TOWARDS INTERPRETABLE EMOTION RECOGNITION: IDENTIFYING KEY FEATURES WITH MACHINE LEARNING

Yacouba Kaloga* **Ina Kodrasi**

Signal Processing for Communication Group, Idiap Research Institute, Martigny, Switzerland

ABSTRACT

Unsupervised methods, such as wav2vec2 and HuBERT, have achieved state-of-the-art performance in audio tasks, leading to a shift away from research on interpretable features. However, the lack of interpretability in these methods limits their applicability in critical domains like medicine, where understanding feature relevance is crucial. To better understand the features of unsupervised models, it remains critical to identify the interpretable features relevant to a given task. In this work, we focus on emotion recognition and use machine learning algorithms to identify and generalize the most important interpretable features for this task. While previous studies have explored feature relevance in emotion recognition, they are often constrained by narrow contexts and present inconsistent findings. Our approach aims to overcome these limitations, providing a broader and more robust framework for identifying the most important interpretable features.

Keywords: *Important Features, Emotion Recognition, Interpretability.*

1. INTRODUCTION

Effective communication relies on the accurate mutual perception of emotions between interlocutors. A significant portion of information exchanged in a conversation is conveyed through emotional cues—visually, as well as through voice, intonation, and other auditory features. When emotional perception is impaired, as in certain pathologies [1], communication becomes particularly challenging, underscoring the importance of preserving these cues in auditory interactions. At the same time, a growing share of our audio interactions now occurs through electronic systems, including remote meetings, phone calls, and hearing aids. While these systems process sound to optimize factors such as signal quality, signal intelligibility, energy efficiency, or latency, the preservation of emotional cues is rarely considered. Evaluating the impact of various processing schemes on emotional cues through human subjective testing is impractical, due to high costs and limited scalability. Automated systems that use machine learning models for speech emotion recognition (SER) across different sound processing schemes can provide a viable alternative. Previous studies have shown similarities between human and machine emotion perception, with machines typically relying on handcrafted acoustic features to identify key emotional cues [2, 3]. The impact of various processing schemes (or perturbations) on emotion perception can then be quantified either by analyzing how the handcrafted acoustic features are affected [4, 5] or by measuring performance degradation in models that depend on these features [4, 6]. Such approaches, however, typically rely on a narrow range of acoustic features and limited number of datasets, lacking generalizability and highlighting the need for a broader set of emotionally relevant acoustic features that can be applied across various settings.

Our work aims to identify relevant acoustic features for SER using various automatic classifiers on various datasets. More specifically, we employ six classification models, a significantly higher number compared to what is typically seen in the SER literature. Additionally, we

****Corresponding author***: *yacouba.kaloga@idiap.ch.*

use six different datasets in four different languages, with models trained multiple times using different splits of the same dataset. Utilizing multiple models (trained multiple times on different splits of the data) and datasets provides several advantages. First, extracting the relevant acoustic features across multiple models (resp. datasets) increases robustness. If a particular feature is consistently important across different models (resp. datasets), it indicates its reliability and reduces the influence of individual model (resp. dataset) characteristics. Additionally, using multiple models helps mitigate the influence of underperforming models whose feature importances are unreliable due to poor performance, as well as models that overperform by relying on features highly specific to a particular dataset. Secondly, testing feature importance algorithms on multiple models and different datasets enables us to gauge the generalizability of feature rankings across different learning paradigms and datasets. Finally, running the analysis multiple times can help mitigate the challenges posed by correlated features. Specifically, our contributions are as follows:

- We derive the most relevant acoustic features for SER across different classification models and datasets.
- We propose an approach to combine the relevance of features from each individual model and dataset, resulting in a single, robust, and generalizable list of key acoustic features.
- We experimentally demonstrate the advantages of our approach for establishing key acoustic features, showing that it outperforms using a single classifier or a single dataset in terms of robustness and generalizability.

2. RELATED WORK

Research on speech emotion perception has primarily explored key acoustic features and their impact on human perception. In this section, we review these studies and their limitations. Further, we examine how machine learning can offer more general findings by comparing its performance to human perception. Finally, we discuss why current studies leveraging this connection remain limited.

2.1 Acoustic Cues and Human Emotion Perception

Numerous studies have established a strong connection between specific acoustic features and human perception of emotions. Pitch variations play a crucial role, with studies showing that higher mean pitch values are associated with high-arousal emotions such as anger, fear, and excitement, while lower values are linked to low-arousal emotions like boredom [7–11]. Temporal aspects, such as speech rate and rhythm, and spectral measures, including formants and spectral flux, are also critical in conveying emotions [2, 12]. Additional acoustic cues, such as prosody contour, intensity, and energy-related features, further enrich the emotional expressiveness of speech [13, 14]. Importantly, these acoustic features do not function in isolation but interact in complex ways to shape emotional perception. Despite these insights, existing studies face limitations. Most experiments rely on small sample sizes due to the high cost and time required for participant-based evaluations. This leads to findings that are highly context-dependent and difficult to generalize [9]. Additionally, some studies report contradictory results, likely due to differences in methodology, dataset composition, and speaker variability (for example, [15] has inconsistent findings on sadness with respect to [16–19]). Given these constraints, machine learning offers a promising avenue for studying emotional perception, as it allows for the analysis of large-scale datasets and the identification of underlying patterns that may be difficult to discern from human-limited studies.

2.2 Machine-based SER and Human Emotion Perception

Recent studies have directly compared machine learning models to human emotion perception. For instance, [4] evaluated classification models (i.e., multilayer perceptron (MLP) and support vector machine (SVM)) against human perception under both clean and noisy conditions. Their study introduced emotion distractors, i.e., emotions present in the test set but unseen during training, to ensure that models were performing true recognition rather than simple discrimination. Their results showed striking similarities between human and machine perception, including comparable overall SER performance, similar relative accuracy across different emotions, consistent performance degradation in noisy environments, and similar confusion matrices, particularly for well-defined emotions like anger and sadness. These findings align with earlier literature [3,6], which also reported strong similarities between human and machine SER. Additionally, regression models predicting valence/arousal positioning have shown that human-based emotion placement can effectively predict machine performance on the same task [2]. This suggests that both humans and machine learning models rely on similar acoustic cues when interpreting emotions [3–5, 20].

2.3 Acoustic Cues and SER

To the best of our knowledge, a limited number of studies have quantitatively explored the most important acoustic cues for SER in machine learning-based models. Such work is carried out using mainly regression and fixed effect regression [2,5,15]. Since features importance results may typically depend on the classifier/regressor use, the generalizability of their conclusions is limited. Additionally, the number of features considered is relatively low, and some key features may be overlooked in these analyses. The study context is often limited to a small number of speakers, a few datasets, and a restricted linguistic environment. In [20], it is highlighted that models perform worse than humans in cross-dataset settings, emphasizing the lack of generalizability in these constrained setups. Furthermore, due to individual authors' choices—such as using different emotions or acoustic cues—aggregating findings to draw broader conclusions remains challenging. Given these limitations, our objective is to derive the most important acoustic cues for SER that generalize well across multiple representative datasets and models.

3. PROPOSED APPROACH

We consider D datasets, each consisting of n_d utterances $\boldsymbol{x}_i, i = 1, \dots, n_d$. Each utterance $\boldsymbol{x}_i$ is associated with an emotion label y_i, such as e.g., *happiness*, *anger*, *sadness*, etc. The SER task is a supervised task, where we train a model to learn to predict the correct emotion label y of an input utterance $\boldsymbol{x}$. In this section, we propose an approach for identifying the most relevant acoustic features for SER using various datasets and various models.

3.1 Feature Set

Each utterance $\boldsymbol{x}_i$ is represented by a fixed-size feature list $f = [f_1(\boldsymbol{x}_i), \dots, f_Q(\boldsymbol{x}_i)] \in \mathbb{R}^Q$, designed to capture diverse acoustic characteristics. To construct this feature list, we use the *ComParE_2016* feature set [21], a large collection of features ($Q = 6373$) not specifically designed for SER but known to capture a broad range of acoustic information. This set is based on extracting 65 low-level descriptors (LLDs) such as e.g., pitch, spectral energy, and auditory spectrum, along with an additional 65 LLDs derived as their temporal differences. Since LLDs are time-varying and utterances are of different lengths, a wide range of functional descriptors are then applied to these LLDs (such as various distribution moments, local extrema, and different types of quantiles) in order to construct the fixed-size feature list.

3.2 Deriving Key Acoustic Features

To identify the most important acoustic features for SER, we use classification models that inherently provide feature importance scores, quantifying each feature's contribution to the models' decision. To reliably leverage these importance scores to derive the most important acoustic features, achieving strong classification performance is essential. Additionally, special consideration must be given to correlated features. In many feature importance models, a single variable may be assigned high importance while its correlated counterparts, despite being equally relevant, may be overlooked. Conversely, importance may be spread across a cluster of correlated variables, creating the misleading impression that individual features lack significance. Addressing this challenge is a key aspect of feature importance analysis. tm This is also why our approach, which involves multiple models, datasets, and experiment repetitions, is relevant. If two features convey the same information, they should exhibit similar average or median importance across all these experimental conditions.

To identify the most important acoustic features for SER, we use M classification models and K datasets. While $M = 4$ and $K = 6$ are used in Section 5.2, the proposed approach is applicable to any number of models M and number of datasets K. Every time a model m is trained on a dataset d, with $m = 1, \dots, M$ and $d = 1, \dots, D$, the outcome is the classification performance p_d^m and the importance score of each feature $s_{q,d}^m$. Preliminary experiments have shown that aggregating all feature importance scores in one step, such as e.g., using $s_q = \frac{1}{DM} \sum s_{q,d}^m$, is inefficient. Specifically, the resulting selection of features is not as effective as using the selection of features from a single model with respect to the evaluation metrics introduced in Section 4.5. Instead, we propose to find the key acoustic features for SER by proceeding as following.

First, we aggregate the normalized feature importance scores across all considered models for each individual dataset. The aggregated importance scores $s_{q,d}$ are computed as

$$s_{q,d} = \text{median}\left\{ \frac{s^{m}_{q,d}}{\overset{\mathrm{Q}}{\underset{q=1}{\max}} \left\{ s^{m}_{q,d} \right\}} \right\}_{m=1}^{M} . \qquad (1)$$

Then, for each dataset d, we construct a new feature list f^{order}_{d} by ordering the features in f in decreasing order of importance according to the scores $s_{q,d}$. Each model is then retrained using subsets of the top-ranked features, varying from 0.5% to 20% in increments of 0.5%. This process is stopped as soon as the model achieves a performance that matches or exceeds the original performance p^{m}_{d} obtained using all features. Let $pt(m)$ denote the minimum percentage of top-ranked features required for model m to reach at least par performance. We then define the multiset M (allowing repeated elements) of all features that were used to reach this threshold

$$M = \{ f^{\text{order}}_{q,d} \mid \forall q \in [\![1, Q]\!], \forall m \in [\![1, \mathrm{M}]\!], q \leq pt(m) \}. \qquad (2)$$

Since the features correspond to statistics of LLDs, we identify the most important acoustic cues for SER by counting the occurrences of each LLD statistic in the multiset M, and ranking them based on frequency.

4. EXPERIMENTAL SETTINGS

In this section, we present the SER datasets, classification models, and feature importance algorithms employed. Finally, we conclude this section with a detailed description of model training and evaluation.

4.1 Datasets and Emotions

To derive a generalizable list of important acoustic cues, we employ six datasets commonly used in the speech emotion recognition (SER) literature, i.e., *CaFE* [22], *Emovo* [23], *EmoDB* [24], *DEMoS* [25], *Tess* [26], and *Gemep-5emo* [27]. As summarized in Table 1, these datasets span four languages—Italian, German, French, and English. According to the emotion mapping shown in Table 2, we consider utterances labeled as *neutral*, as well as those representing six core emotions, i.e., *happiness*, *(hot) anger*, *fear*, *sadness*, *surprise*, and *disgust*.

4.2 Performance Evaluation

The standard evaluation metric used in the SER literature is the Unweighted Average Recall (UAR), which is derived from the individual recalls of each emotion. A model's recall for a specific emotion is calculated by dividing the number of correctly identified instances of that emotion by the total number of instances of that emotion in the dataset. This metric is preferred over the conventional accuracy metric since it focuses on the model's capability to recognize a particular emotion without emphasizing discrimination among other emotions. The UAR is simply the arithmetic mean of the recalls across all emotions [1] .

4.3 Models

The classifiers used to compute the most important acoustic features for SER include the Linear Support Vector Machine (L-SVM) [28], Logistic Regression (LGRG) [29], Random Forest Fourier (RFC) [30], and Light Gradient Boosting Machine (LGBM) [31] classifiers. These classifiers were chosen for their inherent ability to determine feature importance. Findings are then validated using two additional classifiers, i.e., the Radial Basis Function SVM (RBF-SVM) [28] and Multilayer Peceptron (MLP) [32] (which do not offer a straightforward method for determining feature importance). These classifiers were selected because they exhibit similarities to human perception in SER [4]. Each classification model has a set of hyperparameters, which we optimize using grid search focused on key parameters, balancing training efficiency with strong performance (cf. Section 4.4).

4.4 Selection of Hyperparameters

In this section, we outline the process of selecting optimal hyperparameters for each considered model. The selection is carried out in three key steps, i.e.,:

1. *Split datasets.* To ensure an unbiased evaluation with respect to speaker and sex specificity, we carefully split each dataset into training and test sets based on two key principles. First, all utterances from a given speaker are assigned exclusively to either the training or the test set, preventing any speaker from appearing in both. Second, we strive for the best possible male-female balance among speakers in the test set. Following these principles, the splitting strategy is as follows: for datasets with more than 12 speakers, 20% of the speakers are assigned to the test set. For datasets with fewer than 12 speakers, two speakers are assigned to the test set. Exceptions are made for *Emovo* (with six

[1] Also known as the Macro Recall.

Table 1: Summary of the used emotion recognition datasets and their characteristics. Neutral utterances denote utterances with an emotionally neutral meaning, chunks refer to segments extracted (based on syntax or prosody of the speaker) from sentences, whereas non-sense utterances are plausible sequences of language sounds without any meaning. The Male (F) column summarizes the number of male (female) speakers in each dataset. French C. denotes French Canadian. The Utterances column presents the total number of utterances in each dataset (#), the number of distinct sentences uttered in various emotions when applicable (# Sentences), and the mean (μ) and standard deviation (σ) of the duration of utterances (in seconds) in each dataset.

	Datasets Characteristics			**Speakers**			**Utterances**		
Datasets	**Language**	**Acted**	**Utterance Types**	**# Male (F)**	**Native Tongue**	**Actor**	**#**	**# Sentences**	**Duration μ/σ (s)**
CaFE	French	✓	Neutral	6 (6)	French C.	✓	936	6	4.4/0.8
DEMoS	Italian	Induced	Chunks	45 (23)	Italian	✗	9697	-	2.9/1.3
EmoDB	German	✓	Neutral	5 (5)	German	✓	494	10	2.8/1.0
Emovo	Italian	✓	Neutral	3 (3)	Italian	✓	588	14	3.1/1.4
Gemep-5emo	French	✓	Non-sense	5 (5)	French	✓	50	2	2.6/1.2
Tess	English	✓	Neutral word	0 (2)	English	✓	2800	200	2.1/0.3

Table 2: Summary of the emotions and number of utterances for each emotion in the considered datasets.

	Emotion						
Datasets	**Happiness**	**Anger**	**Fear**	**Sadness**	**Neutral**	**Disgust**	**Surprise**
CaFE	120	120	120	120	60	120	120
DEMoS	1395	1477	1156	1530	332	1678	1000
EmoDB	71	127	69	62	79	46	✗
Emovo	84	84	84	84	84	84	84
Gemep-5emo	10	10	10	10	✗	✗	✗
Tess	400	400	400	400	400	400	✗
Total	2080	2218	1839	2206	955	2382	1204

speakers in total) and *Tess* (with two speakers in total), where only one speaker is used for testing.

2. *Validation.* To select the optimal hyperparameters, we use stratified 5-fold cross-validation. For each hyperparameter configuration p, the model is trained on four folds and evaluated on the remaining fold. This process is repeated five times, ensuring that each fold serves as the validation set once. The results from these five runs are then averaged, and the configuration yielding the highest average performance is selected as the optimal hyperparameter configuration p^*. This cross-validation strategy enables robust assessment of different hyperparameter combinations and ensures effective selection for each model. For the larger DEMoS dataset, due to its size, a single evaluation is performed using one fold as a dedicated validation set.

3. *Test.* Finally, we train each model on the entire training set using the hyperparameters p^* that were determined in step 2. Then, we evaluate this trained model on the test set to assess its performance on unseen data.

To further reduce bias, this procedure is repeated 3 times for each dataset (i.e., using 3 different splits of the data into training and test sets), and the average of these three runs corresponds to the model's performance on that particular dataset.

4.5 Features Importance Evaluation

In order to validate the procedure above in the experimental part, we need to be able to determine which feature importance list is better. To do so, we rank the features in order of importance for both lists. We then retrain our model using a growing percentage of features in their importance order. The feature selection that achieves the same performance as using all features with a smaller percentage of features will be considered better.

Table 3: Mean and standard deviation of the UAR across three runs for each considered model and dataset. We highlight in purple all values within one percent of the best value in each row.

Datasets	RBF-SVM	MLP	RFC	LGBM	L-SVM	LGRG	Avg. across models
CaFE	48.6±4.0	50.6±2.8	40.1±1.6	44.6±4.0	49.8±3.8	49.6±2.6	47.2±3.7
EmoDB	78.8±2.4	80.3±1.5	70.8±0.9	75.3±0.1	82.7±2.3	83.9±1.8	78.6±4.5
DEMoS	74.2±0.1	74.9±0.8	51.6±0.6	63.8±0.8	65.9±0.3	66.0±0.3	66.1±7.7
Emovo	43.2±5.8	36.7±5.0	43.5±1.7	44.2±3.3	36.1±5.1	35.7±4.7	39.9±3.8
Gemep-5emo	83.3±8.2	80.0±14.1	80.0±12.2	76.7±8.2	80.0±7.1	76.7±4.1	79.5±2.3
Tess	58.9±0.0	61.7±0.9	57.0±0.4	57.6±0.4	69.9±0.0	69.6±0.0	62.5±5.4
Average across datasets	64.5±15.2	64.0±16.2	57.2±14.2	60.4±13.0	64.1±16.5	63.6±16.3	62.5±15.5

5. EXPERIMENTAL RESULTS

In this section, we present the results of the experiments described previously, followed by an analysis of the most important features. Note that when referring to the selection of important features, we mean selecting those with the highest importance scores from a ranked list of feature importance coefficients.

5.1 Performance of all models across for all datasets using the complete feature set

Table 3 summarizes the unweighted average recall (UAR) achieved by each model on the considered datasets, using the training and validation procedure described in Section 4.4.

From a dataset-level perspective, it is evident that classification performance is significantly influenced by the specific dataset. Notably, the number of speakers appears to affect the variance observed across the three runs (each involving different speaker combinations in the test set). Datasets with fewer speakers generally show higher variance (with the Tess dataset being a notable exception). This increased variability may stem from the limited number of speakers, which restricts the model's ability to generalize and leads to overfitting on speaker-specific traits present in the training data. This observation highlights the importance of using a diverse and extensive set of datasets, as done in this study. Beyond dataset size, no clear correlations are observed between classification performance and other dataset characteristics (as listed in Table 1), such as the speakers' native language, the spoken language, or the nature of the utterances.

From a model-level perspective, the RBF-SVM, L-SVM, MLP, and LGRG models generally yield the best overall performance. However, on the EMOVO dataset, the RFC and LGBM models considerably outperform these models. Whether this is due to overfitting, where certain models capture dataset-specific cues, or simply because some models are better suited to particular datasets, this observation underscores the value of our multi-model approach. Its key strength lies in mitigating the limitations or over-specialization that any single model might exhibit.

5.2 Feature selection: multiple vs one model

To illustrate the advantage of our proposed approach for aggregating feature importance scores across models, in this section we compare the performance when models are trained under two conditions: (i) using top-ranked features derived from the aggregated importance scores across all models, as described in Section 3.2, and (ii) using top-ranked features based solely on each individual model's importance scores. These performances are compared to the baseline performance when all features are used. As outlined in Section 4.4, each experiment is repeated three times. Figure 1 shows the difference from the baseline performance for each dataset, averaged across all considered models [2] using various thresholds of top-ranked features based on either the aggregated feature importance scores or the individual feature importance scores. The presented results demonstrate the advantage of using aggregated feature importance scores across multiple models. Specifically, even with just the top-ranked 0.5% features, performance is higher for each dataset when the features are selected based on aggregated importance scores,

[2] Note that the MLP and RBF-SVM models are also included in the average in the right figure. The outcome remains largely unchanged without MLP & SVM. The marked performance enhancement isn't attributable to their inclusion.

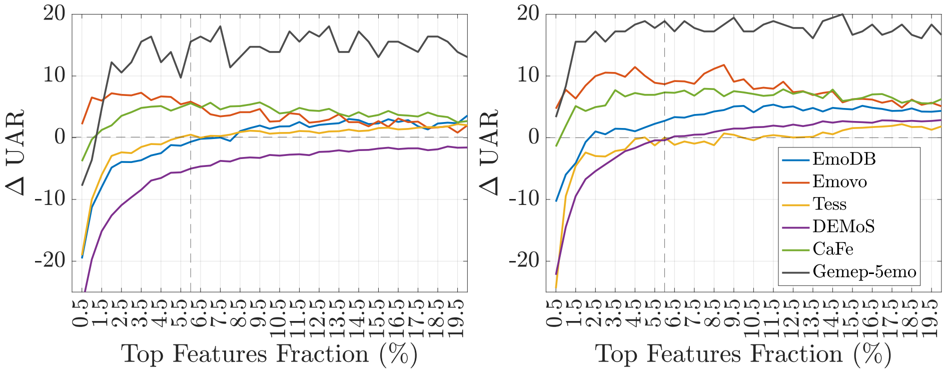

Figure 1: Average UAR difference across all models for each dataset when retraining models using top-ranked features at various thresholds, compared to baseline performance (i.e., when models are trained using all features). Left: Top-ranked features selected based on each model's own feature importance scores. Right: Top-ranked features selected using the proposed aggregated feature importance scores across all models.

as compared to using individual model rankings. Furthermore, when selecting the top-ranked 6% of features using aggregated importance scores, baseline performance is achieved for all datasets. In contrast, when selecting the top-ranked 6% of features using individual importance scores for each model, baseline performance is achieved only for three datasets.

5.3 Most important acoustic cues for SER

The results in Section 5.2 show that using 6% of top-ranked features through aggregated feature importance scores yields the same or better performance than using the complete feature set for all datasets (cf. Figure 1). Retaining these 6% of top-ranked features, we obtain a multiset of $2,282$ features, among which $1,687$ are distinct. This considerable number of distinct features underscores the complexity of the emotion recognition task, where information lies on a diverse and extensive set of features. Since each feature stems from an LLD, we determine the importance of an LLD by counting the number of times features derived from it appear in the top-ranked features (with ΔLLD features also attributed to the specific LLD it is extracted from). Figure 2 presents the normalized occurrence distribution of these LLDs. It can be observed that a small number of LLDs, such as *F0final* and *audspec_lengthL1norm*, are highly important. While the entire list of LLDs is significant, as each LLD represents an important acoustic cue for SER, the distribution shows a sharp decline, followed by a gradual, steady slope. The cutoff point, indicated by the dashed line, represents 50% of the total occurrences, corresponding to only 16 LLDs. Although this cutoff is used for illustrative purposes and does not have intrinsic significance, it highlights the top 16 LLDs that are particularly important. In the remainder of this section, we provide insights into these top 16 LLDs, which are crucial for SER, as determined using our proposed approach.

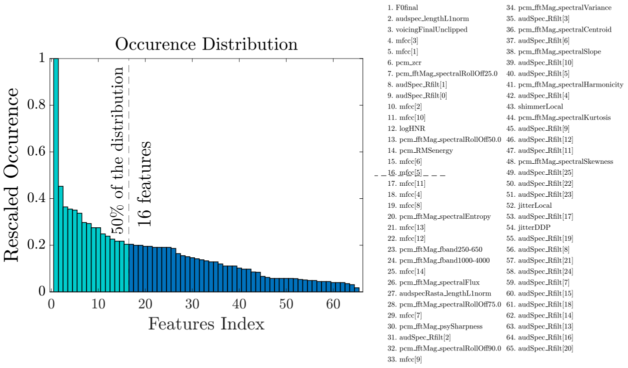

Figure 2: Normalized occurrence of LLDs in the top-ranked 6% features.

F0Final. The fundamental frequency, commonly referred to as pitch, stands out as the LLD on which models heavily rely for effective emotion recognition. This observation aligns with existing literature on emotion recognition (see Section 2.3), where the significance of pitch is widely acknowledged.

AudSpec. The auditory spectrogram is a transformation of the linear-frequency spectrogram that takes into account the non-linear frequency resolution of human hearing. Various characteristics of the auditory spectrogram as captured by the LLDs *audspec_lengthL1norm*, *audSpec_Rfilt[0]*, and *audSpec_Rfilt[1]*, appear to be important acoustic cues for SER. The LLD *audspec_lengthL1norm* is the magnitude of the l_1 norm of the auditory spectrum, broadly corresponding to the perceived loudness. The LLDs *audSpec_Rfilt[0]* and *audSpec_Rfilt[1]* are the first coefficients of the Rasta transformation used to make to make auditory spectrograms more resilient to noise, adverse conditions and other factors that can affect speech perception and analysis.

VoicingFinalUnclipped. This LLD represents the probability that F0Final is voiced. Emotional states can significantly affect phonation and voice intensity, both of which are key factors in vocal sound production. Since these factors directly influence whether F0Final is voiced, this variable is expected to vary with the speaker's emotional involvement.

MFCCs. Mel-Frequency Cepstral Coefficients (MFCCs) are derived from the cosine transform applied to lagirlogarithmic Mel scale representation of the input signal spectrum. These coefficients offer a compact representation of the overall spectral contour and are robust to variations in recording conditions and speaker characteristics. The lower MFCC coefficients, which correspond to formants, tend to be more important for SER than the higher MFCC coefficients, which capture finer spectral details and rapid variations in the spectrum.

LogHNR. The logarithm of the harmonic-to-noise ratio indicates voicing characteristics, and is hence, also relevant for SER.

PCM. Apart from the aforementioned features, the last category of important features includes Pulse Code Modulation (PCM) features related to variation and energy derived from the signal or its spectrum. Even without incorporating human auditory specificity, these features are relevant for the SER task. More specifically:

- The *pcm_RMSenergy* measure is closely associated with loudness, which is an important acoustic cue for SER.

- Emotions have significant effects on spectral energy distribution. For instance, emotions like sadness are often associated with low harmonic energy above 1 kHz, whereas emotions like anger, happiness, or fear, tend to manifest as high harmonic energy within this frequency range [13]. Hence, it's unsurprising to find several features linked to such spectral characteristics to be important for SER. These include the *pcm_fftmag_spectralRollOff25.0*, which designates the frequency demarcating 25% of the signal's energy, and the *pcm_fftmag_spectralRollOff50.0*, which designates the frequency demarcating 50% of the signal's energy.

- The (*pcm_zcr*) measure, which is the zero crossing rate quantifing the number of sign alternations in a signal, also ranks among the crucial features. Indeed, emotion-related changes in vocal expression, intensity, and timbre might affect the temporal variations captured by the zero crossing rate.

These are the characteristics emphasized by our study. The exact computation of these LLDs can be found in [33]. The next phase of this project will involve developing a method to explicitly or implicitly preserve some of these features and determine the effectiveness of such an approach.

6. CONCLUSION

In this paper, we have presented a comprehensive analysis employing multiple models and datasets to identify the most significant and robust features for the SER task. To the best of our knowledge, this is the first study to utilize such a diverse array of datasets and models to establish feature importance in the SER literature. Despite the complexities inherent in SER, we have identified several key features that are highly relevant in this context. In the future, we will extend our findings to specific emotions by repeating the experiments using recall for each emotion as the evaluation metric. Additionally, we will investigate whether the degradation of these features under adverse conditions such as background noise, channel variations, or speaker diversity, correlates with the performance drop of automatic models and human perception of emotions.

7. ACKNOWLEDGMENTS

This work was partially supported by Sonova AG, Stäfa, Switzerland.

8. REFERENCES

[1] F. Y. Leung, V. Stojanovik, M. Micai, C. Jiang, and F. Liu, "Emotion recognition in autism spectrum disorder across age groups: A cross-sectional investigation of various visual and auditory communicative domains," *Autism Research*, vol. 16, no. 4, pp. 783–801, 2023.

[2] C. Lima, T. Alves, S. Scott, and S. Castro, "In the ear of the beholder: How age shapes emotion processing in nonverbal vocalizations," *Emotion (Washington, D.C.)*, vol. 14, pp. 145–160, 11 2014.

[3] E. Coutinho and N. Dibben, "Psychoacoustic cues to emotion in speech prosody and music," *Cognition & emotion*, vol. 27, 10 2012.

[4] E. Parada-Cabaleiro, A. Batliner, M. Schmitt, M. Schedl, G. Costantini, and B. Schuller, "Perception and classification of emotions in nonsense speech: Humans versus machines," *PLoS One*, vol. 18, p. e0281079, Jan. 2023.

[5] J. Koemans, "Comparing cross-lingual automatic and human emotion recognition in background noise," Master's thesis, Radboud University, XXX, 2020.

[6] E. Parada-Cabaleiro, M. Schmitt, A. Batliner, S. Hantke, G. Costantini, K. Scherer, and B. Schuller, "Identifying emotions in opera singing: Implications of adverse acoustic conditions," 09 2018.

[7] S. Mozziconacci, "Speech variability and emotion : production and perception," *Transport Logistics - Transport Logist*, 01 1998.

[8] M. Schroder, "Expressing degree of activation in synthetic speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1128 – 1136, 08 2006.

[9] X. Luo, Q. J. Fu, and J. J. Galvin, "Vocal emotion recognition by normal-hearing listeners and cochlear implant users [published correction appears in trends amplif," *Trends Amplif*, vol. 11, no. 3, pp. 301–315, 2007.

[10] K. Hammerschmidt and U. Jürgens, "Acoustical correlates of affective prosody," *J. Voice*, vol. 21, pp. 531–540, Sept. 2007.

[11] E. Rodero, "Intonation and emotion: influence of pitch levels and contour type on creating emotions," *J. Voice*, vol. 25, pp. e25–34, Jan. 2011.

[12] S. Mozziconacci and D. Hermes, "Expression of emotion and attitude through temporal speech variations," 11 2001.

[13] M. Guzman, S. Correa, D. Muñoz, and R. Mayerhoff, "Influence on spectral energy distribution of emotional expression," *Journal of Voice*, vol. 27, no. 1, pp. 129.e1–129.e10, 2013.

[14] C.-H. Wu, J.-F. Yeh, and Z.-J. Chuang, *Emotion Perception and Recognition from Speech*, pp. 93–110. 01 2009.

[15] O. Scharenborg, S. Kakouros, and J. Koemans, "The effect of noise on emotion perception in an unknown language," pp. 364–368, 06 2018.

[16] E. Parada-Cabaleiro, A. Baird, A. Batliner, S. Hantke, and B. Schuller, "The perception of emotions in noisified nonsense speech," pp. 3246–3250, 08 2017.

[17] W. Thompson and L.-L. Balkwill, "Decoding speech prosody in five languages," *Semiotica*, vol. 2006, 01 2006.

[18] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, "Spectral moment features augmented by low order cepstral coefficients for robust asr," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 551–554, 2010.

[19] M. Pell, L. Monetta, S. Paulmann, and S. Kotz, "Recognizing emotions in a foreign language," *Journal of Nonverbal Behavior*, vol. 33, pp. 107–120, 06 2009.

[20] J. Jeon, D. Le, R. Xia, and Y. Liu, "A preliminary study of cross-lingual emotion recognition from speech: Automatic classification versus human perception," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2837–2840, 01 2013.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the munich versatile and fast Open-Source audio feature extractor," in *Proc. ACM Multimedia (MM), ACM*, pp. 1459–1462, 2010.

[22] P. Gournay, O. Lahaie, and R. Lefebvre, "A canadian french emotional speech dataset," pp. 399–402, 06 2018.

[23] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO corpus: an Italian emotional speech database," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, (Reykjavik, Iceland), pp. 3501–3504,

European Language Resources Association (ELRA), May 2014.

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," vol. 5, pp. 1517–1520, 09 2005.

[25] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. Schuller, "DEMoS: an Italian emotional speech corpus. Elicitation methods, machine learning, and perception," Feb. 2019.

[26] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020.

[27] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the geneva multimodal expression corpus for experimental research on emotion perception," *Emotion*, vol. 12, pp. 1161–1179, Oct. 2012.

[28] V. N. Vapnik, *The nature of statistical learning theory*. New York, USA: Springer-Verlag, 2000.

[29] D. H. Hosmer and S. Lemeshow, *Applied logistic regression*. New Jersey, USA: John Wiley & Sons, Inc., 2000.

[30] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[31] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: a highly efficient gradient boosting decision tree," in *Proc. International Conference on Neural Information Processing Systems*, p. 3149–3157, 2017.

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[33] F. Eyben, *Real-time speech and music classification by large audio feature space extraction*. Springer, 2015.