speakers are H1H2, log-likelihood of consonantal phonemes and descriptors of F0. For female speakers, happiness alters the features of NAQ, labial phonemes, and kurtosis of the tilt of energy at the unvoiced segments.

Results show that for male speakers, H1H2 and maximum tilt of F0 get higher, meaning that they tend to use more distinct pitches while speaking happily. For female speakers, on the other hand, the standard deviation of NAQ gets higher, meaning that they use distinctive intensities for some phonemes, especially labial and dental classes. (Table VI).

6) Sadness: The most informative features for sad speech by male speakers are descriptors of pause duration in the speech, HRF, and descriptors of F0. For female speakers with sad speech, the most informative features are log-likelihood of strident and vocalic phonemes along with descriptors of F0 and HRF.

Feature	Happy Mean	Neutral Mean	p-value	M Mean	FM Mean	p-val M-FM
H1H2	10.3847	12.5532	**p < .05**	12.4206	9.1354	**p < .05**
F0tiltMin	-888.4789	-619.7270	**p < .05**	-1071.7582	-776.0120	**p < .05**
Consonantal	0.5764	0.7731	**p < .05**	0.6437	0.5352	**p < .05**
NAQ	0.0174	0.0111	**p < .05**	0.0150	0.0189	**p < .05**
Labial	-0.1386	-0.0991	p > .05	-0.1056	-0.1588	p > .05

TABLE VI: Statistics of the Most Informative Features, Class = "Happiness", M=Male, FM= Female

Results show that male speakers are not easily distinguishable from female speakers in sad speech. Although the disparity of most features are not distinctive, male speakers lower their tonal variance more than female speakers. For female speakers on the other hand, their pronunciation on strident and vocalic phonemes gets more emphasised while speaking sadly (Table VII).

Feature	Sad Mean	Neutral Mean	p-value	M Mean	FM Mean	p-val M-FM
Dur. Pause	0.3812	0.0327	**p < .05**	0.3216	0.4214	**p < .05**
std HRF	3713.8988	2915.2193	p > .05	339.9793	5993.5741	p > .05
mse F0	235.6563	293.1819	p > .05	120.5276	313.4459	**p < .05**
Strident	0.1352	-0.1427	**p < .05**	0.0701	0.1792	**p < .05**
kurt. Vocalic	0.7297	-0.4862	**p < .05**	0.3097	1.0135	**p < .05**

TABLE VII: Statistics of the Most Informative Features, Class = "Sad", M=Male, FM= Female

B. Reduced Features

For each emotion and gender, we re-trained our models with only a subset of the emotions. While processing time, CPU usage, and input size decreased, we recorded the accuracy, model size, and input size. These subsets were taken in a way that while we reduce the input size, we aimed to keep the variance under the scope. As SHAP values differed among models, we decided to keep the five most informative features for each experiment (See Appendix A for full feature subsets for each model).

Each classifier ended up with 12 experiments with the full set and 12 with a subset. Results show that selectively reducing the input does not have a significant effect on accuracy (p > .05) on four of five models (Table VIII). Thus, we can achieve corresponding results while we drastically reduce the input size and computation power for extracting the features (Figure 4).

Model	Full Set Mean	Sub Set Mean	Ind. T-Test Value	p Significance
rf	0.8613	0.7728	t = 2.3870	**p < .05**
dt	0.6890	0.770	t = -1.6942	p > .05
knn	0.6090	0.5701	t = 0.7682	p > .05
mlp	0.6024	0.5462	t = 1.0809	p > .05
svm	0.6880	0.6253	t = 1.5898	p > .05

TABLE VIII: Individual t-test results for accuracy of each model, Full Set vs Sub Set

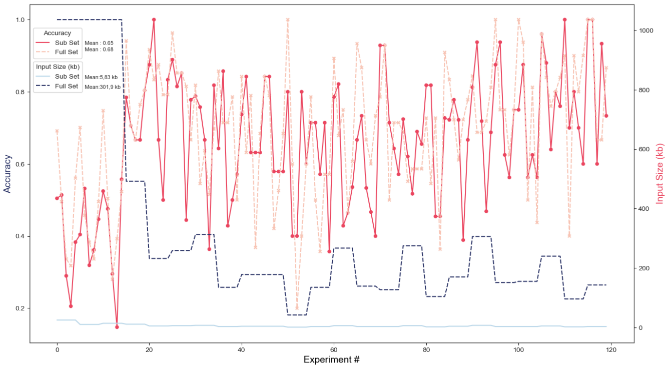

Fig. 4: Accuracy (red) and Input Size (blue), Full Set vs Subset

C. Feature Robustness Against Genders and Emotions

Our experiments with paralinguistic speech features show that male and female speakers express emotions in different ways. Although there are no absolute features that are explicitly informative for specific genders, some features still remain more informative for specific emotions and genders. The activation level and positiveness of emotions seem to be relative to the pronunciation of phonemes.

For male speakers, HRF is more informative for most emotional classes compared to female speakers. As HRF is related to the monotonousness of speech, this yields the information that male speakers tend to alter their excitement more potently. For negative emotions, NAQ remains informative as it is associated with the pressure of general speech. The amount of pauses and the combination of unvoiced segments and pauses in speech provides more information for some emotional classes for males. In reverse to mean of changes in some phonemes, tilt and kurtosis of those changes yield more influential information for some emotional classes.

Female speakers also have distinctive features for certain emotion classes. Changes to the means of strident, voice, and nasal phonemes produce decisive information. This yields the information that females tend to alter the presence of pronunciation on certain phonemes. Contrasting with males, NAQ is informative for positive emotions in female speakers. The amount of voiced and unvoiced segments along with their sequential combinations are consistently more informative for negative emotion classes.

For all the emotion classes and genders, the base frequency and its descriptors stand highly informative. Notably, MSE and tilt of the F0 are highly informative as they are a perspective of how the base speaking frequency is altered by the speaker. HRF is also informative for all emotion classes as it displays the eventfulness of a speech, therefore, it changes a lot with the emotional changes. As the amount of arousal decreases, the amount of change in most phonemes becomes more decisive for both genders. The difference in GCI between genders increases when a negative emotion is expressed. Even though GCI is not highly informative for classifying fear yet it is still significant between genders.

VI. DISCUSIONS

In this study, we performed an analysis of intelligible paralinguistic speech features for SER among different gender and emotion classes. We found out that the gender of the speaker and the emotion to be expressed influence the speech in an observable way. Changes in glottal, phonological, and prosodical speech features are deterministic for different emotions and genders. Negative emotions tend to have common feature importance on the descriptors of F0 and HRF. The levels of pleasantness and activation of emotions have a coherent relationship with intelligible paralinguistic speech features such as HRF, energy, and, amplitude quotients. Pronunciations of vowels are altered significantly when the stress in speech changes. As the activation level increases, tilt and other descriptors of F0, and H1H2 become more informative according to SHAP values. For negative emotions (i.e. sadness), the duration and amount of pauses in the speech become more informative.

Our experiments on reducing the input size based on feature informativeness show that if prior information about the speaker is known such as gender, it is possible to reduce the input size drastically while not compromising on accuracy. This enables efficient, personalized, and context-adaptive models while adding to the affective behaviour summarisation discussion for SER.

Analyses on feature robustness to gender show that there is an opportunity for development in feature engineering and selection in SER. Some emotions are easier to capture from a speech by implementing bias. While some frequency-related paralinguistic speech features remain informative, some features become redundant for certain emotions and genders. Our findings add to the discussion of contextualized modelling of SER for use cases in which only the occurrence of a set of emotions is concerned such as customer service calls.

Our observations have theoretical and practical implications. Researchers can use our results to further study the effect and explainability of speech features on SER. A new data fusion approach can be established by employing paralinguistic features along with acoustic features for more generalizable and personalized SER. The developers of SER solutions can use our results for feature engineering to train lightweight models which enables these models to run on edge devices. Additionally, our results can guide practitioners in creating more resilient SER models, that are robust against discriminating biases based on gender and ethnicity.

VII. CONCLUSIONS

In this study, we investigated SER with paralinguistic speech features to analyze differences between the emotional expressions of different genders. Our results show that intelligible paralinguistic speech features can be informative for SER, and there are significant differences in certain speech features between genders. Additionally, under circumstances where prior knowledge about the speaker is available, it is possible to shrink the input size, processing time, and possibly processing power. Also, if the presence of a specific emotion in a speech is pondered for a context-aware application, it is possible to reduce the model complexity by utilizing a rule-based decision layer of feature selection. Finally, the informativeness of features depends on the expressed emotion and the speaker's gender. Therefore, feature selection can also be done considering speaker characteristics.

A. Future Work

Obscure paralinguistic speech features are altered by the age of the speaker [35]. This might make the employment of intelligible paralinguistic speech features even more essential for the assistance of elderly people in hospitals and smart home environments. Similarly, the expertise of the speaker affects the way the words are spoken [36]. Using paralinguistic speech features on periodic verbal quizzes in classroom environments could yield information about students' learning experiences and understanding of the subject.

Recognizing emotions under noisy conditions is a challenging task for SER applications. As most of the intelligible paralinguistic speech features are expected to be independent of recording quality contrary to obscure paralinguistic features, it is possible to implement these features for SER under noisy conditions and low-quality audio communications such as phone calls. Also, it is still possible to extend the performance and capabilities by implementing gender information and speaker-specific speech habits.

While our study's findings remain consistent, it is essential to acknowledge that we solely utilized one acted dataset for our analysis. In order to enhance the generalizability of our results, it is recommended to incorporate a diverse range of datasets including both acted and non-acted settings. Furthermore, to strengthen the robustness and reliability of the results, conducting the study with multiple random seeds and applying corresponding significance tests would be beneficial. By adopting these measures, we can achieve a more comprehensive understanding of the phenomena under investigation and maintain the overall validity of our conclusions.

VIII. Ethical Impact Statement

In this study, we used EmoDB, which is a public dataset that has been commonly used in affective computing studies. The dataset contains no personally identifiable attributes, posing no privacy threats. The data have been collected from individuals who responded to a call for participation. Twenty participants contributed, equally representing both sexes. The dataset consists of German speech which might affect the generalizability of our findings. Specifically, some of the intelligible paralinguistic speech features might be language dependent, thus, they potentially differ in distribution based on the spoken language. Further research is required to explore the effect of spoken language on the distribution of intelligible paralinguistic speech features.

Our study shows that a subset of intelligible speech features represents particular emotions better. Also, the distribution of some of such features is significantly different between genders. These findings lead to several contributions; Firstly, SER models can be trained with a subset of features that are much smaller yet comparably effective. Secondly, with prior knowledge of the gender of a user, gender-specific SER models can be utilized, thus, yielding better performance. Moreover, our findings can be used to select features that are informative regarding emotions but uninformative regarding gender. This might be very useful in scenarios where gender is considered sensitive. The second contribution comes with an indirect risk. Our findings indicate that better SER models can be developed when the gender of the speaker is known. This might motivate practitioners to collect gender information purely to improve SER performance, and the collected gender information might be used in discriminating against users. However, it should also be noted that in any scenario where speech signals are analyzed, gender can be inferred without difficulty. Thus, we consider our work alleviating this issue by enabling practitioners to train SER models with gender-unspecific speech features.

Appendix

Appendix : Table of Feature Informativeness and Statistical Tests (Anonymized). [37]

References

[1] R. Cowie, "Perceiving emotion: towards a realistic understanding of the task," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, pp. 3515–3525, Dec. 2009.

[2] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression.," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[3] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Transactions on Affective Computing*, vol. 10, pp. 85–99, Jan. 2019.

[4] J. Zhao, R. Li, J. Liang, S. Chen, and Q. Jin, "Adversarial domain adaption for multi-cultural dimensional emotion recognition in dyadic interactions," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, ACM, Oct. 2019.

[5] P. P. Dahake, K. Shaw, and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and support vector machine," in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, IEEE, Sept. 2016.

[6] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.

[7] S. G. Koolagudi, N. Kumar, and K. S. Rao, "Speech emotion recognition using segmental level prosodic analysis," in *2011 International Conference on Devices and Communications (ICDeCom)*, IEEE, Feb. 2011.

[8] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 112–118, 2018.

[9] T.-W. Sun, "End-to-end speech emotion recognition with gender information," *IEEE Access*, vol. 8, pp. 152423–152438, 2020.

[10] W. Fitch, "The evolution of speech: a comparative review," *Trends in Cognitive Sciences*, vol. 4, no. 7, pp. 258–267, 2000.

[11] L. Zhang and H. Xing, "A study on lexical knowledge and semantic features of speech act verbs based on language facts," in *Lecture Notes in Computer Science*, pp. 187–197, Springer International Publishing, 2022.

[12] "Disvoice - references." https://disvoice.readthedocs.io/en/latest/reference.html. Accessed: 2022-04-13.

[13] D. Mitrović, M. Zeppelzauer, and C. Breiteneder, "Features for content-based audio retrieval," in *Advances in Computers*, pp. 71–150, Elsevier, 2010.

[14] M. Swain, A. Routray, and P. Kabisatpathy, “Databases, features and classifiers for speech emotion recognition: a review,” *International Journal of Speech Technology*, vol. 21, pp. 93–120, Jan. 2018.

[15] T. Giannakopoulos, “pyAudioAnalysis: An open-source python library for audio signal analysis,” *PLOS ONE*, vol. 10, p. e0144610, Dec. 2015.

[16] T.-Y. Huang, J.-L. Li, C.-M. Chang, and C.-C. Lee, “A dual-complementary acoustic embedding network learned from raw waveform for speech emotion recognition,” in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, Sept. 2019.

[17] H. Lee, Y. Largman, P. Pham, and A. Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS’09, (Red Hook, NY, USA), p. 1096–1104, Curran Associates Inc., 2009.

[18] I. Chaturvedi, T. Noel, and R. Satapathy, “Speech emotion recognition using audio matching,” *Electronics*, vol. 11, p. 3943, Nov. 2022.

[19] X. Zhang, Y. Sun, and S. Duan, “Progress in speech emotion recognition,” in *TENCON 2015 - 2015 IEEE Region 10 Conference*, IEEE, Nov. 2015.

[20] W. Han, H. Ruan, X. Chen, Z. Wang, H. Li, and B. Schuller, “Towards temporal modelling of categorical speech emotion recognition,” in *Interspeech 2018*, ISCA, Sept. 2018.

[21] J. Guillory, J. Spiegel, M. Drislane, B. Weiss, W. Donner, and J. Hancock, “Upset now?: emotion contagion in distributed groups,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, May 2011.

[22] E. A. Belalcázar-Bolaños, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, T. Haderlein, and E. Nöth, “Glottal flow patterns analyses for parkinson’s disease detection: Acoustic and nonlinear approaches,” in *Text, Speech, and Dialogue*, pp. 400–407, Springer International Publishing, 2016.

[23] J. Vásquez-Correa, P. Klumpp, J. R. Orozco-Arroyave, and E. Nöth, “Phonet: A tool based on gated recurrent neural networks to extract phonological posteriors from speech,” in *Interspeech 2019*, ISCA, Sept. 2019.

[24] N. Dehak, P. Dumouchel, and P. Kenny, “Modeling prosodic features with joint factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2095–2103, Sept. 2007.

[25] T. Drugman, B. Bozkurt, and T. Dutoit, “A comparative study of glottal source estimation techniques,” 2020.

[26] P. Alku, T. Bäckström, and E. Vilkman, “Normalized amplitude quotient for parametrization of the glottal flow,” *The Journal of the Acoustical Society of America*, vol. 112, pp. 701–710, Aug. 2002.

[27] A. I. Koutrouvelis, G. P. Kafentzis, N. D. Gaubitch, and R. Heusdens, “A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 316–328, Feb. 2016.

[28] R. S. Sudhakar and M. C. Anil, “Analysis of speech features for emotion detection: A review,” in *2015 International Conference on Computing Communication Control and Automation*, IEEE, Feb. 2015.

[29] C. Busso, M. Bulut, and S. Narayanan, “Toward effective automatic recognition systems of emotion in speech,” in *Social Emotions in Nature and Artifact*, pp. 110–127, Oxford University Press, Nov. 2013.

[30] C. K. Yogesh, M. Hariharan, R. Ngadiran, A. H. Adom, S. Yaacob, C. Berkai, and K. Polat, “A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal,” *Expert Systems with Applications*, vol. 69, pp. 149–158, Mar. 2017.

[31] M. Gupta, S. S. Bharti, and S. Agarwal, “Emotion recognition from speech using wavelet packet transform and prosodic features,” *Journal of Intelligent & Fuzzy Systems*, vol. 35, pp. 1541–1553, Aug. 2018.

[32] J.-L. Li and C.-C. Lee, “Attention learning with retrievable acoustic embedding of personality for emotion recognition,” in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, Sept. 2019.

[33] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech,” in *Interspeech 2005*, ISCA, Sept. 2005.

[34] J. A. Russell, “A circumplex model of affect.,” *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, Dec. 1980.

[35] A. Dehqan, R. C. Scherer, G. Dashti, A. Ansari-Moghaddam, and S. Fanaie, “The effects of aging on acoustic parameters of voice,” *Folia Phoniatrica et Logopaedica*, vol. 64, no. 6, pp. 265–270, 2012.

[36] S. Scherer, N. Weibel, L.-P. Morency, and S. Oviatt, “Multimodal prediction of expertise and leadership in learning groups,” in *Proceedings of the 1st International Workshop on Multimodal Learning Analytics*, ACM, Oct. 2012.

[37] Huseyin Ediz Yildirim and D. İren, “Appendix - informative speech features based on emotion classes and gender in explainable speech emotion recognition,” 2023.