Indonesian Speech Emotion Recognition using Cross-Corpus Method with the Combination of MFCC and Teager Energy Features

Oscar Utomo Kumala[1], Amalia Zahra[2]
Computer Science Department, Bina Nusantara University
Jakarta, Indonesia 11480

***Abstract*—Emotion recognition is one of the widely studied topics in speech technology. Emotions that come from speech can contain useful information for many purposes. The main aspects in speech emotion recognition are speech features, speech corpus, and machine learning algorithms as the classifier method. In this paper, cross-corpus method is used to conduct Indonesian Speech Emotion Recognition (SER) along with the combination of Mel Frequency Cepstral Coefficients (MFCC) and Teager Energy features. Using Support Vector Machine (SVM) as classifier, the experiment result shows that applying cross-corpus method by adding corpora from other languages to the training dataset improves the emotion classification accuracy by 4.16% on MFCC Statistics feature and 2.09% on Teager-MFCC Statistics feature.**

Keywords—Cross corpus; Indonesian speech emotion recognition; Mel Frequency Cepstral Coefficients; Teager Energy

I. Introduction

Nowadays we are experiencing a rapid growth on Information Technology (IT) sectors, especially in mobile devices area. The interaction between user and mobile devices is getting smoother each day so that it can assist user’s daily activities. One important means to achieve this is speech voice. In speech voice there are a lot of information that can be extracted and analyzed. One important aspect of the information is emotion which can contain many additional information, such as the speaker’s condition (physical state or mood), the meaning of the speech, and many more.

Different kind of emotion contained in a speech may cause a different response on the person or device the speaker is talking to. A virtual assistant with emotion recognition feature [1] will have advantage of obtaining capability to give different answers or responses depending on the emotion contained in the speech or order. One simple application is the virtual assistant will compile a (song) playlist that is comforting the user if there is sad emotion recognized in the speech.

Because of this high potential of use, it is necessary to further analyze the emotion recognition process itself. From the studies written in literatures, there are three main factors in Speech Emotion Recognition (SER). The first one is speech features, which consist of acoustic features, lexical features, sound volume and frequencies, vocabularies, languages, speaker’s background (nationality, ethnic, age, etc.), and many more. The next factor is the availability of corpus which will be used as training and testing set. The last factor is the methods of machine learning algorithm used to classify the emotion in the speech.

Studies have been carried out at international level for different languages, features [2], corpora [3], methods [4], and algorithms [5] showing various results. Only in recent years, studies for Indonesian have been rising. In this study, we will focus on the Indonesian SER. We need to distinguish SER in Indonesian with other languages because each language and culture has its own characteristic in the SER process [6].

The first main topic in this study is the use of cross-corpus method [7] for the Indonesian SER. We decide to use cross-corpus because of the limitation of the available Indonesian corpus. A SER will achieve better result with larger training dataset. That is the reason we include the corpus from other languages to the training dataset. The corpus used in this study is Berlin Database of Emotional Speech (Berlin EmoDB), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and Surrey Audio-Visual Expressed Emotion (SAVEE). Thus, there are three corpora: one German corpus and two English corpora.

Another main topic is the combination of two speech features, Mel Frequency Cepstral Coefficients (MFCC) features and Teager Energy features. MFCC features are one of the most used speech features in the SER studies [8]. The features will be combined with Teager Energy features [9] to hopefully achieve better result. These speech features are extracted from the corpus and used along with their statistical values. The combination of the features will be tested together with the cross-corpus method and Support Vector Machine (SVM) classifier. The results of the test will be analyzed thus a number of conclusions can be drawn.

The remainder of this paper is as follows. In Section II, we discuss the previous research and studies that are relevant to this topic. We describe the configurations of our experiments in Section III. In Section IV, we describe the results of our experiments and perform analysis on them. Section V concludes the paper.

II. Related Works

Studies on SER started at international level. One of the first studies was conducted by Shah & Hewlett [10] to detect emotion by extracting and analyzing speech features which consist of pitch, MFCC, and Formants. The study used SVM with Linear kernel and k-Mean as classifiers. It showed that the emotion recognition rate is higher for male speakers compared to female ones, and there are some similar emotions: happy, elation, and interest; agitated and subdued.

A framework was created by Pfister [11] to detect emotion by analyzing speech in real time, extracting speech features using OpenSMILE algorithm: energy, volume, voice quality, mel-spectra, MFCC, and some calculations such as mean, extreme, peak, percentile, and deviation. Using SVM with Radial Basis (RBF) kernel, it achieved 70% to 89% accuracy rate depending on the selected method with low delay (0.046 to 0.110 second for 1- to 5-second sentence). The notable finding from the framework is that emotion recognition can be performed in real time, which is vital to human-computer interaction applications.

Studies on Indonesian SER have started in recent years [12]. There is a study by Lubis, Lestari, Purwarianti, Sakti, & Nakamura [13] which succeeded in forming the first Indonesian emotional speech corpus, namely Indonesian Emotional Speech Corpus (IDESC). From the study, it can be concluded that the recognition of angry emotion has a relatively higher accuracy than satisfied emotion. In general, active emotions are easier to recognize than passive emotions. The study achieved 68.31% accuracy for the classification of four emotion classes. Another study presents a speaker-independent emotion recognition [14]. The study also found that disgust is the most difficult emotion to detect, followed by sad emotion.

One of the latest Indonesian SER studies was conducted by Gunawan & Idananta [15] by testing sound signals in Indonesian. Existing sound signals are analyzed using the MFCC features combined with the Teager Energy feature. The emotions were classified using SVM into four classes, namely angry, fear, happy, and sad. The speech corpus was created by recording conversations from four amateur actors and actresses. They speak 15 Indonesian sentences for four times, each based on the emotion requested. This speech corpus will also be used in this study as Indonesian corpus.

The results in [15] show that Teager Energy is an important feature that contributes to the accuracy of emotional classification by approximately 41%. The study also found that happy emotion seemed to be somewhat difficult to distinguish from angry emotion. In addition, the emotions of angry, fear, and sad can be recognized from speech signals with high accuracy.

III. Experiments

We begin the experiments by preparing the training and testing dataset which consist of the corpora aforementioned. After the training and testing dataset are ready, the speech feature extraction process starts, followed by training and testing process, which is followed by evaluation.

A. *Preparing Corpus*

There are four corpora used as training and testing dataset. The first corpus is Berlin EmoDB [16], which is a database of German emotional speech containing 535 audio files with 7 emotion classes. The second one is RAVDESS [17], which is a validated multimodal database of English emotional speech and song with North American accent, containing 1440 audio files with 8 emotion classes. The next corpus, SAVEE [18], is an English audiovisual database with British accent which consists of 480 audio files with 7 emotion classes. The Indonesian corpus is the last corpus used with 4 emotion classes, resulting in 60 audio files.

All the corpora will go through emotion filter process. The experiment will only use 4 emotion classes: angry, fear, happy, and sad. All emotion classes outside those 4 will not be included in the training and testing dataset. After this process, there is 61% data left on Berlin EmoDB, 54% data left on RAVDESS, 50% data left on SAVEE, and 100% data on Indonesian corpus.

The next process is data standardization process, which balances the data amount of all emotion classes. This process is necessary because the data amount of each emotion class is not the same for each corpus. Besides standardizing the data amount, the audio bitrate of all audio files is also standardized into 256 kbps with the assistance of Audacity desktop application.

Through emotion filter process and data standardization processes, the data amount of each class emotion becomes 61 on Berlin EmoDB, 174 on RAVDESS, 60 on SAVEE (same as original), and 60 on Indonesia corpus (same as original). There are two data excluded from RAVDESS because we are unable to extract Teager Energy feature from the audio files. The final amount of data that can be used for training and testing dataset is 1418.

After the previous processes, all corpora will be combined and then divided into three corpus groups: 100% corpus group, 80% corpus group, and 20% corpus group. 100% corpus group is the fully combined corpus. 80% corpus group consists of 80% data of combined corpus, where the data are picked manually, and used as part of training dataset. Likewise, 20% corpus group consists of 20% data of combined corpus, where the data are picked manually, and used as part of testing dataset. By performing such a grouping, the ratio of 80% and 20% for each corpus can be achieved.

B. *Extracting the Speech Features*

There are two kinds of speech features that will be used for training and testing process: MFCC features and Teager Energy features. We begin the speech feature extraction process by reading the corresponding audio file with wav file extension. Reading the audio file will give us the signal and rate values which will be used to calculate the MFCC and Teager Energy values.

The MFCC feature extraction process is carried out using python_speech_features library [19]. A simple function in the library takes the signal and rate values as parameters and return the MFCC feature values in the form of a 2-dimensional array. The size of the array is 13 times the number of sound frames. The number 13 is obtained from the number of frequency bands in a speech voice, and the number of sound frames produced depending on the duration of the corresponding audio file. In this study, the number of sound frames taken is 75, which is the smallest number of sound frames from all corpus data. Thus, we will get 975 MFCC feature values saved into a database in the form of a 1-dimensional array.

In addition to the MFCC feature values, statistical values will also be calculated for each frequency band. These statistical values include mean (average array value), min (smallest array value), max (largest array value), std (array standard deviation), and median (array middle value). Thus, 5 statistical values will be obtained for each of the 13 frequency bands so that a total of 65 MFCC statistical values will be saved into the database in the form of a 1-dimensional array.

Similar to the MFCC feature extraction process aforementioned, the Teager Energy feature extraction process is also carried out using a library. The library has a function to return the values of Teager-Kaiser Operator (also known as the Nonlinear Energy Operator) and Envelope Derivative Operator (EDO). The Teager-Kaiser Operator values will be used in this study as the Teager Energy feature values. The values are returned in 1-dimensional array with the size depending on the duration of the corresponding speech file. In this study, the first 1000 arrays will be retrieved and divided into two features (500 arrays each), namely Teager 1 feature and Teager 2 feature. Both features are saved to the database in the form of a 1-dimensional array.

In addition to the Teager Energy feature values, the statistical values are also calculated. Similar to the MFCC statistical values, Teager Energy statistical values also include mean (average array value), min (smallest array value), max (largest array value), std (array standard deviation), and median (array middle value). Thus, we will obtain five Teager Energy statistical values saved to the database in the form of a 1-dimensional array.

C. Configuring Corpus for Training and Testing

Training and testing processes are interconnected. A training process produces a model used for one or more testing processes. Both processes are conducted to as many corpus combinations as possible so we can obtain many testing results for this study.

From the initial four corpora, there will be two additional corpora formed from the combination of those corpora. The first one is International corpus which consists of three corpora in non-Indonesian language: Berlin EmoDB, RAVDESS, and SAVEE. Another one is a combined corpus which consists of all initial four corpora. Thus, there are six corpora in total for training and testing processes.

With such corpus combinations, we need a proper configuration for the use of corpus grouping in certain training and testing scenarios. Table I shows the possible scenarios and the configuration.

First scenario is when the training and testing processes use the same corpus. Here we will use 80% corpus group for training and 20% corpus group for testing. The next scenario is the opposite of first scenario, when the training and testing processes use the different corpora. We will use 100% corpus group for both processes.

The third scenario is when the training process uses a combined corpus which consists of a single corpus that is also used for testing process. In this case, we will use 80% corpus group at the training process and 20% corpus group at the testing process. The other corpus of combined corpus at the training process will use 100% corpus group.

The final scenario is the opposite of the third scenario, when training process uses a single corpus which is also used as part of a combined corpus at testing process. In this case, for that single corpus we will use 80% corpus group at the training process and 20% corpus group at the testing process. The other corpus of combined corpus at the testing process will use 100% corpus group.

D. Conducting Training and Testing

In this study, we conduct the training using SVM with RBF kernel. First step of the training process is retrieving speech features values. There are five speech features: MFCC feature, MFCC Statistics feature, Teager Energy 1 feature, Teager Energy 2 feature, and Teager Energy Statistics feature.

At this point, we add a new speech feature which is formed from the combination of Teager Energy Statistics feature with MFCC Statistics feature, namely Teager-MFCC Statistics feature. Thus, there are six speech features values which can be used for next processes.

All speech feature values obtained need to go through normalization process. This normalization process is called scaling in Python, where all values will be normalized to the range of -1 to 1. Normalization will be carried out on the training speech feature values so that the minimum value is -1 and the maximum value is 1. The normalization will produce a scale which will be applied to the testing speech features values.

The next process is building model. This process is carried out by importing 'RandomizedSearchCV' from 'sklearn.model_selection' in Python. This module aims to form the best model by looking for random combinations of parameter values from several predetermined parameter values. The mentioned parameters are C and Gamma values. In this module, 10-fold cross validation is applied to the training dataset. The 'RandomizedSearchCV' process is repeated 100, 250, and 500 times. The next process is testing. Each model built from the previous step will be tested using the testing dataset to classify emotions.

TABLE I. CONFIGURATION FOR THE USE OF CORPUS GROUPING IN CERTAIN TRAINING AND TESTING SCENARIOS. # IS NUMBER OF SCENARIO, TRC IS TRAINING CORPUS, TSC IS TESTING CORPUS, TRG IS TRAINING CORPUS GROUPING, TSG IS TESTING CORPUS GROUP

#	TRC	TSC	TRG	TSG
1	A	A	80% A	20% A
2	A	B	100% A	100% B
3	A B C	A	80% A 100% B 100% C	20% A
4	A	A B C	80% A	20% A 100% B 100% C

IV. Results and Discussion

We divide the testing results into three parts. The first part is the testing results for the same corpus. The next part is the testing results for different corpus. The final part is the analysis of all testing results.

A. Testing Result for the Same Corpus

Table II shows the accuracies of testing using the same corpus. All six available corpora will go through testing process with six speech features values. The results shown in the table are those that achieve the best average accuracy among three numbers of iteration (i.e. 100, 250, and 500). The configuration of corpus grouping is applied here.

B. Testing Result for different Corpus

In contrast with testing using the same corpus in Table II, The result of testing using different corpus is shown in Table III. All six available corpora will go through testing process with several corpus combinations and five speech features values. The Teager Energy 2 feature is excluded here because the result is very similar to that using the Teager Energy 1 feature. The results shown in the table are those that achieve the best average accuracy between three numbers of iteration (i.e. 100, 250, and 500). The configuration of corpus grouping is also applied here.

TABLE II. Testing result for same Corpus. C is Corpus Name, F1 is MFCC Feature, F2 is MFCC Statistics Feature, F3 is Teager 1 Feature, F4 is Teager 2 Feature, F5 is Teager Statistics Feature, F6 is Teager-MFCC Statistics Feature, Avg is Average Accuracy, C1 is Berlin EmoDB, C2 is RAVDESS, C3 is SAVEE, C4 is Indonesian, C5 is International, C6 is Combined Corpus

C	F1	F2	F3	F4	F5	F6	Avg
C1	68.75%	87.50%	47.92%	41.67%	47.92%	85.42%	**63.20%**
C2	41.18%	83.82%	25.74%	20.59%	45.59%	79.41%	**49.39%**
C3	54.17%	58.33%	39.58%	39.58%	31.25%	58.33%	**46.87%**
C4	54.17%	79.17%	43.75%	47.92%	66.67%	83.33%	**62.50%**
C5	45.69%	80.17%	29.31%	30.17%	39.66%	81.90%	**51.15%**
C6	46.43%	82.50%	25.71%	26.43%	40.00%	80.36%	**50.24%**
Avg	**51.73%**	**78.58%**	**35.34%**	**34.39%**	**45.18%**	**78.13%**	**53.89%**

TABLE III. Testing Result for different Corpus. TR is Training Dataset, TS is Testing Dataset, F1 is MFCC Feature, F2 is MFCC Statistics Feature, F3 is Teager 1 Feature, F5 is Teager Statistics Feature, F6 is Teager-MFCC Statistics Feature, Avg is Average Accuracy, C1 is Berlin EmoDB, C2 is RAVDESS, C3 is SAVEE, C4 is Indonesian, C5 is International, C6 is Combined Corpus

TR	TS	F1	F2	F3	F5	F6	Avg
C1	C2	25.50%	28.53%	25.07%	31.84%	26.51%	**27.06%**
C1	C3	35.83%	36.25%	29.17%	26.67%	22.08%	**31.17%**
C1	C4	42.50%	49.17%	17.08%	23.33%	37.92%	**29.83%**
C1	C5	29.84%	34.11%	26.88%	36.25%	33.20%	**30.53%**
C1	C6	30.36%	29.54%	24.55%	35.02%	28.81%	**28.69%**
C2	C1	25.00%	50.82%	25.41%	24.59%	47.13%	**30.41%**
C2	C3	23.33%	26.25%	25.00%	24.17%	27.92%	**24.75%**
C2	C4	23.33%	29.58%	25.00%	27.08%	32.92%	**26.00%**
C2	C5	27.10%	46.94%	24.52%	25.65%	44.52%	**29.97%**
C2	C6	26.51%	42.33%	25.12%	29.65%	41.63%	**29.84%**
C3	C1	25.00%	39.75%	42.21%	16.80%	32.79%	**32.87%**
C3	C2	22.19%	25.36%	24.93%	29.11%	25.36%	**25.30%**
C3	C4	25.83%	26.25%	14.17%	20.42%	26.25%	**20.33%**
C3	C5	27.28%	30.93%	30.43%	26.17%	32.45%	**28.88%**
C3	C6	26.35%	35.73%	27.49%	25.69%	33.93%	**28.39%**
C5	C1	66.67%	89.58%	45.83%	54.17%	89.58%	**59.17%**
C5	C2	34.56%	84.56%	24.26%	43.38%	84.56%	**42.20%**
C5	C3	43.75%	75.00%	37.50%	27.08%	79.17%	**43.33%**
C5	C4	34.17%	48.33%	16.25%	25.83%	44.58%	**28.25%**
C6	C4	58.33%	83.33%	35.42%	43.75%	85.42%	**46.25%**
C6	C5	44.40%	79.74%	26.72%	37.07%	74.14%	**43.10%**
Avg		**33.36%**	**47.24%**	**27.48%**	**29.93%**	**45.28%**	**34.81%**

C. Analysis

Analysis is conducted to the results obtained from both testing using the same corpus and that using different corpus. From the result of testing using the same corpus, we can highlight several points. The first point is related to speech features that achieve the highest accuracy for each corpus. MFCC Statistics feature achieves the highest accuracy on Berlin EmoDB (87.50%), RAVDESS (83.82%), SAVEE (58.33%, tied with Teager-MFCC Statistics feature), and Combined corpus (82.50%). Meanwhile, Teager-MFCC Statistics feature achieves the highest result on SAVEE, Indonesian corpus (83.33%), and International corpus (81.90%).

Next point is that Berlin EmoDB has the highest average accuracy among all corpora (63.20%), while SAVEE has the lowest one (46.87%). The last point to highlight is that MFCC Statistics feature achieves the highest average accuracy (78.58%) followed tightly (0.45%) by Teager-MFCC Statistics feature (78.13%). On the contrary, Teager Energy 2 feature has the lowest average accuracy (34.39%) followed by Teager Energy 1 feature (35.34%).

We can see different highlights on the testing result for different corpus. Combined corpus achieves the highest average accuracy (44.68%) followed tightly (1.44%) by International corpus (43.24%). From the result, it is shown that corpus which consists of many single corpora (International corpus and Combined corpus) has higher average accuracy than single corpus (Berlin EmoDB, RAVDESS, SAVEE, and Indonesian corpus). Additionally, similar to the result of testing using the same corpus, MFCC Statistics feature achieves the highest average accuracy (47.24%) followed tightly (1.96%) by Teager-MFCC Statistics feature (45.28%). Teager Energy 2 feature once again has the lowest average accuracy (25.57%) followed by Teager Energy 1 feature (27.48%).

From both testing results, we can highlight several points. The order of the speech features that achieve from the highest to the lowest average accuracy is MFCC Statistics feature, Teager-MFCC Statistics feature, Teager Statistics feature, MFCC feature, Teager Energy 1 feature, and Teager Energy 2 feature. Even though MFCC Statistics feature achieves the highest average accuracy, in some cases Teager-MFCC Statistics produces a better result.

Another point to highlight is that statistical features (MFCC Statistics, Teager Statistics, and Teager-MFCC Statistics) achieve higher average accuracy than non-statistical features (MFCC, Teager Energy 1, and Teager Energy 2). The average accuracy of testing using the same corpus is still higher than that using different corpora.

For specific case where Indonesian corpus used as a testing dataset, we obtained a higher accuracy rate when using combined corpus as the training dataset compared to that achieved by using the same corpus. We achieved the accuracy of 83.33% and 79.17% from testing using the MFCC Statistics feature for the first and latter scenario, respectively, whereas using Teager-MFCC Statistics feature achieved the accuracy of 85.42% and 83.33% for such scenarios, respectively. It means that there is an accuracy improvement by 4.16% using MFCC Statistics feature and 2.09% using Teager-MFCC Statistics feature.

V. Conclusion

Based on the results and highlights described in the previous section, we can conclude three main points. First, we can see that applying cross-corpus method by adding corpora from other languages to the training dataset can improve the overall performance of the emotion recognition, including the Indonesian SER. From the corpora, we can also see that English and German languages have good compatibility with Indonesian in emotion classification aspect.

Next, we can see that when applying cross-corpus method in Indonesian SER, the speech features that achieve the best results are MFCC Statistics feature and Teager-MFCC Statistics feature. And finally, the accuracy improvement of 4.16% using MFCC Statistics feature and 2.09% using Teager-MFCC Statistics feature could be a good start for Indonesian SER and can be further improved in the future.

Potential improvement in future studies may include the use of more complex speech features complemented with feature selection method and other classification methods apart from SVM. A good classifier that is worth a try is Extreme Learning Machine (ELM) [20] which has proven to achieve good results in some studies.

References

[1] G. Iannizzotto, L. Lo Bello, A. Nucita, and G. M. Grasso, “A Vision and Speech Enabled, Customizable, Virtual Assistant for Smart Environments,” Proc. - 2018 11th Int. Conf. Hum. Syst. Interact. HSI 2018, no. November, pp. 50–56, 2018.

[2] T. Özseven, “A Novel Feature Selection Method for Speech Emotion Recognition,” Appl. Acoust., vol. 146, pp. 320–326, 2019.

[3] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases,” Pattern Recognit., vol. 44, no. 3, pp. 572–587, Mar. 2011.

[4] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, “Speech Emotion Recognition Using Fourier Parameters,” IEEE Trans. Affect. Comput., vol. 6, no. 1, pp. 69–75, 2015.

[5] L. Yu, B. Wu, and T. Gong, “A Hierarchical Support Vector Machine Based on Feature-driven Method for Speech Emotion Recognition,” pp. 901–908, 2013.

[6] N. Kamaruddin, A. Wahab, and C. Quek, Cultural Dependency Analysis For Understanding Speech Emotion, Expert Syst. Appl., vol. 39, no. 5, pp. 5115–5133, 2012.

[7] H. Kaya and A. A. Karpov, “Neurocomputing Efficient and effective strategies for cross-corpus acoustic emotion recognition,” Neurocomputing, vol. 275, pp. 1028–1034, 2018.

[8] C. S. Ooi, K. P. Seng, L. M. Ang, and L. W. Chew, A new approach of audio emotion recognition, Expert Syst. Appl., vol. 41, no. 13, pp. 5858–5869, 2014.

[9] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M. A. Mahjoub, and C. Cleder, Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO, Speech Commun., vol. 114, no. May, pp. 22–35, 2019.

[10] R. Shah and M. Hewlett, “Emotion detection from speech, Final Proj. cs, p. 229, 2007.

[11] T. Pfister, Emotion Detection from Speech, Gonv. Caius Coll., 2010.

[12] F. Kasyidi and D. P. Lestari, “Identification of Four Class Emotion from Indonesian Spoken Language Using Acoustic and Lexical Features,” J. Phys. Conf. Ser., vol. 971, no. 1, 2018.

[13] N. Lubis, D. Lestari, A. Purwarianti, S. Sakti, and S. Nakamura, Emotion recognition on Indonesian television talk shows, pp. 466–471, 2014.

[14] M. Kurniawati, Pipin; Lestari, Dessi Puji; Leylia Khodra, Speech emotion recognition From Indonesian spoken language using acoustic and lexical features, no. November, pp. 1–3, 2017.

[15] F. E. Gunawan and K. Idananta, Predicting the level of emotion by means of Indonesian speech signal, TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 15, no. 2, p. 665, 2018.

[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, A database of German emotional speech, 9th Eur. Conf. Speech Commun. Technol., no. January, pp. 1517–1520, 2005.

[17] 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, vol. 13, no. 5. 2018.

[18] P. Jackson and S. ul haq, Surrey Audio-Visual Expressed Emotion (SAVEE) database, 2011. [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/. [Accessed: 26-May-2020].

[19] James Lyons et al., “jameslyons/python_speech_features: release v0.6.1,” 2020.

[20] W.-H. Cao, M. Wu, J.-P. Xu, Z.-T. Liu, G.-Z. Tan, and J.-W. Mao, “Speech Emotion Recognition Based on Feature Selection and Extreme Learning Machine Decision Tree,” Neurocomputing, vol. 273, pp. 271–280, 2017.