REPORT

on

Cross Language Emotion Detection

Submitted as part of Speech Understanding
Programming Assignment 1

by

Rosalind Margaret Paulson(M23CSA526)

Indian Institute of Technology, Jodhpur
February 2025

Abstract

Cross-Language Emotion Detection is the task of identifying and analyzing emotions from text, speech, or other modalities in multiple languages. Unlike monolingual emotion detection, it aims to generalize across languages, enabling models trained in one language to work effectively in others. Cross-language speech emotion recognition (SER) is a challenging yet essential task for practical applications, especially in multilingual environments.

Chapter - 1

Introduction

1.1 What is Cross-Language Emotion Detection?

Cross-language emotion detection is the task of identifying and analyzing emotions from text, speech, or other modalities across multiple languages. Unlike monolingual emotion detection, this approach seeks to generalize across languages, allowing models trained in one language to perform effectively in others. Cross-language speech emotion recognition (SER) is a particularly challenging yet crucial task, especially for practical applications in multilingual environments.

1.2 Importance in the real world

. Speech emotion recognition (SER) is particularly useful for applications requiring natural human-machine interaction, such as web-based media and computer-assisted tutorials, where system responses depend on detected emotions. It also plays a crucial role in in-car systems, where assessing the driver's mental state can enhance safety measures. Additionally, SER can serve as a diagnostic tool for therapists and improve automatic translation systems by preserving the speaker's emotional intent, which is essential for effective communication. SER has also been widely applied in call center analytics and mobile communication [3]. Cross-language speech emotion recognition extends these capabilities to multilingual environments, enabling applications in human-robot interaction, behavioral analysis, virtual reality, and other real-world scenarios involving speakers from diverse cultural and linguistic backgrounds [6]. Developing generalized emotion recognition systems that work across different or previously unseen languages has significant practical

implications, particularly for multilingual settings [4][5][13].

1.3 Summary

This chapter provides an overview of cross-language emotion detection, highlighting its significance in real-world applications. Emotion detection plays a crucial role in various domains, including human-computer interaction, sentiment analysis, mental health monitoring, customer feedback analysis, and multilingual communication systems. With the increasing globalization of digital interactions, the ability to accurately recognize emotions across different languages is essential for improving user experience and decision-making processes.

Chapter - 2

Literature Survey

2.1 State of the art models or tools in terms of methodologies and approaches

Many studies focus on acoustic features such as cochleagrams, spectrograms, mel-cepstrograms, and fractal dimension-based representations to effectively capture emotional cues from speech. These features provide a comprehensive analysis of speech emotions by modeling frequency, temporal, and nonlinear characteristics. Ziyang Ma et al.[11] proposed emotion2vec, a universal emotion representation model that can be used to extract speech features for diverse emotion tasks.
A spectrogram is generated by computing the short-time spectrum of successive signal frames, using linear frequency scaling to represent frequency components over time. In contrast, the mel-frequency scale is a quasi-logarithmic scale that more closely approximates the resolution of the human auditory system. A cepstrogram is derived by first processing the DFT-based spectrum through a filter bank that models human auditory perception. This is followed by a logarithmic transformation and a discrete cosine transform (DCT), capturing key spectral features. A cochleagram represents how the brain processes auditory information received from the ear. It serves as a computational model of the peripheral auditory system and is constructed using auditory filtering techniques. Additionally, the fractal dimension is used to characterize and differentiate speech emotions by analyzing the irregularity, self-similarity, and nonlinearity in speech signals.[6][15]

Feature selection techniques such as Mel Frequency Cepstral Coefficients (MFCCs) and the Synthetic Minority Oversampling Technique (SMOTE) enhance the accuracy of cross-language emotion recognition models [8]. Lan-

guage agnostic approaches, such as combining mel-scaled and temporal modulation spectral representations, have proven effective in handling multiple languages [12]. Additionally, leveraging phonetic commonalities, such as shared vowel sounds across languages, improves cross-lingual SER by serving as anchors for emotion transfer [17]. Some models integrate domain adaptation strategies with Bag-of-Words (BoW) and data augmentation techniques to improve performance across languages [7]. Traditional machine learning techniques, including Gaussian Mixture Models (GMMs) and high-order Markov random fields, are used to optimize cross-language SER by effectively capturing emotional patterns across different languages. These models help in statistical pattern recognition and probabilistic modeling of speech emotions.

Advanced models such as Deep Belief Networks, Bidirectional Long Short-Term Memory(BLSTM) networks, hybrid systems combining CNNs with Long Short-Term Memory (LSTM) networks and transfer learning algorithms have achieved high accuracy in cross-lingual SER [15][16][14][10].

Quang-Anh N.D. et al. [2] proposed the Dynamic Convolutional Block Attention Module (Dynamic-CBAM) within an Attention-GRU Network to classify emotions from human audio signals. Similarly, Seunghyun Yoon et al. [18] developed a model using dual recurrent neural networks (RNNs) to encode audio and text sequences, integrating both modalities for emotion classification.

A novel approach involves the Adversarial Dual Discriminator (ADDi) network, which utilizes a three-player adversarial game to learn generalized representations without requiring labeled target data. An extension, self-supervised ADDi (sADDi), incorporates pre-training on unlabeled data and synthetic data generation to enhance domain-invariant and emotionally discriminative feature learning [9]. Another method, introduced by Mirko Agarla et al.[1], applies Semi-Supervised Learning (SSL) with Transformer-based models for cross-lingual emotion recognition. The model adapts to new languages using pseudo-labeling, where both hard and soft pseudo-labels are explored to improve classification when only a few labeled examples are available in the target domain.

Ensemble learning has been investigated for cross-corpus, multilingual SER, using a majority voting technique to improve recognition accuracy across different datasets [19]. Studies suggest that different classifiers perform best on different corpora, making ensemble learning advantageous by leveraging the strengths of multiple models rather than relying on a single classifier. A hybrid feature extraction approach combines prosodic and spec-

tral features to enhance emotion recognition. Experiments on public speech databases show that random decision forests, when trained on these hybrid acoustic features, significantly improve classification performance in speech emotion recognition [20].

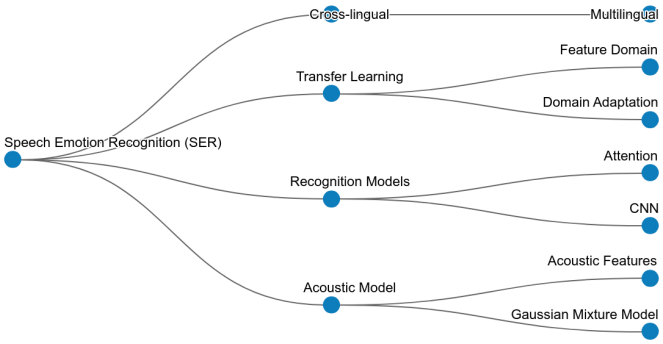

Figure 2.1: Credits: Scopus

2.2 Strengths and Limitations

2.2.1 Strengths

- **Dynamic-CBAM in Attention-GRU Network:** Efficiently captures nuanced emotional cues in audio signals by combining dynamic convolutional attention mechanisms with GRU networks, improving emotion classification accuracy . [2].
- **Dual RNN Model (Audio and Text):** Achieves 71.8% WAP on IEMOCAP, resolving "neutral class bias". Integrates audio and text sequences for comprehensive emotion prediction, leveraging complementary information from both modalities [18].
- **Deep Learning with RF and XGBoost:** Achieves high accuracy (91.25%) for Urdu data using MFCC features combined with Random Forest and XGBoost, showcasing robust performance in multilingual settings [6].
- **Multilingual Foundation Models (e.g., Wav2Vec2, Whisper):** Outperform traditional CNN-LSTM models by leveraging large-scale pre-trained models for cross-lingual SER, especially effective with limited target data and noisy environments [13].

- **Three-Layer Model for Human Perception:** Effectively estimates emotion dimensions (valence, activation, dominance) across languages using a model that aligns closely with human perception, ensuring cross-lingual applicability [4].
- **Hybrid Acoustic Features with Ensemble Learning:** Combines prosodic and spectral features in a random decision forest framework, significantly enhancing emotion recognition accuracy [20].
- **Phonetic Anchor-Based Transfer Learning:** Utilizes phonetic commonalities (like shared vowels) to anchor emotion transfer across languages, achieving better emotion classification without labeled target data [17].
- **Self-Supervised Adversarial Domain Adaptation (sADDi):** Employs self-supervised learning and synthetic data generation to create domain-invariant representations, improving cross-language SER even without labeled target data [9]. Also employs adversarial learning to create domain-invariant features, enhancing cross-language adaptability.
- **Semi-Supervised Learning with Pseudo-Labeling:** Uses a Transformer based model with pseudo-labeling strategies, demonstrating a 40% increase in unweighted accuracy for cross-lingual SER, especially with minimal labeled data [1]. Utilizes hard and soft pseudo-labels to adapt to new languages, improving accuracy with minimal labeled data in target languages.
- **Ensemble Learning for Multilingual SER:** Increases accuracy across multiple languages by leveraging ensemble techniques which combines multiple classifiers, showing significant improvements (up to 15%) in cross-corpus scenarios [19].
- **Language-Agnostic GMM-Based Approach:** Combines mel-scaled and temporal spectral representations using Gaussian Mixture Models, achieving robust emotion classification across Italian and German without language-specific tuning [12].
- **Cross-Corpus Feature Combination (MFCC + Teager Energy):** Enhances accuracy by integrating multiple feature types (MFCC and Teager Energy) combined with techniques like SMOTE, showing clear performance gains in cross-corpus settings [8].

- **CNN and BLSTM with Time-Distributed Flatten Layer:** Achieves state-of-the-art results (86.9% accuracy) for Bangla and English datasets by combining CNN and BLSTM networks, capturing both temporal and sequential emotional cues [14]. Effectively capture spatial hierarchies in acoustic features, offering high accuracy in cross-linguistic contexts [15, 16].
- **Hyperparameter Optimization for SVM:** Fine-tunes SVM parameters using hyperparameter optimization, effectively handling overlapping emotions across languages and datasets, increasing classification efficiency [16].
- **Transfer Learning with Deep Belief Networks:** Shows strong performance gains in cross-language SER by applying transfer learning techniques, indicating the benefit of leveraging multiple language datasets for training [10].
- **Bag-of-Words (BoW) with Domain Adaptation and Data Augmentation:** Combines BoW representation with domain adaptation (e.g., N-CORAL) and data augmentation, providing robustness in cross-language and noisy environments [7].
- **emotion2vec (Self-Supervised Pre-Training):** Outperforms WavLM and HuBERT with 77.64% UA on IEMOCAP. Generalizes across 10+ languages and tasks (e.g., song emotion recognition) [11].

2.2.2 Limitations

Generalization and Dataset Diversity

- **Limited Linguistic Diversity:** Current datasets often lack linguistic diversity, which hampers the generalization of models to new languages and emotional expressions. Models requires labeled emotion data for fine-tuning.[2]
- **Cross-Corpus Performance:** Models trained on specific corpora tend to perform poorly when applied to different corpora due to variations in recording conditions, languages, and accents.[10]

Semantic Disparities

- **Language Bias:** Multilingual pre-trained models are often biased towards English, leading to inaccurate emotion predictions in other languages due to semantic misinterpretations.

Complexity and Computational Resources

- **High Computational Cost:** Advanced models, especially those integrating multiple architectures or using large-scale benchmarks, require significant computational resources, which can be a barrier for real-time applications [11]. [18] depends on accurate ASR for text extraction, leading to error propagation and high computational cost due to dual RNNs.
- **Model Complexity:** The complexity of hybrid and ensemble models can make them difficult to implement and fine-tune for specific applications.[14] [19]

Model	**Strengths**	**Limitations**
CNN	Efficient for feature extraction	Struggles with long-term dependencies
RNN/LSTM	Good for sequential data processing	Computationally expensive
Transformers	Superior accuracy, less manual feature engineering	Requires large-scale data and high computational power
CNN-LSTM	Effective for dynamic speech patterns	Complexity increases training time
Dynamic Attention-GRU	High accuracy for specific tasks	Limited cross-lingual generalization
MDRE	Resolves class bias with multimodal fusion	Heavy computational requirements
emotion2vec	Generalizes across languages and tasks	High pre-training data and computational cost

Table 2.1: Comparison of different models based on strengths and limitations

2.3 Summary

The chapter explores the challenges of emotion recognition across languages, such as linguistic diversity, cultural differences in emotional expression, and the scarcity of annotated multilingual datasets. Additionally, it discusses state-of-the-art approaches, including machine learning and deep learning techniques, that facilitate effective emotion detection in multilingual settings.

Chapter - 3

Evaluation Metrics

- **Accuracy**: Measures the proportion of correctly identified emotions out of the total instances.
- **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two. It assess class-wise performance.
- **Confusion Matrix:** Visualizes misclassification patterns.
- **Mean Squared Error (MSE) & Root Mean Square Error (RMSE):** Evaluates regression-based emotion models.
- **Weighted Accuracy (WA):** Reflects overall accuracy but biased toward majority classes.
- **Unweighted Accuracy (UA):** Balances class-wise performance, critical for imbalanced datasets.
- **Unweighted Average Recall(UAR):** Calculated by averaging the recall for each class in a multi-class classification problem. It is particularly useful when the classes in your dataset are not equally represented, as it gives each class the same weight when computing the average recall.
- **Mean absolute error (MAE):** MAE between the predicted values of emotion dimensions and the corresponding average value given by human subjects is calculated.

3.1 Strengths and Limitations

1. Accuracy

- **Strengths:**
  - Suitable when classes are balanced and when you want an overall performance measure.
- **Limitations:**
  - Misleading in the case of imbalanced datasets, as the model could be biased toward the majority class.
  - Doesn't give insights into the performance across different classes.

2. F1-Score

- **Strengths:**
  - Balances precision and recall, which is important when dealing with imbalanced datasets.
  - Especially useful in classification tasks where false positives and false negatives are equally important.
- **Limitations:**
  - Can be tricky to interpret in isolation since it combines two metrics (precision and recall).
  - May not fully reflect performance in cases where you care more about the performance in specific classes (without class weighting).

3. Confusion Matrix

- **Strengths:**
  - Provides a clear visualization of misclassifications, highlighting where the model is making errors.
  - Useful for understanding class-wise performance.
- **Limitations:**
  - Doesn't summarize performance into a single number; you have to interpret it carefully.
  - Can be hard to interpret with a large number of classes.

4. Mean Squared Error (MSE) & Root Mean Square Error (RMSE)

- **Strengths:**
  - Effective in regression problems, measuring the average squared difference between predicted and actual values.
  - RMSE is more interpretable than MSE because it has the same units as the target variable.
- **Limitations:**
  - Sensitive to outliers, since the squaring process exaggerates large errors.
  - Doesn't work well when errors are not normally distributed.

5. Weighted Accuracy (WA)

- **Strengths:**
  - Gives more importance to the majority classes, reflecting the overall performance when such classes dominate.
- **Limitations:**
  - Biased toward majority classes, which may mask poor performance on minority classes.

6. Unweighted Accuracy (UA)

- **Strengths:**
  - Averages performance across all classes, regardless of class frequency.
  - Useful for imbalanced datasets since it ensures all classes contribute equally.
- **Limitations:**
  - May not fully reflect performance when the overall accuracy (weighted) is more important.
  - Less informative if the model's primary goal is to perform well on the majority class.

7. Unweighted Average Recall (UAR)

- **Strengths:**

  - Ensures equal treatment of each class's recall, preventing bias towards majority classes.
  - Useful when you're interested in how well the model is identifying each class, particularly when classes are imbalanced.

- **Limitations:**
  - Does not take precision into account, so it may not capture performance in the context of false positives.
  - Might not be a great metric when you care about precision, especially in multi-class settings.

8. Mean Absolute Error (MAE)

- **Strengths:**
  - Easy to interpret and compute.
  - Less sensitive to outliers than MSE since it uses absolute values.

- **Limitations:**
  - Does not penalize large errors as strongly as MSE, which may be a limitation in some contexts.
  - Less sensitive to variance compared to RMSE.

3.2 Summary

These metrics provide a comprehensive framework for evaluating cross-language speech emotion detection systems, ensuring robust and generalizable performance across different languages and datasets.

Chapter - 4

Open Challenges and Future Opportunities

1. **Data Scarcity**
    - High-quality emotional speech datasets are limited, making it difficult to train robust models.
    - Labeled emotion datasets are often small, expensive to collect, and hard to generalize across different domains.
2. **Cross-Language and Cross-Cultural Variability**
    - SER models tend to fail in generalizing across languages. Models trained on datasets such as IEMOCAP (focused on Western languages) struggle when applied to non-English languages (e.g., Bangla in SUBESCO).
    - Linguistic and cultural variations pose significant challenges, as emotions may be expressed differently or understood in varying contexts across languages.
    - Inconsistencies in annotation methods, emotion definitions, and interaction scenarios further complicate multilingual systems.
3. **Real-World Noisy Environments**
    - SER models tend to perform poorly in noisy real-world settings (e.g., call centers, mobile devices, background chatter).
    - Noise robustness remains a significant challenge, affecting the accuracy of emotion recognition in practical applications.
4. **Multimodal Emotion Recognition**
    - Integrating multiple cues such as facial expressions, textual input, and physiological signals (e.g., heart rate) alongside speech can improve accuracy.

- For example, combining audio-text fusion in systems like MDRE (Multimodal Deep Recognition of Emotion) leads to better performance in emotion classification.

5. **Lightweight Models for Edge Devices**
   - There is a growing demand to deploy emotion recognition models on edge devices, which require optimization for low-resource hardware without sacrificing accuracy.
6. **Cross-Cultural Generalization**
   - Models trained on data from Western contexts may not perform well on non-Western cultures due to differences in emotional expression, social norms, and language use.
   - This issue underscores the need for more diverse training datasets that capture a wide range of cultural contexts.
7. **Noise Robustness**
   - Real-world settings often include background noise or disturbances that degrade model performance.
   - Improving noise robustness in SER systems is essential to deploy them in uncontrolled environments.
8. **Opportunities for Improvement**
   - **Self-Supervised Models:** Self-supervised learning methods, such as emotion2vec, offer promising solutions by reducing reliance on labeled data.
   - **Multimodal Fusion:** As mentioned, combining speech with other modalities (e.g., text, video, physiological signals) could substantially improve the model's accuracy and robustness.
   - **Ethical Concerns:** SER models need to address biases, such as cultural stereotypes in emotion labeling, as well as privacy risks associated with emotional data collection.
9. **Quality of Recording Devices**
   - Variations in recording equipment and elicitation techniques (e.g., controlled vs. spontaneous interactions) can lead to inconsistencies in data quality, further complicating emotion recognition tasks.

10. **Accents and Speaker Diversity**

    - Variations in accents, dialects, and speaking styles (e.g., gender or age differences) also affect model performance, highlighting the need for diverse training data.

4.1 Summary

Cross-language speech emotion recognition faces challenges related to linguistic and cultural variations, annotation methods, and recording conditions.

Bibliography

[1] Mirko Agarla, Simone Bianco, Luigi Celona, Paolo Napoletano, Alexey Petrovsky, Flavio Piccoli, Raimondo Schettini, and Ivan Shanin. Semi-supervised cross-lingual speech emotion recognition. *Expert Systems with Applications*, 237:121368, 2024.

[2] Quang-Anh N. D., Manh-Hung Ha, Thai Kim Dinh, Minh-Duc Pham, and Ninh Nguyen Van. Emotional vietnamese speech-based depression diagnosis using dynamic attention mechanism, 2024.

[3] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[4] Reda Elbarougy and Masato Akagi. Cross-lingual speech emotion recognition system based on a three-layer model for human perception. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2013*, 2013. Cited by: 21; Conference name: 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2013; Conference date: 29 October 2013 through 1 November 2013; Conference code: 102470.

[5] Vladimir Hozjan and Zdravko Kačič. Context-independent multilingual emotion recognition from speech signals. *International Journal of Speech Technology*, 6(3):311 – 320, 2003. Cited by: 56.

[6] Amjad Khan. Improved multi-lingual sentiment analysis and recognition using deep learning. *Journal of Information Science*, 2023. Cited by: 20.

[7] Shruti Kshirsagar and Tiago H. Falk. Cross-language speech emotion recognition using bag-of-word representations, domain adaptation, and data augmentation. *Sensors*, 22(17), 2022.

[8] Oscar Utomo Kumala and Amalia Zahra. Indonesian speech emotion recognition using cross-corpus method with the combination of mfcc and teager energy features. *International Journal of Advanced Computer Science and Applications*, 12(4):163 – 168, 2021. Cited by: 7; All Open Access, Gold Open Access.

[9] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn Schuller. Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):1912–1926, 2023.

[10] Siddique Latif, Rajib Rana, Shahzad Younis, Junaid Qadir, and Julien Epps. Transfer learning for improving speech emotion classification accuracy. In Sekhar C.C., Rao P., Ghosh P.K., Murthy H.A., Yegnanarayana B., Umesh S., Alku P., Prasanna S.R.M., and Narayanan S., editors, *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2018-September, page 257 – 261. International Speech Communication Association, 2018. Cited by: 101; Conference name: 19th Annual Conference of the International Speech Communication, INTERSPEECH 2018; Conference date: 2 September 2018 through 6 September 2018; Conference code: 139961; All Open Access, Green Open Access.

[11] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation, 2023.

[12] Stavros Ntalampiras. Toward language-agnostic speech emotion recognition. *AES: Journal of the Audio Engineering Society*, 68(1-2):7 – 13, 2020. Cited by: 23.

[13] Moazzam Shoukat, Muhammad Usama, Hafiz Shehbaz Ali, and Siddique Latif. Breaking barriers: Can multilingual foundation models bridge the gap in cross-language speech emotion recognition? In *Proceedings - 2023 10th International Conference on Social Networks Analysis, Management and Security, SNAMS 2023*. Institute of Electrical and Electronics Engineers Inc., 2023. Cited by: 0; Conference name: 10th International Conference on Social Networks Analysis, Management and Security, SNAMS 2023; Conference date: 21 November 2023 through 24 November 2023; Conference code: 196145.

[14] Sadia Sultana, M. Zafar Iqbal, M. Reza Selim, Md. Mijanur Rashid, and M. Shahidur Rahman. Bangla speech emotion recognition and cross-

lingual study using deep cnn and blstm networks. *IEEE Access*, 10:564 – 578, 2022. Cited by: 44; All Open Access, Gold Open Access.

[15] Gintautas Tamulevičius, Gražina Korvel, Anil Bora Yayak, Povilas Treigys, Jolita Bernatavičienė, and Bożena Kostek. A study of cross-linguistic speech emotion recognition based on 2d feature spaces. *Electronics (Switzerland)*, 9(10):1 – 13, 2020. Cited by: 31; All Open Access, Gold Open Access.

[16] Anuja Thakur and Sanjeev Kumar Dhull. Language-independent hyperparameter optimization based speech emotion recognition system. *International Journal of Information Technology (Singapore)*, 14(7):3691 – 3699, 2022. Cited by: 13.

[17] Shreya G. Upadhyay, Luz Martinez-Lucas, Bo-Hao Su, Wei-Cheng Lin, Woan-Shiuan Chien, Ya-Tse Wu, William Katz, Carlos Busso, and Chi-Chun Lee. Phonetic anchor-based transfer learning to facilitate unsupervised cross-lingual speech emotion recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2023-June. Institute of Electrical and Electronics Engineers Inc., 2023. Cited by: 7; Conference name: 48th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2023; Conference date: 4 June 2023 through 10 June 2023; Conference code: 193814.

[18] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118, 2018.

[19] Wisha Zehra, Abdul Rehman Javed, Zunera Jalil, Habib Ullah Khan, and Thippa Reddy Gadekallu. Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems*, 7(4):1845–1854, Aug 2021.

[20] Kudakwashe Zvarevashe and Oludayo Olugbara. Ensemble learning of hybrid acoustic features for speech emotion recognition. *Algorithms*, 13(3), 2020.