Language- and Speaker- Independent Emotion Recognition System Using XGBoost

XNVZ7

Supervised by Dr Christopher Carignan

Key words: emotion recognition, acoustic features, cross-lingual, XGBoost, speaker-independent

September 2023

Submitted in partial fulfilment of the MSc Speech and Language Sci-

ences Division of Psychology & Language Sciences

University College London

Word Count: 8332

Copyright notice

Project storage statement

All MSc SLS dissertations will be stored by the department electronically for administrative purposes. The department will, on request, make copies of dissertations available to UCL students for the purposes of non-commercial research and private study. Dissertations will be made available only to registered UCL students on an individual basis and will not be published publicly. Dissertations are protected by copyright and the copyright belongs to the author in each case. Should you wish that your dissertation is NOT made available to other UCL students then please check the box below:

I request that my dissertation is NOT made available to other UCL students ☐

Abstract

Background: This study emerges in the realm of acoustic emotion recognition, an extensively studied domain that primarily focuses on specific languages and a limited range of emotions. While substantial progress has been achieved in understanding emotional cues within speech, the prevailing research landscape largely revolves around single-language contexts, limiting applicability in scenarios involving multiple languages or cross-lingual interactions. Prior investigations have shed light on acoustic features linked to emotions in languages like English and Spanish, illuminating the intricate relationship between sound patterns and emotional expression. However, a noteworthy gap persists in the development of cross-lingual or multilingual emotion recognition systems. Current models, concentrated on particular languages and emotional categories, often struggle to adapt to different linguistic and cultural contexts.

Aim: This study endeavors to bridge significant lacunae within the domain of acoustic emotion recognition by devising an encompassing and versatile emotion recognition framework. The overarching goal is to surmount the limitations intrinsic to prevailing monolingual and constrained emotion recognition models, and to establish a system that is both language-independent and speaker-agnostic. By capitalizing on the aptitude of XGBoost, this study is also dedicated to finding out the most significant acoustic features cross-linguistically in identifying emotion categories.

Methods & Procedures: The study seeks to achieve this by amassing data from disparate languages and dialects, fostering a robust and adaptable emotion recognition system capable of encapsulating a spectrum of linguistic and cultural variations. The overarching intention is to transcend the confines of language, culminating in an emotion recognition paradigm primed for effective deployment in multilingual settings. Moreover, the proposed framework underscores speaker independence, addressing the challenges inherent to speaker-dependent models, thus instilling a more pragmatic and extensible solution for emotion recognition across diverse speakers.

Outcomes & Results: The study fine-tuned the emotion recognition system and trained the XGBoost model, achieving an accuracy of 54.3%, surpassing random probability of 25%. Notably, Anger displayed the highest accuracy at 0.735, while other emotions—Sadness, Neutral, and

Happiness—also exhibited promising accuracies. The top 10 acoustic features contributing to predictive performance were identified, including average alpha ratio, MFCC, F3 amplitude, and F0-related features.

Conclusions & Implications: In summary, this study advances the development of a generalizable emotion recognition system across languages and speakers. While achieving promising results, areas for improvement include addressing data disparities and expanding emotion categories. Future research can enhance cross-linguistic generalizability through techniques like data augmentation and transfer learning. The study's insights into shared acoustic features and underexplored attributes open avenues for refining multilingual emotion recognition. Implications span human-computer interaction and speech therapy, with potential for transformative applications.

What this paper adds

Building upon a backdrop of research that has predominantly focused on specific languages, this study 's major achievement lies in the development of a language- and speaker-independent emotion recognition framework that effectively surmounts the constraints of monolingual models and constrained emotions. This accomplishment underscores the model's cross-linguistic or generalizable capabilities, presenting a promising direction for multilingual emotion recognition systems. Equally significant is the study's dedicated exploration of the most influential acoustic features, specifically tailored for emotion recognition across diverse languages. By leveraging the power of XGBoost, the research delves into the shared acoustic underpinnings that facilitate accurate emotion recognition, paving the way for comprehensive and adaptable systems.

Introduction

Emotions play a crucial role in human communication, influencing our interactions, social relationships, and overall well-being. The ability to accurately recognize and understand emotions from speech is of great importance in various fields, including psychology, human-computer interaction, and social robotics. While extensive research has been conducted on emotion recognition systems, most existing models focus on specific languages or rely on language-specific acoustic features. This limitation hampers the development of a generalizable and language-independent emotion recognition system.

The aim of this dissertation is to build a robust emotion recognition system across five different languages, namely Canadian French, English, German, Mandarin, and Portuguese. By leveraging the power of XGBoost, a widely used machine learning algorithm, we seek to create a model that can effectively recognize emotions across diverse linguistic and cultural contexts. Moreover, we aim to investigate which acoustic features are most influential in distinguishing between different emotions, thereby contributing to the understanding of the underlying acoustic cues that are universal across languages.

The motivation behind this research stems from the fundamental human ability to recognize emotions regardless of the language spoken. Despite the variations in linguistic structures and phonetic characteristics, humans possess an innate capability to perceive emotions accurately, i.e., emotion recognition, and identify different individuals based on their acoustic signals, which is speaker recognition. This suggests the presence of underlying acoustic features that are inherently associated with different emotional states, transcending linguistic boundaries.

By developing a language-independent emotion recognition system, we aim to overcome the limitations of existing approaches that heavily rely on language-specific models or feature extraction techniques. Such a system has the potential to facilitate cross-cultural and cross-linguistic research, enable emotion recognition in multilingual environments, and enhance the performance of human-computer interaction systems in diverse linguistic contexts.

In order to achieve these objectives, we will employ a comprehensive dataset consisting of speech samples from individuals across different languages and emotional states. By leveraging the power of XGBoost, a gradient boosting algorithm known for its ability to handle complex patterns and high-dimensional data, we expect to build a robust model that effectively generalizes across languages and accurately predicts emotions.

The findings from this research have implications for various domains, including affective computing, speech technology, and cross-cultural communication. Understanding the universal acoustic features that underlie emotion recognition can enhance the design and development of emotionally intelligent systems, enabling more natural and context-aware human-machine interactions.

Theories of Emotional Tones

Emotional tones, also referred to as affective tones or affective states, are subjective experiences characterized by a range of psychological and physiological changes. Emotion, on the other hand, can be defined as a complex psychological state that involves subjective feelings, physiological arousal, cognitive appraisal, and behavioral expressions (Ekman, 1992). Emotion is often experienced in response to specific stimuli or events, and it influences our thoughts, actions, and interactions with others (Mullennix et al., 2002; Coutinho et al., 2013; Banse et al., 1996).

One prominent approach to the analysis of emotions is the dimensional theory of emotion. This theory, proposed by Russell (1980), suggests that emotions can be represented along two primary dimensions: valence and arousal. Valence refers to the pleasantness or unpleasantness of an emotion, ranging from positive (e.g., happiness) to negative (e.g., sadness). Arousal, on the other hand, represents the level of physiological activation associated with an emotion, ranging from low arousal (e.g., calmness) to high arousal (e.g., excitement). The circumplex model, derived from the dimensional theory, represents emotions as points in a two-dimensional space, with valence and arousal as orthogonal axes (Russell, 1980). This model provides a systematic way to categorize and compare emotions based on their positions within the space, allowing for a more nuanced understanding of emotional experiences.

Another influential theory in the field of emotion research is the theory of basic emotions. According to this theory, proposed by Ekman and Friesen (1971), there are a set of universal, biologically innate emotions that are shared across cultures. These basic emotions include happiness, sadness, anger, fear, surprise, and disgust. Basic emotions are considered to be distinct from one another and are characterized by specific facial expressions and physiological responses. Recent studies proposed that the basic emotions may include fear, anger, joy and sadness based on their examination of facial expressions and brain imaging (Gu et al., 2016, 2018 & 2019).

In addition to basic emotions, there is the concept of primary emotions. Primary emotions are the fundamental building blocks from which more complex emotional states can be derived. For example, happiness and sadness can be seen as primary emotions, from which other related emotions such as joy, contentment, or grief can arise (Plutchik, 1980). This current experiment includes four basic emotions to build the emotion recognition system, i.e., sadness, anger, happiness and neutral tones.

It is important to note that the acoustic characteristics may vary across different emotions (New et al., 2003; Leinonen et al., 1997; Murray et al., 1993). While there is some variation, certain patterns have been observed. For example, anger is often associated with higher fundamental frequency, increased intensity, and shorter duration. Happiness is typically linked to higher fundamental frequency and longer duration. Conversely, sadness is often associated with lower fundamental frequency, decreased intensity, and longer duration (Li et al., 2011; Lin et al., 2012; Chang et al., 2010; Scherer et al., 2003; Pell et al., 2009). In a study using Mandarin, it is much easier to identify negative emotions than positive ones (Wang and Lee, 2015; Liu and Pell, 2012).

On the other hand, machine learning-based methods utilize algorithms that can automatically learn patterns and features from data. One popular machine learning algorithm for emotion recognition is XGBoost (eXtreme Gradient Boosting), which is known for its ability to handle complex and high-dimensional data. XGBoost has been successfully applied in various domains, including speech and text-based emotion recognition (Chen & Guestrin, 2016). In addition to rule-

based and machine learning-based approaches, there are also multimodal approaches that integrate information from multiple modalities such as facial expressions, vocal cues, and textual content to enhance emotion analysis accuracy (Schuller et al., 2018). These multimodal approaches recognize that emotions are conveyed through various channels and that integrating information from multiple sources can provide a more comprehensive understanding of emotional states.

The Influence of Emotion on Acoustic Features

Several studies have demonstrated that even in emotionally charged speech, linguistic information about the spoken language is conveyed in a manner similar to normal speech (Batliner et al., 2011). This suggests that despite the presence of emotional expression, the acoustic features related to linguistic content are preserved. This finding is important for the development of language- and speaker-independent emotion recognition systems, as it suggests that linguistic information can be utilized alongside emotional cues for accurate recognition.

As for machine learning, statistical models are commonly used to cluster samples into distinct qualitative emotions like happiness, anger, and sadness. Deep Neural Network (DNN) and Convolutional Neural Network (CNN) are typically used for emotion recognition. Bertero and Fung (2018) introduced a convolutional neural network capable of detecting emotions of anger, happiness, and sadness with an accuracy of 66.1%. To classify these emotions in machine learning, it is necessary to model them using features derived from speech. This is typically achieved through the extraction of various categories of prosody, spectral features, and voice quality. Classifying certain emotions can be achieved through various categories, but each has its own advantages and limitations. However, for machine learning, prosody features are sometimes insufficient in accurately distinguishing between angry and happy emotions (Lee and Narayanan., 2005). Research also found that the performance of machine learning differs in different emotions. The recalls are the highest in anger and sadness in both DNN and CNN, followed by neutral emotion and happiness (Lee et al., 2011). They found that the emotion such as neutral showed a high level of confusions in machine learning (Lee et al., 2011).

Fundamental frequency (F0), also known as pitch, is a crucial acoustic feature influenced by emotion. Empirical studies have shown that different emotional states are associated with variations in F0. For example, high arousal emotions like anger and excitement tend to be accompanied by higher F0 values, reflecting a rise in vocal pitch (Banse & Scherer, 1996). In contrast, low arousal emotions such as sadness and calmness are often characterized by lower F0 values, indicating a drop in vocal pitch.

Mean amplitude, which represents the overall energy or loudness of speech, is another acoustic feature affected by emotion. Emotional states characterized by high arousal, such as anger or surprise, are typically accompanied by louder speech compared to low arousal emotions like sadness or contentment (Banse & Scherer, 1996). Variations in speech rate and rhythm have also been found to be influenced by emotion. For example, speech produced during joyful or excited states tends to be faster and more energetic, while speech associated with sadness or depression may exhibit slower tempo and reduced rhythmic patterns.

Moreover, spectral features such as formant frequencies and energy distribution have been observed to vary across different emotional states. Formant frequencies play a crucial role in speech intelligibility and are known to change depending on emotional expression (Laukka et al., 2005). Emotional states such as anger and fear have been associated with increased energy in higher frequency bands, while sadness and relaxation have shown greater energy in lower frequency regions.

Related Work

Acoustic emotion recognition, a field dedicated to automatically detecting and understanding emotions expressed in speech signals, has been extensively researched. However, most of the current studies in acoustic emotion recognition have emphasized on mono-language scenarios, with limited attention given to cross-corpus or multilingual speech emotion recognition.

While numerous studies have explored emotion recognition, most acoustic emotion recognition systems have primarily concentrated on analyzing emotional speech within one or two

languages. For example, Bassi et al. (2006) conducted studies using Spanish speech data. Some research has expanded to bilingual emotion recognition. Heracleous and Yoneyama (2018 & 2019) conducted a bilingual emotion recognition experiment using English and German corpora. However, further exploration and improvement are needed for cross-lingual or multilingual emotion recognition systems. Polzehl et al. (2010) compared mono- and multi-lingual anger recognition and found that even within a single emotion category, cross-lingual and multi-lingual recognition performed significantly poorer than mono-lingual recognition. When encountering a new language, severe degradation was observed in the system's performance (Polzehl et al., 2010). Polzehl et al. also developed a single bilingual anger recognition system using American English and German, and found that within only one emotion, the performance of multi- and mono-lingual emotion recognition system is quite similar. These investigations have provided valuable insights into the acoustic cues associated with emotions in specific languages, resulting in impressive accuracies for recognizing certain emotional states. Nevertheless, the monolingual nature of these systems presents significant challenges when applied to real-world scenarios involving multilingual or cross-lingual contexts. Such systems encounter difficulties in adapting to new languages or cultural contexts, limiting their usability and practicality. Further research is required for multi-lingual emotion recognition across several emotions.

In addition, individual different in speech may also affect the identification of emotions, which has seldom been emphasized in previous research. The identification of emotions from a speaker's voice is primarily based on voice quality features. However, these features vary from one speaker to another, which makes it challenging to utilize them in a setting where the speaker is unknown (Gobl et al., 2003). Therefore, this study is also centered on enabling the system to generalize to different speakers and be not influenced by various acoustic qualities across different individuals.

In addition to acoustic emotion analysis, cross-lingual emotion/sentiment analysis has been explored in other aspects of natural language processing. Dong and de Melo (2019) investigated

cross-lingual sentiment analysis using self-learning and multilingual BERT models. Their study focused on document classification and Chinese sentiment analysis, demonstrating the effectiveness of cross-lingual techniques in sentiment analysis tasks. Xu and Yang (2017) addressed cross-lingual text classification through a parallel corpus and adversarial feature adaptation. Their research encompassed text classification across multiple languages, including English, German, French, Japanese, and Chinese, highlighting the potential of cross-lingual approaches for handling diverse languages and achieving effective text classification in different language contexts. Their study centered around sentiment analysis in textual data, showcasing the applicability of cross-lingual techniques for sentiment analysis tasks.

In conclusion, while acoustic emotion recognition has seen significant research advancements, the field primarily focuses on single-language scenarios and a limited set of emotions. This limitation hampers the generalizability of current systems, especially in multilingual or cross-lingual contexts. Exploring and developing cross-lingual or multilingual emotion recognition models is crucial for building more versatile and culturally sensitive emotion recognition systems.

Current Work

The current work aims to bridge the research gaps in acoustic emotion recognition by developing a language- and speaker-independent emotion recognition system using XGBoost. As demonstrated before, studies that explore emotion recognition in a multilingual context involving numerous languages are scarce. Besides, most bilingual and cross-lingual recognition systems in previous studies didn’t reach a satisfying outcome. This approach restricts the generalizability of the systems, as they struggle to adapt to different languages and capture the diverse range of emotions expressed by individuals across cultures (Chen et al., 2014). Therefore, there is a pressing need to investigate and develop an approach that can effectively capture and recognize emotions in diverse languages.

By addressing these research gaps, the current work seeks to develop a more comprehensive and generalizable emotion recognition system. Firstly, the recognition system will be designed

to transcend language barriers and operate across multiple languages. This experiment will extract data from 5 completely different languages, i.e., Canadian French, English, Mandarin, Portuguese and German, and train the machine to recognize four basic emotions, which are sadness, anger, happiness as well as neutral emotion.

Secondly, the system will be speaker-independent, accommodating a wide range of speakers and their unique speech characteristics. The limitations of speaker-dependent models in adapting to new speakers and capturing individual differences pose significant challenges in real-world applications (Alluhaidan et al, 2023). By embracing speaker independence, the proposed system will provide a more practical and adaptable solution for emotion recognition tasks across various speakers.

Furthermore, the task of speech emotion recognition is inherently challenging due to the complexity and ambiguity surrounding the most discriminative acoustic features for distinguishing emotions (Chen et al., 2016). In the 2010 study of Polzehl et al, they discovered that when combining two single-language feature rankings for bilingual classification, Mel Frequency Cepstral Coefficients (MFCC) statistics are the most prominent in the merged sets. Among these, the maximum and average values derived from voiced sections are particularly significant. Regarding pitch, the most crucial aspects are the derivatives. The current work aims to further explore and identify the most significant speech/acoustic features that play a crucial role in accurate emotion recognition. By leveraging the characteristics of XGBoost as a decision-tree network, the proposed system will employ machine learning techniques to effectively capture and model these distinguishing features, enabling more robust and precise emotion recognition.

There is great potential in developing a more generalizable speech emotion recognition system that works across different language and speakers. For instance, one application can be in the development of more natural and effective human-computer interfaces, where the system can recognize the user's emotional state and respond appropriately. Another application is in the development of more effective language learning tools, where the system can provide feedback

on the user's pronunciation and intonation based on their emotional state. Multi-lingual emotion recognition could also be used in the development of more effective customer service systems, where the system can recognize the customer's emotional state and respond appropriately to their needs.

In conclusion, the current work addresses the research gaps in acoustic emotion recognition by developing a more generalizable emotion recognition system that transcends language barriers and operates across diverse speakers. By encompassing multiple languages, adapting to new speakers, and exploring powerful speech features, the proposed system will contribute to a more comprehensive and applicable solution for emotion recognition tasks.

Method

Datasets

In order to create a more language-independent emotion recognition system, this experiment used data across five distinct languages, which are Canadian French, English, German, Mandarin and Portuguese. We extracted our data respectively from five different language datasets as below.

For the data in Canadian French, this experiment selected the Canadian French Emotional speech dataset (CaFe) as our dataset, and this study only chose 12 Canadian French speakers inside the dataset. The CaFe dataset (Gournay et al., 2018) focuses on emotion recognition from acoustic and linguistic features in spontaneous conversations. It consists of audio recordings of conversations between pairs of speakers, capturing natural and spontaneous emotional expressions. The dataset covers various emotions, including anger, disgust, fear, happiness, sadness, and surprise. It provides transcriptions, annotations of emotional segments, and acoustic features extracted from the recordings. This allows us to investigate the interplay between acoustic cues, linguistic patterns, and contextual factors in natural conversational settings. The Cafe dataset offers a valuable resource for understanding the complexities of

emotional speech in real-life interactions and developing emotion recognition systems that account for conversational dynamics.

Secondly, the MSP-IMPROV dataset is used for collecting English data for this experiment. It is designed for studying spontaneous emotional speech in improvisational acting scenarios (Cowie et al., 2003). It comprises recordings of improvised dialogues performed by actors. It provides audio recordings, transcriptions, and emotion annotations, enabling researchers to analyze the acoustic characteristics and linguistic patterns associated with different emotional states. From the MSP-IMPROV dataset, audio data from 12 English speakers were used to establish English acoustic data for this experiment.

Data from 10 German speakers were extracted from the Berlin Emotional Speech Database dataset (EMO-DB). The Emo-DB dataset is a widely used benchmark dataset for studying emotional speech recognition (Schuller et al., 2009). It comprises approximately 800 sentences, which included 7 emotions, 10 actors, and 10 sentences. A group of 10 actors, consisting of 5 females and 5 males, imitated various emotions and created 10 German phrases that could be used in everyday conversations and understood in all emotional contexts (Burkhardt et al., 2005). These phrases included 5 short and 5 longer sentences. The emotions covered in the dataset include anger, boredom, disgust, fear, happiness, sadness, and neutral expressions. The Emo-DB provides audio recordings, label files indicating syllables and phones, and extensive information about perception tests evaluating emotion recognition, naturalness, syllable stress, and strength of displayed emotions. It also includes measurements of fundamental frequency, energy, loudness, duration, stress, and rhythm. The comprehensive nature of the Emo-DB enables researchers to investigate acoustic cues, linguistic features, and perceptual aspects of emotional speech across a range of emotions and actors.

Moreover, Mandarin data is from the recording of 10 Mandarin speakers in the Emo-Speech Dialogue (ESD) dataset. The Emo-Speech Dialogue dataset focuses on emotional speech within the context of human-computer interaction (Zhou et al., 2022). It comprises recorded dialogues between participants and a computer system, providing an authentic setting for emotional

expression. The dataset covers a wide range of emotions, including anger, happiness, sadness, surprise, and neutral expressions. It offers valuable audio recordings, transcriptions, and emotion labels for each dialogue, facilitating in-depth analysis of acoustic characteristics, linguistic patterns, and prosodic features associated with different emotional states. It helps researchers to study emotional speech in dynamic interactive scenarios.

The acoustic data of 12 Portuguese speakers are extracted from the VERBO dataset (Neto et al., 2018). It is a multilingual emotional speech database designed to facilitate cross-lingual research on emotion recognition. It includes recordings of emotional expressions in various languages, such as English, German, and Portuguese, while in this experiment, only the Portuguese emotional acoustic data are from the VERBO dataset. The dataset covers a diverse set of emotions, including anger, happiness, sadness, and surprise, providing a rich resource for cross-cultural and cross-lingual studies. It offers audio recordings, transcriptions, and emotion labels, enabling researchers to explore the impact of language on emotional expression and investigate the role of acoustic and prosodic features across different languages. The VERBO dataset contributes to the development of more robust and culturally sensitive emotion recognition systems.

Data Extraction

Since the multi-lingual are audio/video from different datasets, we need to extract the acoustic features using the same method so that we can get the acoustic data of the same form for the further model training and testing. In this experiment, OpenSMILE is used to extract audio-based features from speech signals. OpenSMILE (Open-Source Speaker Multimodal Interaction Learning Environment), a widely used and powerful open-source toolkit, is designed to facilitate the analysis of speech and audio data in various applications, including speech emotion recognition, speaker identification, and voice activity detection, among others (Eyben et al., 2010). The toolkit is written in C++ but provides Python bindings to make it accessible and easy to use for Python users. OpenSMILE operates by taking an audio file as input and extracting a wide range of acoustic features from the speech signal. These features are designed to capture various aspects of the speech signal, including prosody, spectral

characteristics, voice quality, and rhythm. It employs a modular approach to feature extraction. It provides a set of pre-defined feature extraction modules, each designed to compute specific types of features. Users can choose the modules they want to use and configure various parameters to customize the feature extraction process. The toolkit reads the audio signal from the input file, applies various signal processing techniques, and computes the selected features over short time frames (e.g., using sliding windows). The extracted features are then saved to a file or can be directly used as input for further processing, such as emotion recognition or speaker identification.

In this experiment, OpenSMILE is used to extract acoustic features such as: 1) Mel Frequency Cepstral Coefficients (MFCCs): MFCCs aim to capture the spectral shape in a human-like manner. By warping the frequency axis to the mel scale, taking the log, MFCCs produce a set of coefficients that compactly represent the spectral envelope. These coefficients have been shown to be highly effective for modeling speech and audio signals. MFCCs have become a staple in speech and audio processing due to their high performance and perceptual relevance; 2) Fundamental frequency (F0): F0 represents the pitch of the speech signal, which corresponds to the rate of vocal cord vibrations. It is an essential feature for speech prosody analysis and emotion recognition; 3) Formants: which represent the resonant frequencies of the vocal tract and play a crucial role in speech production and phonetic analysis; 4) Harmonics-to-noise ratio (HNR): HNR measures the ratio of harmonic components to noise in the speech signal and can provide insights into voice quality and articulation; 5) Zero crossing rate (ZCR): ZCR is a measure of the frequency content of a signal that quantifies the rate at which the signal changes from positive to negative values or vice versa. It is calculated by determining the number of times a signal crosses the zero axis within a given frame. Signals with more high-frequency content will tend to have a higher ZCR, as the signal will change signs more often; 6) Pitch-related features: such as pitch variance and pitch range, which describe the variations in pitch during speech.

After feature extraction using OpenSMILE, the IDs of the speakers are unified, which are 56 in total, meaning there are altogether 56 different speakers in this dataset across different

languages. Moreover, four emotions, i.e., neutral, happy, sad, and angry, are coded using numbers. In this experiment, neutral emotion is coded with 0, happiness with 1, sadness is represented with 2, and anger with 3. The final raw dataset containing acoustic features of all 5 languages now has the structure of (23948, 95), which means there are 23,948 samples with 95 features each, including 7 categories and 88 acoustic features (Figure 1):

```
print(data.shape)
print(data.dtypes)
```

```
(23948, 95)
emotion                          int64
ID                               int64
dataset                         object
file                            object
speaker                         object
                                 ...
MeanVoicedSegmentLengthSec      float64
StddevVoicedSegmentLengthSec    float64
MeanUnvoicedSegmentLength       float64
StddevUnvoicedSegmentLength     float64
equivalentSoundLevel_dBp        float64
Length: 95, dtype: object
```

Figure 1. Data Shape and Acoustic Features

Data Preprocessing

After extracting acoustic features from speech samples in the five languages and amalgamating them into a unified dataset for the experiment, a critical issue arises: the presence of language category imbalance. While the data distribution among the four emotion categories (neutral, happy, sad, angry) is generally equitable (as illustrated in Figure 2), we have observed a significant imbalance in the distribution of language classes (depicted in Figure 3). Specifically, the Mandarin language exhibits a substantial representation with 14,000 data points, whereas German is severely underrepresented, comprising merely 339 data points. This imbalance poses a considerable challenge in ensuring unbiased and accurate model performance across all languages during the emotion recognition process. If not addressed, the model might become disproportionately skilled at recognizing emotions from the Mandarin language due to its large representation, while struggling to accurately recognize emotions from languages like German with fewer data points. As a result, the model's effectiveness in cross-lingual emotion

recognition could be compromised, as it might not generalize well to languages with limited representation in the dataset. This scenario undermines the model's overall reliability and its ability to provide consistent and accurate emotion recognition results across diverse languages.

```
3    6449
1    6148
2    5908
0    5443
Name: emotion, dtype: int64
```

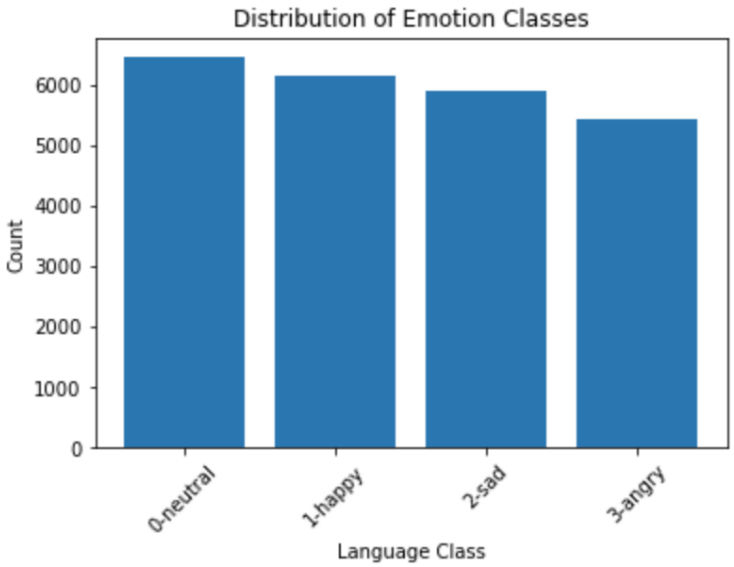

Figure 2. Distribution of Emotion Classes

```
Mandarin           14000
English             8438
Portuguese           667
Canadian French      504
German               339
Name: language, dtype: int64
```

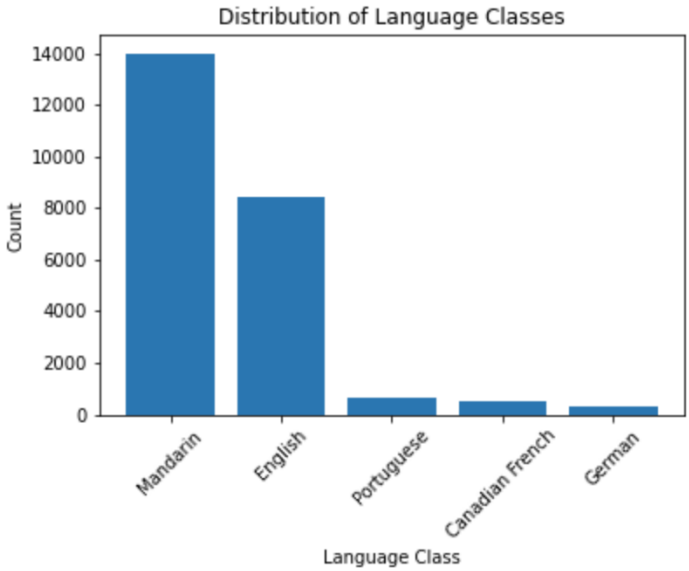

Figure 3. Distribution of Language Classes

Consequently, to address the language category imbalance during model training, we introduce class weights for the language classes. As this scenario involves a "multiclass" case, we must consider the weight parameter in XGBoost on a per-instance basis rather than per class. Thus, we need to assign weights to each data point according to the language class to which it belongs. Given the current dataset, we have identified five imbalanced classes with the following ratios:

Class Mandarin = 58.5%
Class English = 35.2%
Class Portuguese = 2.8%
Class Canadian French = 2.1%
Class German = 1.4%

To calculate the weight for each instance within each class, we divide the ratio of the German class (smallest class) by the ratios of other classes. Specifically, the weight for each instance in each class will be determined as follows:

Weight for Mandarin Class = 1.4% / 58.5% = 0.024
Weight for English Class = 1.4% / 35.2% = 0.040
Weight for Portuguese Class = 1.4% / 2.8% = 0.500
Weight for Canadian French Class = 1.4% / 2.1% = 0.667
Weight for German Class = 1.0 (since it is the smallest class and serves as the reference class)

By assigning appropriate weights to each instance based on its language class, considering the German class as the reference class, we aim to rectify the language category imbalance during the training process. This approach will promote a more balanced and reliable emotion recognition model across all languages.

Moreover, to enhance the generalizability of the emotion recognition model to new speakers, it is crucial to control the condition of speakers (IDs) during the experimental setup. By carefully managing the speaker distribution between the training and test datasets, we ensure that speakers who have already been encountered and used for training purposes do not reappear in the test set. This practice is commonly referred to as "speaker-independent evaluation".

The rationale behind this approach is to simulate a more realistic scenario in real-world applications, where the emotion recognition system will encounter speakers, it has not encountered during training. By excluding speakers from the test set if they were present in the training set, we can better assess the model's generalization capabilities to previously unseen speakers. This speaker control strategy helps to mitigate any potential bias or overfitting that might occur when the model becomes too reliant on specific speaker characteristics present in the training data. By challenging the model to recognize emotions from speech samples of entirely new speakers, we can assess its robustness and adaptability to different speaker variations, leading to a higher ability to generalize to novel speaker identities.

Overall, the speaker control technique adopted in this experiment strengthens the reliability and real-world applicability of the emotion recognition system by enabling it to handle previously unseen speakers effectively. This controlled approach aligns with best practices in acoustic emotion recognition research and contributes to producing more trustworthy and versatile emotion recognition models.

Model Selection and training

XGBoost was chosen as the primary algorithm for building the emotion recognition model. This choice is motivated by XGBoost's proven effectiveness in various machine learning applications, particularly classification problems.

XGBoost, shorthand for eXtreme Gradient Boosting, is an optimized implementation of gradient boosting machines (Chen & Guestrin, 2016). Gradient boosting is an ensemble learning technique that combines multiple weak learners, typically decision trees, to create a

stronger model. The primary concept behind gradient boosting is to iteratively add new models to the ensemble, with each new model attempting to correct the errors made by previous models. This iterative process continues until a predefined stopping criterion is met, such as reaching a maximum number of iterations or achieving a desired level of accuracy.

One of the key reasons for choosing XGBoost is its adaptability and flexibility. The algorithm can handle various types of data, including numerical, categorical, and textual data, making it suitable for emotion recognition tasks involving diverse data sources (Chen & Guestrin, 2016). Furthermore, XGBoost has been shown to outperform other machine learning algorithms in numerous benchmark datasets and competitions, indicating its superior predictive performance (Chen & Guestrin, 2016).

XGBoost offers several advantages that make it an ideal choice for the emotion recognition system. First, it employs a regularized learning approach, which helps control overfitting by adding penalty terms to the objective function (Chen & Guestrin, 2016). This regularization improves the generalization capabilities of the model, resulting in better performance on unseen data. The emotion recognition system aims to accurately predict emotions across different languages and speakers, requiring a model that generalizes well.

Second, XGBoost is computationally efficient due to its parallelization and cache-aware block structure (Chen & Guestrin, 2016). This allows the algorithm to scale effectively to large datasets and quickly produce accurate results. The emotion recognition system uses a large dataset containing numeric data across five languages, necessitating an efficient algorithm that can handle such a sizable amount of information.

Moreover, XGBoost provides a flexible framework that allows for easy integration of custom loss functions and evaluation metrics, making it adaptable to various problem domains (Chen & Guestrin, 2016). This customizability allows the XGBoost model to be tailored for the nuances of the emotion recognition dataset and task. The ability to optimize hyperparameters and loss functions helps maximize model performance for the specific emotion prediction task.

Last but not least, the capability of XGBoost to extract feature importance from the final model significantly enhances its value for this research. During the training of an XGBoost model, decision trees are built sequentially by selecting the feature that provides the best split of the data based on a criterion such as the reduction in impurity or mean squared error. The importance of a feature is calculated by aggregating its contribution to all trees, and the feature importance score is computed as the average gain or improvement in the chosen criterion. Features that consistently lead to higher improvements in the chosen criterion are assigned higher importance scores, indicating their stronger influence on the model's predictions. The importance scores are then normalized to sum up to 1, providing a relative ranking of feature importance. By quantifying the influence of various acoustic features, this feature importance analysis provides valuable insights into the underlying patterns that drive emotion recognition performance. This aspect aligns perfectly with the goals of the study, delivering a comprehensive understanding of the acoustic cues that contribute to successful emotion recognition across languages and speakers.

In conclusion, XGBoost's proven effectiveness in classification tasks, adaptability to diverse data types, and numerous advantages such as regularization, computational efficiency, and flexibility make it a suitable choice for building the language- and speaker- independent emotion recognition system. The use of XGBoost in this context demonstrates its potential for creating accurate and robust models for emotion recognition across different languages and speakers. The algorithm's efficiency, regularization techniques, and flexibility help address the challenges of the large multilingual dataset and complex emotion prediction task.

Before training the model, I also employed Grid Search to find the best parameters for the model performance. Grid Search is a popular hyperparameter tuning technique used in machine learning to optimize the performance of a model by exhaustively searching through a predefined set of hyperparameter values (Bergstra & Bengio, 2012). The primary goal of hyperparameter tuning is to identify the optimal combination of hyperparameters that yields the best performance on a given dataset. Grid Search operates by evaluating the model's performance for each combination of hyperparameters in the search space and selecting the set that produces

the highest evaluation metric, such as accuracy or F1-score.

The advantages of Grid Search include its simplicity and comprehensiveness. Since it evaluates all possible combinations of hyperparameters in the search space, it guarantees that the best combination will be found, assuming that the search space contains the optimal values. This exhaustive search can provide a high level of confidence in the resulting model's performance.

Grid Search was also employed to identify the best hyperparameters for the XGBoost model. This study finetuned two primary parameters: max_depth of the decision tree and the subsample. By fine-tuning these hyperparameters, it is possible to improve the model's predictive performance and reduce overfitting. Utilizing Grid Search ensures that the model is well-optimized for the specific task of emotion recognition across multiple languages, which can ultimately lead to more accurate predictions. Below shows the best parameters for this model (Figure 4). Moreover, this study used the “multi:mlogloss” as the objective function, which calculates the log loss for each class and sums them up, providing an overall measure of how well the model is performing across all classes. To evaluate the performance, this experiment used accuracy as the evaluation metric, which is the ratio of correctly predicted instances to the total instances.

```
Fitting 5 folds for each of 40 candidates, totalling 200 fits
Best parameters:  {'max_depth': 9, 'subsample': 0.8}
Best score:  0.7382202710166104
```

Figure 4. Results for Grid Search

However, it is essential to note that Grid Search can be computationally expensive, especially when dealing with a large search space and complex models (Bergstra & Bengio, 2012). In such cases, alternative techniques like Random Search or Bayesian Optimization can be considered, as they can potentially find good hyperparameter combinations with fewer iterations (Bergstra & Bengio, 2012; Snoek, Larochelle, & Adams, 2012).

Subsequently, cross validation was also run as a double-check before actually training the model, with the outcome of around 74% accuracy under ‘mlogloss’ evalution metric during the model training. Therefore, the actual value of accuracy is exp(-0.74), i.e., around 47%This indicated that the model was trained well above the chance level (25%) on the current dataset across five languages and different individuals.

```
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 8 concurrent workers.
[Parallel(n_jobs=-1)]: Done   2 out of   5 | elapsed:  3.2min remaining:  4.9min
[Parallel(n_jobs=-1)]: Done   5 out of   5 | elapsed:  3.3min finished
Cross-validation scores: [0.73835732 0.7358171  0.73722834 0.73976856 0.73997741]
```

Figure 5. Cross-validation Result of the XGBoost Model

Results

Upon incorporating the optimal hyperparameters identified through Grid Search, additional parameter tuning was performed to mitigate the risks of overfitting and underfitting. This fine-tuning process aimed to strike a balance between model complexity and generalization, ensuring that the final model could accurately capture the underlying structure of the data without being overly sensitive to noise or overly simplistic.

The training outcomes, as illustrated in Figure 6, demonstrate a consistent reduction in loss for both the validation and training sets as the number of iterations increased. This trend indicates that the model was successfully learning from the data and improving its predictive performance with each subsequent iteration. Notably, the decrease in loss began to level off after approximately 150 iterations, suggesting that the model had reached a point of convergence and further training was unlikely to yield significant improvements in performance. This stabilization in loss reduction can be attributed to the effective tuning of hyperparameters and

model complexity, which prevented overfitting and underfitting while maximizing the model's ability to generalize to unseen data.

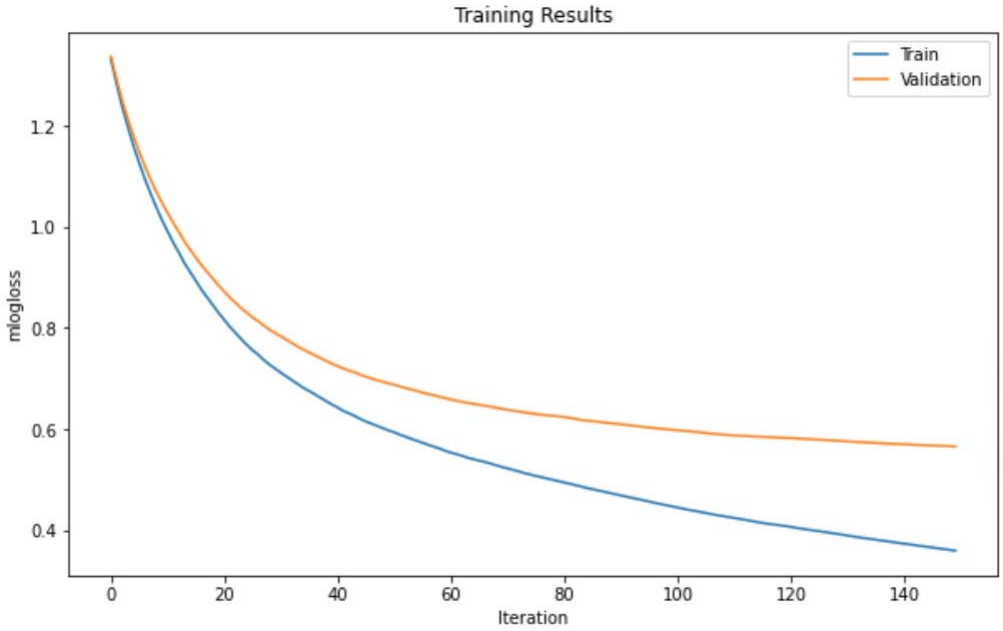

Figure 6. Training Results of the Model

The achieved test accuracy of the XGBoost-based emotion recognition system was 54.1%, which is similar but a little bit higher than the training results and still considerably higher than the random probability (25%) associated with the four emotion categories. This indicates good training of the model, that the model has become relatively generalizable not only on training set but also on test set, which means it can also perform good prediction on unseen data/language.

To gain a deeper understanding of the model's predictive performance within each of the four emotion categories, a confusion matrix was generated and is presented in Figure 7. The matrix reveals that the majority of instances within each emotion category were correctly classified, with only a small number of misclassifications scattered across the remaining categories. This observation is a promising indication of the model's generalizability across different languages, highlighting its potential for real-world applications.

Upon closer examination, it is evident that the Anger category exhibited the highest prediction

accuracy, reaching 0.735. In contrast, the accuracies of the other emotions were relatively lower, with Sadness at 0.530, Neutral at 0.491, and Happiness at 0.473. Interestingly, the model rarely misclassified instances from other emotion categories as Anger, except for instances in the Happiness category. Additionally, despite the high accuracy achieved for the Anger category, it was relatively common for instances of Anger to be misclassified as Happiness (109 cases compared to only 20 as Neutral and 21 as Sad).

In summary, the XGBoost-based emotion recognition system demonstrated a test accuracy significantly higher than random probability, showcasing its potential in recognizing emotions across various languages.

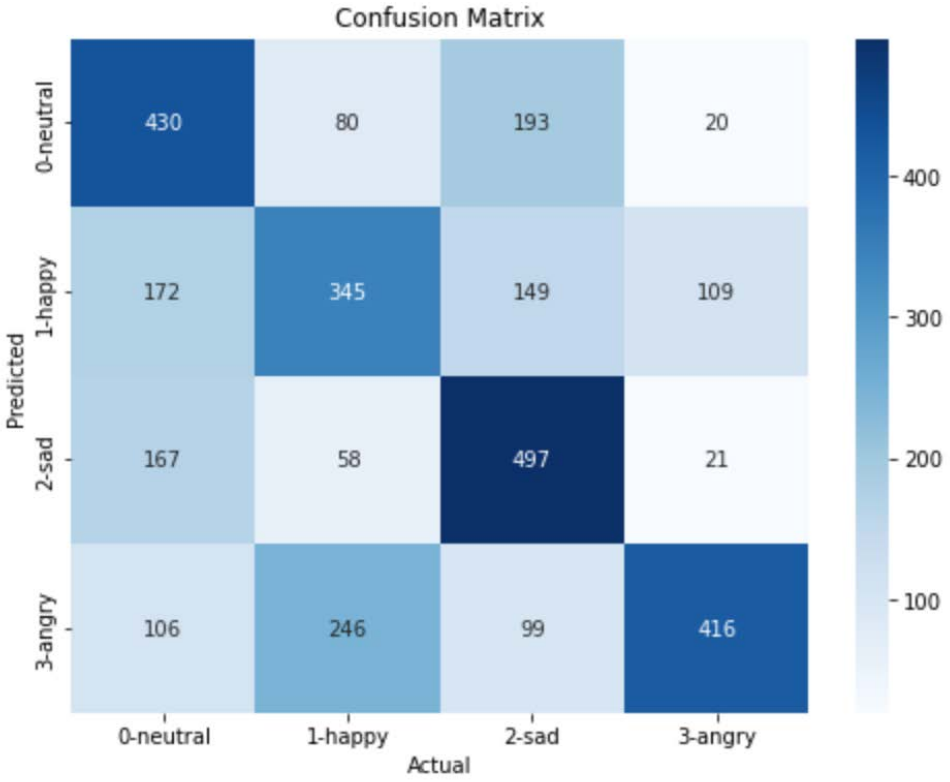

Figure 7. Confusion Matrix of the prediction result across 4 emotion categories

The top 10 important acoustic features were then extracted using the “feature_importances_” attribute of the model, which contains the importance scores for each feature in the dataset. A dictionary of feature importance was created based on that and sorted in descending order. After plotting the bar chart, + during the model prediction process, such as the average alpha ratio,

Mel-frequency cepstral coefficients (MFCC), F3 amplitude and fundamental frequency (F0). The importance of the average alpha ratio, accounts for over 0.07, which is the highest across all 88 acoustic features. It quantifies the average ratio of low-frequency energy to high-frequency energy in a signal, is crucial for distinguishing between various types of speech and emotions due to its sensitivity to vocal tract dynamics and the energy distribution across the frequency spectrum. MFCC1 and MFCC2, as the second and third most important feature, capture the spectral shape of the signal, which is related to the vocal tract shape, and are essential in analyzing different emotional states as they reflect the unique articulatory patterns associated with each emotion. The F3 amplitude provides information about the speaker's emotional state by capturing the resonance characteristics of the vocal tract. To explain this in more detail, the length and shape of the vocal tract would change while producing different speech sounds, leading to the creation of formants in the speech spectrum. As for another important feature F0, it is associated with the perceived pitch of a speaker's voice and serves as a vital indicator in emotion recognition due to its strong correlation with arousal and valence dimensions. The remaining also contributed to the model's performance, albeit to a lesser extent. These features include F0 semitone range, F0 semitone mean, Harmonics-to-Noise Ratio (HNR), F0 semitone 80th percentile, and normalized standard deviations of alpha ratio and F1 amplitude.

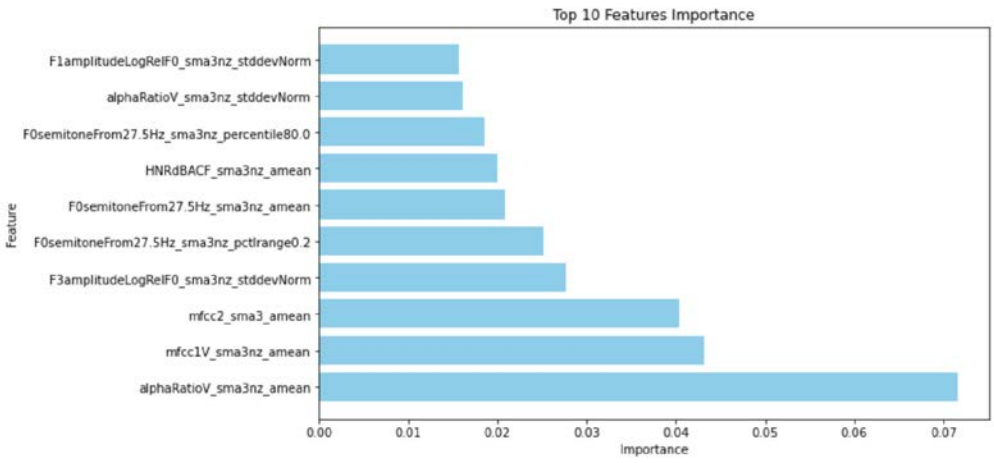

Figure 8. Top 10 Features Importance

Discussion

From the perspective of language categories, the results of this study demonstrate the progress made in cross-linguistic acoustic emotion recognition systems. In our experiment, the accuracy of the cross-linguistic emotion model doubled compared to random probability, indicating a substantial overall improvement and satisfactory performance for a cross-linguistic model. However, the majority of past research has focused on analyzing emotions in one or two languages or on identifying a specific emotion across different languages (Chen et al., 2016). Cross-linguistic or multilingual emotion recognition systems still require further development and refinement. For instance, Polzehl et al. (2010) compared single-language and multilingual emotion recognition systems for the anger emotion and found that multilingual systems performed noticeably worse than single-language systems, with limited generalizability. In contrast, our study fed the model acoustic data from five languages belonging to different language families and possessing distinct tonal systems, significantly enhancing the model's generalization capabilities and achieving performance and accuracy levels nearly on par with single-language models. This finding suggests that languages with different acoustic systems share similar acoustic features when expressing the same emotion. By leveraging these shared acoustic features, emotion recognition can be generalized across multiple languages and even cross-linguistic contexts. This not only advances our understanding of the underlying acoustic properties governing emotion perception but also provides a solid foundation for future research endeavors aimed at enhancing the performance of multilingual emotion recognition systems and their real-world applications.

Nonetheless, there are still some problems in the cross-linguistic data used in this study. The experimental data for different languages come from various databases, resulting in significant disparities in data volume between languages. For example, there are approximately 14,000 Mandarin Chinese samples while only around 340 German samples. This could result in satisfactory performance on German training data but suboptimal performance when encountering new, unlearned German data. Future research should aim to balance data distribution among languages to further improve the performance and generalizability of cross-linguistic emotion recognition models. This experiment addressed this issue by assigning

weights during model training. However, if the original dataset can be more balanced across different categories, even better results might be achieved.

Another noteworthy aspect of this study is that it not only focuses on enhancing the model's generalizability across languages but also strives to achieve generalization at the individual level, i.e., speaker-independent generalization. This aspect has been less emphasized in previous research. In our experiment, we ensured the generalization capability at the speaker level by dividing the training and testing datasets in such a way that the participants (speakers) in the testing dataset did not appear in the training dataset. Consequently, the model's performance on the testing set is based on its performance on unlearned speakers. The relatively high prediction results indicate that the model can accurately classify emotional speech for new, unknown speakers without being affected by individual differences in emotional expression. This finding is significant, as it demonstrates the potential of our cross-linguistic emotion recognition model to be applied in real-world scenarios where it may encounter a wide range of speakers with varying vocal characteristics and emotional expression styles.

Additionally, from the perspective of emotion categories, the prediction results for the "angry" emotion in this study were the most remarkable, indicating that it exhibits more distinct acoustic features and is thus relatively easier for machine learning models to recognize. This finding is consistent with previous research using deep neural networks (Lee et al., 2011). Furthermore, our study discovered a higher rate of misclassification between "angry" and "happy" emotions, with most misclassifications within the "happy" category being predicted as "angry," and vice versa. This suggests that there may be some similarity in the acoustic features exhibited by these two emotions, a conclusion that aligns with findings from earlier studies. For instance, both "angry" and "happy" emotions are associated with higher F0 values, reflecting an upward shift in voice pitch (Banse and Scherer, 1996), which is characteristic of high-arousal emotions. Simultaneously, they exhibit higher average amplitudes (Banse and Scherer, 1996), faster speech rates, or more rhythmic variations (Scherer, 1986). However, these acoustic theories have rarely been confirmed in cross-linguistic contexts in the past and have predominantly been observed in single-language or acoustically similar bilingual settings.

This study advances our understanding of these shared acoustic features by providing evidence from a cross-linguistic machine learning model that supports the notion that these characteristics are common across different languages. This not only contributes to the existing body of knowledge on emotion recognition but also highlights the potential for developing more accurate and generalizable multilingual emotion recognition systems based on shared acoustic properties.

In terms of acoustic features, the complexity and ambiguity of distinguishing vocal characteristics for emotions make it challenging for machines to recognize emotions in speech. Nevertheless, the significant acoustic features identified in previous research have been supported by our study. For instance, Polzehl et al. (2010) found that MFCCs played a dominant role in emotion recognition. Similarly, in our experiment, MFCCs were among the top ten important features for model prediction. Furthermore, the fundamental frequency (F0) and numerous F0-related acoustic features played a crucial role in the model's emotion recognition capabilities, such as F3 amplitude, F0 semitone range, F0 semitone mean, and F0 semitone 80th percentile.

Simultaneously, our experiment discovered that the mean alpha ratio is the most critical acoustic feature for cross-linguistic emotion recognition. By investigating the significance of the alpha ratio mean and other underexplored acoustic features, researchers can potentially improve the performance of cross-linguistic emotion recognition models and contribute to a more comprehensive understanding of the relationship between acoustic features and emotions across languages. Additionally, examining these acoustic features may provide insights into the underlying mechanisms that enable humans to convey and perceive emotions through speech, which could have implications for various applications, such as human-computer interaction, speech therapy, and emotion analysis in multilingual contexts.

Conclusion and Limitations

This study demonstrates progress in developing a more generalizable emotion recognition system that can transcend language barriers and operate across diverse speakers. By encompassing multiple languages and adapting to new speakers, the proposed system addresses key research gaps in acoustic emotion recognition and contributes to a more comprehensive solution for emotion-related tasks.

The experiment achieved satisfactory performance for a cross-linguistic model, with the accuracy doubling compared to random probability. This indicates a substantial overall improvement and the potential for multilingual emotion recognition systems. However, there are some limitations. First, the disparity in data volume between languages could lead to overfitting, indicating a need to balance data distributions among languages. Second, the study only considered four basic emotions; future work could expand to more complex emotions.

While the study focused on both cross-linguistic and speaker-independent generalization, there is still room for improvement in model performance and generalization capabilities. Future work could investigate techniques to enhance cross-linguistic generalizability, such as data augmentation and transfer learning. Larger and more balanced multilingual datasets would also help improve model performance.

The findings regarding shared acoustic features of emotions across languages and the significance of underexplored features provide valuable insights for future research aimed at enhancing multilingual emotion recognition. Further studies could explore the role of the mean alpha ratio and other important acoustic features in more depth. Additionally, investigating emotion recognition for under-resourced languages could expand the generalizability of such systems.

Overall, this study highlights the potential of developing more comprehensive solutions for emotion-related tasks by encompassing multiple languages and adapting to diverse speakers. Further research efforts in this direction could have implications for human-computer

interaction, speech therapy, and emotion analysis in multilingual contexts. However, additional work is needed to address limitations, improve model performance, and expand the study of cross-linguistic emotion recognition. With continued progress, more generalizable and applicable multilingual emotion recognition systems may become feasible.

Reference

Alluhaidan, A. S., Saidani, O., Jahangir, R., Nauman, M. A., & Neffati, O. S. (2023). Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. Applied Sciences, 13(8), 4750. MDPI AG. http://dx.doi.org/10.3390/app13084750

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70(3), 614–636. https://doi.org/10.1037//0022-3514.70.3.614

Bassi, A., Becerra Yoma, N., & Loncomilla, P. (2006). Estimating tonal prosodic discontinuities in Spanish using HMM. Speech Communication, 48(9), 1112-1125. https://doi.org/10.1016/j.specom.2006.03.006.

Bergstra, James & Bengio, Y.. (2012). Random Search for Hyper-Parameter Optimization. The Journal of Machine Learning Research. 13. 281-305.

Bertero, D., & Fung, P. (2017). A First Look into a Convolutional Neural Network for Speech Emotion Detection. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5115-5119). New Orleans, LA, USA. https://doi.org/10.1109/ICASSP.2017.7953131

Bertero, D., & Fung, P. (2017). A First Look into a Convolutional Neural Network for Speech Emotion Detection. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5115-5119). New Orleans, LA, USA. https://doi.org/10.1109/ICASSP.2017.7953131

Chang, H. S., Young, S. T., & Yuen, K. (Eds.). (2010). Effects of the Acoustic Characteristics on the Emotional Tones of Voice of Mandarin Tones. 20th International Congress on Acoustics, ICA 2010; 2010; Sydney, Australia.

Chen, J., Chen, Z., Chi, Z., & Fu, H. (2014). Emotion Recognition in the Wild with Feature Fusion and Multiple Kernel Learning. In Proceedings of the International Conference on Multimedia (pp. 508-513). https://doi.org/10.1145/2663204.2666277.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Coutinho, E., & Dibben, N. (2013). Psychoacoustic cues to emotion in speech prosody and music. Cognition & Emotion, 27(4), 658–684. https://doi.org/10.1080/02699931.2012.732559

Cowie, R., et al. (2003). Describing the Emotional States That Are Expressed in Speech. Speech Communication, 40(1-2), 5-32.

Dong, X. and de Melo, G. (2019). A robust selflearning framework for cross-lingual text classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6306–6310, Hong Kong, China, November. Association for Computational Linguistics.

Ekman, P. (1992). An argument for basic emotions. Cognition and Emotion, 6(3-4), 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124–129. https://doi.org/10.1037/h0030377

Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE -- The Munich Versatile and Fast Open-Source Audio Feature Extractor. In MM'10 - Proceedings of the ACM Multimedia 2010 International Conference (pp. 1459-1462). https://doi.org/10.1145/1873951.1874246.

Gu, S., Gao, M., Yan, Y., Wang, F., Tang, Y. Y., & Huang, J. H. (2018). The neural mechanism underlying cognitive and emotional processes in creativity. Frontiers in Psychology, 9, 1924. https://doi.org/10.3389/fpsyg.2018.01924

Gu, S., Wang, W., Wang, F., & Huang, J. H. (2016). Neuromodulator and emotion biomarker for stress-induced mental disorders. Neural Plasticity, 2016, 2609128. https://doi.org/10.1155/2016/2609128

Gu, S., Wang, F., Cao, C., Wu, E., Tang, Y. Y., & Huang, J. H. (2019). An integrative way for studying neural basis of basic emotions with fMRI. Frontiers in Neuroscience, 2019:00628. https://doi.org/10.3389/fnins.2019.00628

Gobl, C., & Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. Speech Communication, 40, 189-212.

Gournay, P., Lahaie, O., & Lefebvre, R. (2018). A Canadian French Emotional Speech Dataset. In Proceedings of the 9th ACM Multimedia Systems Conference (MMSys '18) (pp. 399-402). Association for Computing Machinery. https://doi.org/10.1145/3204949.3208121.

Heracleous, P., & Yoneyama, A. (2019). A comprehensive study on bilingual and multilingual speech emotion recognition using a two-pass classification scheme. PloS one, 14(8), e0220386. https://doi.org/10.1371/journal.pone.0220386

Heracleous, P., Takai, K., Yasuda, K., Mohammad, Y., & Yoneyama, A. (2018). Comparative Study on Spoken Language Identification Based on Deep Learning. In Proceedings of EUSIPCO.

Lee, C. C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion Recognition Using a Hierarchical Binary Decision Tree Approach. Speech Communication, 53, 1162-1171. https://doi.org/10.1016/j.specom.2011.06.004

Lee, C., & Narayanan, S. (2005). Toward Detecting Emotions in Spoken Dialogs. IEEE Transactions on Speech and Audio Processing, 13, 293-303. https://doi.org/10.1109/TSA.2004.838534

Leinonen, L., Hiltunen, T., Linnankoski, I., & Laakso, M. J. (1997). Expression or emotional-motivational connotations with a one-word utterance. The Journal of the Acoustical Society of America, 102(3), 1853–1863. https://doi.org/10.1121/1.420109

Li, A., Fang, Q., & Dang, J. (Eds.). (2011). Emotional Intonation in a Tone Language: Experimental Evidence From Chinese. ICPhS XVII; 2011; Hong Kong.

Lin, H. Y., & Fon, J. (Eds.). (2012). Prosodic and Acoustic Features of Emotional Speech in Taiwan Mandarin. 6th International Conference on Speech Prosody; 2012.

Liu, P., & Pell, M. D. (2012). Recognizing Vocal Emotions in Mandarin Chinese: A Validated Database of Chinese Vocal Emotional Stimuli. Behavior Research Methods, 44, 1042–1051.

Mullennix, J. W., Bihon, T., Bricklemyer, J., Gaston, J., & Keener, J. M. (2002). Effects of variation in emotional tone of voice on speech perception. Language and Speech, 45(3), 255–283. https://doi.org/10.1177/00238309020450030301

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America, 93(2), 1097–1108. https://doi.org/10.1121/1.405558

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. Speech Communication, 41(4), 603–623. https://doi.org/10.1016/s0167-6393(03)00099-2

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. Journal of Phonetics, 37(4), 417–435. https://doi.org/10.1016/j.wocn.2009.07.005

Plutchik, R. (1980). A General Psychoevolutionary Theory of Emotion. In R. Plutchik & H. Kellerman (Eds.), Theories of Emotion (pp. 3-33). Academic Press. ISBN 9780125587013. https://doi.org/10.1016/B978-0-12-558701-3.50007-7.

Polzehl, T., Schmitt, A., & Metze, F. (2010). Approaching multi-lingual emotion recognition from speech-on language dependency of acoustic prosodic features for anger detection. In Proceedings of Speech Prosody.

Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161–1178. https://doi.org/10.1037/h0077714

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldmith (Eds.), Handbook of the Affective Sciences (pp. 433–456). New York: Oxford University Press.

Schuller, B., Steidl, S., Batliner, A., Marschik, P. B., Baumeister, H., Dong, F., Hantke, S., Pokorny, F. B., Rathner, E.-M., Bartl-Pokorny, K. D., Einspieler, C., Zhang, D., Baird, A., Amiriparian, S., Qian, K., Ren, Z., Schmitt, M., Tzirakis, P., & Zafeiriou, S. (2018). The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats. Interspeech 2018, 122–126. https://doi.org/10.21437/Interspeech.2018-51

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms (arXiv:1206.2944). arXiv. http://arxiv.org/abs/1206.2944

Torres Neto, J. R., Filho, G. P. R., Mano, L. Y., & Ueyama, J. (2018). VERBO: Voice Emotion Recognition dataBase in Portuguese Language. Journal of Computer Science, 14(11). https://doi.org/10.3844/jcssp.2018.1420.1430

Wang, T., & Lee, Y. C. (2015). Does Restriction of Pitch Variation Affect the Perception of Vocal Emotions in Mandarin Chinese? The Journal of the Acoustical Society of America, 137, EL117–EL123. https://doi.org/10.1121/1.4904916

Xu, R., & Yang, Y. (2018). Cross-lingual Distillation for Text Classification (arXiv:1705.02073). arXiv. http://arxiv.org/abs/1705.02073

Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional voice conversion: Theory, databases and ESD. Speech Communication, 137, 1–18. https://doi.org/10.1016/j.specom.2021.11.006