INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Exploring the English Accent-independent Features for Speech Emotion Recognition using Filter and Wrapper-based Methods for Feature Selection

Nowshin Tabassum, Tasfia Tabassum*, Fardin Saad*, Tahiya Sultana Safa*, Hasan Mahmud, Md. Kamrul Hasan*

Systems and Software Lab (SSL), Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT), Gazipur, Bangladesh
{nowshintabassum, tasfiatabassum, fardinsaad, tahiyasultana, hasan, hasank}@iut-dhaka.edu

Abstract

In Speech Emotion Recognition (SER), significant progress has been made. Despite cutting-edge developments, faultless human-computer interaction remains a distant goal since established SoTA models cannot perceive the speaker's emotional state flawlessly. On the contrary, several studies in SER uncovered the possibility of language and culture-specific differences in this domain. Emotion recognition in speech can vary from person to person based on age, gender, language, and accent, amongst others. In this study, we explore and investigate how assorted accents of the English language influence SER. We employ four different English accents: American, British, Canadian, and Bengali English. Then we extracted a subset of best-performing accent-neutral features by incorporating filter and wrapper-based feature selection methods. Our investigations reveal that pitch, intensity, and MFCC-related features more effectively recognize emotions regardless of accent.

Index Terms: Speech Emotion Recognition, Accent, Accent-Neutral, Feature Selection, Prosodic, Spectral, Voice Quality.

1. Introduction

Emotion is a means of expressing one's perspective or state of mind to others. However, automated detection of an individual's emotional state from their speech remains a significant challenge due to the unique features of each speaker, like gender, age, and culture. There are a few universal emotions that any intelligent system with finite processing resources can be trained to recognize as needed, including Anger, Sadness, Happiness, and Neutral [1]. The emotion identification process involves utilizing spectral, prosodic, and voice quality features extracted from a speech signal. Popular spectral features, such as MFCC, LPC, and LPCC, are used to model emotional responses. Meanwhile, prosodic features, such as amplitude, pitch , rhythm , speech intensity, and voiced factors, provide additional information about the speaker's emotional state [2].

Our study aims to develop an accent-independent Speech Emotion Recognition system (SER) to enhance the human-computer interaction experience. In today's world, SER systems are widely used in customer services, call centers, and educational institutions where individuals with diverse backgrounds and accents come with their queries or problems. While previous studies have explored the impact of language on SER systems[1], this study seeks to investigate the effect of accents on the accuracy of emotion recognition.

To look into how variations of accents affect the Speech Emotion Recognition tasks, we will be using datasets of 4 different accents of English.

The main contributions of our research are -

- A robust Speech Emotion Recognition System that can correctly identify emotions despite such diverse accents.
- Determine feature subsets that are Accent-Neutral

2. Related work

Many researchers have studied existing Speech Emotion Recognition(SER) systems, but more attention needs to be paid to the impact of accents on recognition accuracy. While studying SER systems, we categorized related research into the salient features used in SER systems and feature selection methods, classifiers employed, and factors affecting SER.

2.1. Features Selection Methods and Salient Features

Paralinguistic features extracted from speech signals are used for speech emotion recognition and are independent of the language's semantic structure [3]. Recent research has used several features in speech emotion recognition, including pitch and intensity [4], formants, and Mel-frequency cepstral coefficients (MFCCs) [5]. Kächele et al. [6] developed a feature selection method using a forward-backward algorithm that adds the most promising features and removes the least salient features to obtain the final feature set, achieving an accuracy of 88.97% on the Berlin Database with an SVM classifier. Turgut Özseven et al. [7] proposed a statistical feature selection method based on emotional changes in acoustic features. This method found that reducing the number of features can increase classification success. Zhang et al. [8] used prosodic and voice quality features such as shimmer, jitter, HNR, and the first three formants. A 10% increase in recognition rate was obtained when prosodic and voice quality features were combined, as opposed to prosodic features alone. Kuchibhotla et al. [9] showed that the system's performance is not good when the prosody or spectral features are used individually. Prosodic and spectral features were combined using a feature fusion approach to enhance the speech emotion recognition system's performance.

Research has shown that speech emotion recognition often uses prosodic, spectral, and voice quality features. Moreover, combining these features through feature fusion can result in improved performance. In addition, a range of feature selection methods has been proposed, including the forward-backward algorithm, statistical feature selection, and correlation-based subset evaluator.

2.2. Classifiers Used for SER Systems

Speech Emotion Recognition (SER) involves using various classifiers to identify emotions based on speech signals. These classifiers can be broadly categorized into two groups: feature-

*The first four authors contributed equally to this work.

10.21437/Interspeech.2023-1888

based models and end-to-end models.
Feature-based models rely on a set of hand-engineered features to represent speech signals. These features, such as Mel-frequency cepstral coefficients (MFCCs), pitch, and energy, are then fed into a classifier, such as support vector machines (SVMs) [10, 11, 12], or decision trees [13], to make the final prediction. However, speech signals are often non-stationary, making non-linear classifiers more effective for SER. The most commonly used non-linear classifiers for SER include the Hidden Markov Model (HMM) [14, 15, 16], K-nearest neighbor [17], and the Gaussian mixture model (GMM) [11, 12, 18]. These models have demonstrated promising results, but the hand-engineered features limit their performance.
On the other hand, end-to-end models construct the direct mapping between speech signals and emotions without relying on hand-engineered features. These models typically use deep neural networks (DNNs) [19], such as recurrent neural networks (RNNs) [20], convolutional neural networks (CNNs) with long short-term memory networks (LSTMs) [21], and stacked transformer layers [21, 22], to learn an end-to-end representation of speech signals and emotions. These models have been shown to achieve state-of-the-art performance on SER tasks [23, 24] but require large amounts of data for training, and the learned representations may not be easily interpretable.

2.3. Factors that have an effect on the performance of SER

This section explores the factors that influence Speech Emotion Recognition (SER). Several studies have been conducted to find features that can accurately recognize emotions from speech, regardless of the speaker's background or language.
One factor studied is the effect of age and gender on SER performance. Ftoon et al. [25] built hierarchical classification models and found that using separate models for each gender and age category improved the emotion recognition accuracy compared to using one classifier to classify all the data. TING-WEI SUN [26] proposed a speech emotion recognition algorithm that combines gender information and deep learning algorithms, resulting in improved accuracy in Mandarin, English, and German.
Another aspect that has been studied concerning SER is culture and language. Different cultures express emotions in speech differently, influencing SER systems' accuracy. Fardin et al.[1] conducted a comparative analysis of speech emotion recognition in Bangla and English and discovered that SER in Bangla and English is mostly language-independent, with minor disparities observed for emotions like disgust and fear. While Raju et al. [27] adopted a k-Nearest Neighbor (kNN) classifier to recognize four discrete emotions using acoustical features for three different languages and found that English has a greater recognition rate than Malay and Mandarin.
To study influence of speakers in SER systems , Liu Z T et al. [28] proposed a method to extract a speaker-independent feature set for SER using correlation analysis and Fisher Score Selection Algorithm. The selected features were Fundamental Frequency, Formant frequency, MFCCs, and short-time Energy, achieving an average accuracy of 70% for SER across different speakers.
The approach of studying the different factors that affect Speech emotion Recognition inspired us to focus on the effects of accents on speech emotion recognition despite the same language.

3. Proposed Approach

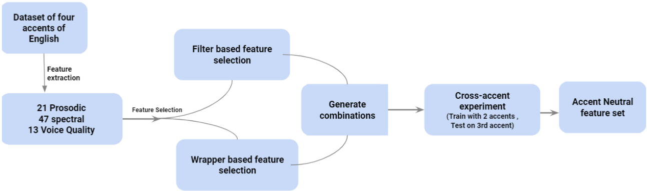
Dataset of four accents of English
Feature extraction
21 Prosodic
47 spectral
13 Voice Quality
Feature Selection
Filter based feature selection
Wrapper based feature selection
Generate combinations
Cross-accent experiment
(Train with 2 accents , Test on 3rd accent)
Accent Neutral feature set

Figure 1: *Proposed approach*

3.1. Dataset

Our study utilized four datasets of four distinct accents of English - TESS for Canadian accent [29], BESS[1] for Bangla accent, RAVDESS for American accent [30], and SAVEE for British accent [31].

3.1.1. Dataset Preprocessing

The dataset we selected for our experiments included a range of emotions - SAVEE has 7 emotions, RAVDESS has 8, TESS has 7, and BESS has 6. To ensure consistency across the datasets, only the six common emotions **angry, disgust, fear, happy, neutral, and sad** are selected after pre-processing the data by discarding surprise and calm. And in order to prevent any form of bias during the training, we took an equal amount of samples from each dataset for each emotion.

3.2. Feature Extraction

In general, four basic types of features are extracted from a speech signal to detect emotions from a speech. which are - Prosodic features, Spectral features, Voice Quality features, and TEO-based features [32].
For our experiment, we selected the prosodic features, spectral features, and voice quality features, which are the most used categories of features, as mentioned in section 2.1. Following is a count of the features that we extracted from each speech signal in our dataset.

- **21 Prosodic features**: Prosodic features provide structural information of speech, a combination of rhythm, intonation and expression, such as Intensity, pitch, and so on.
- **47 Spectral features**: Spectral features are extracted from the frequency domain of speech signals and include properties of the vocal tract [33] such as MFCC, spectral roll-off, and formant.
- **13 Voice Quality Features**: Voice quality features refer to the sound quality of someone's voice such as jitter, shimmer, and the harmonics-to-noise ratio (HNR) that can be used to distinguish emotions [33].

A total of 81 features were extracted from each speech signal in the dataset using tools such as Librosa Library [34], Praat-Parselmouth Library, and PRAAT Software [35]. Librosa is mainly used to extract the Spectral Feature, while Praat-Parselmouth Library from python and the PRAAT Software is used to extract the prosodic and voice quality features.

[1] Saad, F. (2021). A case study on the independence of speech emotion recognition in Bangla and English languages using language-independent prosodic features. arXiv preprint arXiv:2111.10776v3.

3.3. Feature Selection and Analysis

To identify a subset of accent-neutral features for emotion classification, we employed two methods of feature selection-

- Filter-based Feature Selection
- Wrapper-based Feature Selection

3.3.1. *Filter-based Feature Selection*

Filter-based feature selection involves using algorithms to filter out irrelevant or redundant features from the dataset. Filter-based feature selection such as *Fisher Score and ANOVA algorithms* was applied to 81 extracted features from all datasets to identify the top 20 significant features and every possible combination of feature subsets was generated using the top 20 features. Each subset was trained on a Support Vector Machine (SVM) to evaluate its performance for the cross-accent speech emotion recognition task. The feature subset with the highest accuracy was noted for cross-accent experiments.

3.3.2. *Wrapper-based Feature Selection*

The exhaustive search method is a specific implementation of wrapper-based feature selection which involves selecting and eliminating features to generate all possible combinations of feature sets. Our study used the exhaustive search method to find the best-performing feature subset from each of the three fundamental feature divisions - prosodic, spectral, and voice quality features. This involved taking all possible feature subsets from 21 prosodic features and performing cross-accent experiments on an SVM with each subset to find the optimal prosodic feature set that was the most accent-neutral. The exact process is repeated for the 47 spectral and 13 voice quality features to identify the best sets of features for each division. Finally, we combined the three best sets of features to obtain an accent-neutral subset of features for our emotion recognition task.

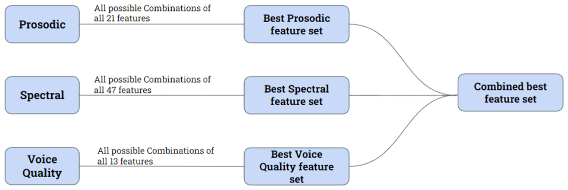
Prosodic
All possible Combinations of all 21 features
Best Prosodic feature set
Spectral
All possible Combinations of all 47 features
Best Spectral feature set
Voice Quality
All possible Combinations of all 13 features
Best Voice Quality feature set
Combined best feature set

Figure 2: *Exhaustive Feature selection approach*

3.4. Classifier

In our study, we utilized the Support Vector Machine (SVM) algorithm for the classification of emotions in speech. SVM is a popular choice for Speech Emotion Recognition tasks, as mentioned in Section 2.2. To ensure that our SVM model was optimized for accuracy, we adjusted the hyperparameters of the algorithm. The choice of hyperparameters can significantly affect the model's performance; therefore, it is crucial to tune them for optimal results. We used a grid search approach to explore the hyperparameter space and identify the optimal values.
After experimenting with various hyperparameters, we found that the *rbf* kernel was the best choice for their SVM model, which is capable of fitting non-linear decision boundaries. We also found that a C value of 10 and a gamma value of 0.01 produced the best results for our classification task. The C value controls the complexity of the decision boundary, while the gamma value determines the shape of the decision boundary and affects the model's sensitivity to input data variations. By optimizing our SVM model's hyperparameters, we obtained the best possible classification accuracy for our Cross-Accent Speech Emotion Recognition task.

4. Experiments and Results Analysis

In our study, we performed two main experiments on our datasets, including four accents. Due to the limited number of samples for each gender in each accent, we conducted separate experiments for men and women. This was done to avoid any potential gender bias in our results as we indicated in section 2.3 that researchers showed gender dependency in SER systems [26]. The process of selecting male and female samples from each accent dataset is illustrated in Figure 3.

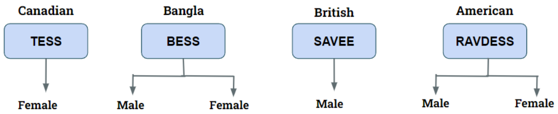

Figure 3: *Gender division of datasets*

The designed experiments are as follows -

1. Same accent experiments.
2. Cross-accent experiments using the one-vs-all approach, which were conducted in two ways:
   (a) Without feature selection, using all 81 extracted features
   (b) With feature selection, using different subsets of selected features.

4.1. Same accent experiments

In the same accent experiment, the accuracy is good when the same accent is used for both training and testing. However, in the Ravdess dataset, which uses American English, some speakers spoke quickly out of fluency, making it difficult for SVM to categorize their emotions accurately. Overall, the results were satisfactory in the same accent experiments. The performance for the same accent experiments can be found in Table 1.

Table 1: *Result from same accent experiments*

Female Samples	
Train -Test with same accent	**Accuracy**
TESS (Canadian)	99.54%
BESS (Bangla)	95.65%
Ravdess (American)	52.83%
Male Samples	
Train -Test with same accent	**Accuracy**
SAVEE (British)	65.87%
BESS (Bangla)	78.95%
Ravdess (English)	50.31%

4.2. Cross accent experiments (One Vs All)

The one-vs-all cross accent experiments, in which we used one accent for testing and others for training, are constructed in six

ways which are illustrated in Table: 2. Here the experiments were conducted on both genders separately. Specifically, three distinct accents were utilized for the test set in each gender category. As a result, there were three experimental setups for female speakers and three for male speakers.
The first experiment was conducted with all 81 extracted features without any feature selection. Later, the same experiments were re-run using a selected set of features obtained from feature selection methods mentioned in section 3.3. A significant change in accuracy was observed after using the selected features.

4.2.1. *Without feature selection*

When the 6 experiments are conducted without any kind of feature selection, accuracy drops by a large scale than that of the same accent experiments indicating that accent variation matters in Speech Emotion Recognition systems. The results for cross-accent experiments without any kind of feature selection are given in Table 2.

Table 2: *Results on cross accent experiments without feature selection*

Female Samples		
Train	**Test**	**Accuracy**
Canadian + American	Bangla	32.45%
Bangla + American	Canadian	16.67%
Bangla + Canadian	American	18.18%

Male Samples		
Train	**Test**	**Accuracy**
British + American	Bangla	26.2%
Bangla + American	British	14.29%
Bangla + British	American	9.09%

4.2.2. *With feature selection, using the selected features*

We conducted the same six experiments on our selected feature sets, chosen using the two approaches mentioned in section 3.3. Based on the filter-based feature selection outlined in section 3.3.1, we obtained the following feature set - **median intensity, q3-pitch, q1-pitch, mean-pitch, mfcc 1, f1-median, mff, min-pitch, fitch-vtl, f1-mean, mfcc 2, mfcc 4, pitch-slope, spectral centroids, average-formant, max-intensity, spectral rolloff, mfcc 3, mfcc 0, q3-intensity**. The exhaustive feature search approach, outlined in section 3.3.2 to find the best-performing feature subset in cross-accent experiments, produced the following feature subset - **q3 pitch, median intensity, q1 intensity, mfcc 7, mfcc 2, mfcc 6, min pitch, stddev pitch, pitch-slope, max intensity, relative min intensity time, relative max intensity time, q3 intensity, min intensity, mfcc 9, mfcc 12, fitch vtl, q1 pitch, stddev hnr, stddev intensity, mfcc 4, mfcc 8**.
After combining the features generated from both approaches, we arrived at a set of 30 features that gave us the best accuracy for our cross-accent experiments. Using this subset of features, we observed significant improvements in accuracy. For example, the accuracy increased from 32.45% to 62.91% when training with Canadian + American Accent and testing with Bangla accent. Similarly, after training SVM with selected features from Bangla + American accent and testing with Canadian accent, accuracy increased from 16.67% to 64.44%, and so on. The results for cross-accent experiments on the generated subset of features are presented in Table 3.

Table 3: *Results on cross accent experiments after feature selection*

Female Samples		
Train	**Test**	**Accuracy**
Canadian + American	Bangla	62.91%
Bangla + American	Canadian	64.44%
Bangla + Canadian	American	35.04%

Male Samples		
Train	**Test**	**Accuracy**
British + American	Bangla	60.43%
Bangla + American	British	44.52%
Bangla + British	American	32.01%

Our goal was to find the accent-neutral feature set. These improvements in performance indicate that the selected feature sets are more accent-neutral and can identify emotions more accurately despite variations in accents.
The analysis of the best feature sets has revealed a specific pattern in the types of features that are most effective in capturing accent independence for Speech Emotion Recognition. The pattern that we found from our results is that all the features that were found to be most prominent belong to the group (local/global version) of **Pitch**, **Intensity**, and **MFCC**. Interestingly, we observed that this pattern of features is consistent across both male and female speech samples. This consistency suggests that these features are not gender-specific and can be used effectively in both male and female speech for Speech Emotion Recognition.

5. Conclusion

Our study aimed to develop a Speech Emotion Recognition (SER) system that performs well despite variations in accents within the English language. Our analysis focused on four accents in the English language: Canadian, British, Bangla, and American. We conducted two main experiments to evaluate the performance of an SVM classifier on same-accent experiments and cross-accent experiments using two feature selection approaches. The results showed that the SER system performed well for the same accent experiments, but the performance dropped significantly with the cross-accent emotion classification due to variations in accents. However, by selecting specific feature sets, we were able to achieve better performance despite accent variations. We observed that the selected feature sets consistently included Pitch, Intensity, and MFCC features. Using our selected features, we improved the performance of our cross-accent SER system by 20-30%.
In our future research, we plan to broaden the scope by including more accents and exploring deep learning-based approaches in future research.

6. References

[1] M. Selvaraj, R. Bhuvana, and S. Padmaja, "Human speech emotion recognition," *International Journal of Engineering & Technology*, vol. 8, pp. 311–323, 2016.

[2] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021.

[3] J. M. Zarate, X. Tian, K. Woods, and D. Poeppel, "Multiple levels of linguistic and paralinguistic features contribute to voice recognition," *Scientific reports*, vol. 5, p. 11475, 06 2015.

[4] J. B. Alonso, J. Cabrera, M. Medina, and C. M. Travieso, "New approach in quantification of emotional intensity from the speech signal: emotional temperature," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9554–9564, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417415005229

[5] B. Ayoub, K. Jamal, and Z. Arsalane, "An analysis and comparative evaluation of mfcc variants for speaker identification over voip networks," in *2015 World Congress on Information Technology and Computer Applications (WCITCA)*, 2015, pp. 1–6.

[6] M. Kächele, D. Zharkov, S. Meudt, and F. Schwenker, "Prosodic, spectral and voice quality feature selection using a long-term stopping criterion for audio-based emotion recognition," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 803–808.

[7] T. Özseven, "A novel feature selection method for speech emotion recognition," *Applied Acoustics*, vol. 146, pp. 320–326, 2019.

[8] S. Zhang, "Emotion recognition in chinese natural speech by combining prosody and voice quality features," in *Advances in Neural Networks-ISNN 2008: 5th International Symposium on Neural Networks, ISNN 2008, Beijing, China, September 24-28, 2008, Proceedings, Part II 5*. Springer, 2008, pp. 457–464.

[9] S. Kuchibhotla, H. D. Vankayalapati, R. Vaddi, and K. R. Anne, "A comparative analysis of classifiers in emotion recognition through acoustic features," *International Journal of Speech Technology*, vol. 17, pp. 401–408, 2014.

[10] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech Language*, vol. 29, 02 2014.

[11] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2010.

[12] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, "Automatic emotion recognition using prosodic parameters," in *Ninth European conference on speech communication and technology*. Citeseer, 2005.

[13] L. Sun, S. Fu, and F. Wang, "Decision tree svm model with fisher feature selection for speech emotion recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, pp. 1–14, 2019.

[14] T. Nwe, S. Foo, and L. De Silva, "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, pp. 603–623, 11 2003.

[15] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 2, 2003, pp. II–1.

[16] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *Eighth European conference on speech communication and technology*, 2003.

[17] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification," in *2005 IEEE international conference on multimedia and expo*. IEEE, 2005, pp. 864–867.

[18] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[19] N. Scheidwasser-Clow, M. Kegler, P. Beckmann, and M. Cernak, "Serab: A multi-lingual benchmark for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7697–7701.

[20] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech 2015*, 2015.

[21] X. Wang, M. Wang, W. Qi, W. Su, X. Wang, and H. Zhou, "A novel end-to-end speech emotion recognition network with stacked transformer layers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6289–6293.

[22] Y. Wang, G. Shen, Y. Xu, J. Li, and Z. Zhao, "Learning mutual correlation in multimodal transformer for speech emotion recognition." in *Interspeech*, 2021, pp. 4518–4522.

[23] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[24] M. R. Ahmed, S. Islam, A. M. Islam, and S. Shatabda, "An ensemble 1d-cnn-lstm-gru model with data augmentation for speech emotion recognition," *Expert Systems with Applications*, vol. 218, p. 119633, 2023.

[25] F. A. Shaqra, R. Duwairi, and M. Al-Ayyoub, "Recognizing emotion from speech based on age and gender using hierarchical models," *Procedia Computer Science*, vol. 151, pp. 37–44, 2019.

[26] T.-W. Sun, "End-to-end speech emotion recognition with gender information," *IEEE Access*, vol. 8, pp. 152 423–152 438, 2020.

[27] R. Rajoo and C. C. Aun, "Influences of languages in speech emotion recognition: A comparative study using malay, english and mandarin languages," in *2016 IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE)*, 2016, pp. 35–39.

[28] Z. Liu, K. Li, D.-Y. Li, L. Chen, and G. Tan, "Emotional feature selection of speaker-independent speech based on correlation analysis and fisher," *2015 34th Chinese Control Conference (CCC)*, pp. 3780–3784, 2015.

[29] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020. [Online]. Available: https://doi.org/10.5683/SP2/E8H2MF

[30] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[31] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[32] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320310004619

[33] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[34] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[35] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.