Simple meta-optimization of the feature MFCC for public Emotional datasets classification

Enrique de la Cal[1], Alberto Gallucci[1], Jose Ramón Villar[1], Kaori Yoshida[2], and Mario Koeppen[2]

[1] University of Oviedo, Computer Science Department, Oviedo, Spain
albertogalluccis@gmail.com, {delacal,villarjose}@uniovi.es
[2] Kyushu Institute of Technology, Departament of Human Intelligence Systems, Fukuoka, Japan
kaori@brain.kyutech.ac.jp, mkoeppen@ieee.org

Abstract. A Speech Emotion Recognition (SER) system can be defined as a collection of methodologies that process and classify speech signals to detect emotions embedded in them [2]. Among the most critical issues to consider in an SER system are: i) definition of the kind of emotions to classify, ii) look for suitable datasets, iii) selection of the proper input features, and iv) optimization of the suitable features. This work, will consider four of the well-known datatset in the literature: EmoDB, TESS, SAVEE and RAVDSS. Thus, this study focuses on designing a low-power SER algorithm based on combining one prosodic feature with six spectral features to capture the rhythm and the frequency, respectively, comparing eleven low-power Classical classification Machine Learning techniques (CML). The main goal will be to optimise the two main parameters of the MFCC spectral feature through the meta-heuristic technique SA: the n_mfcc and the hop_length. The resulting algorithm could be deployed on low-cost embedded systems with limited computational power like a smart speaker, and the proposed SER algorithm will be validated for four selected datasets.
The obtained models for the eleven CML techniques with the optimised MFCC features, outperforms clearly (more than a 10%) the baseline models obtained with the not-optimized MFCC for the studied datasets.

Keywords: Speech Emotion Recognition · SER · Segmental features · MFCC · Emotional Speech DataSets · SER classification algorithms · Optimization

1 Introduction and motivation

A Speech Emotion Recognition (SER) system can be defined as a collection of methodologies that process and classify speech signals to detect emotions embedded in them [2].

Some of the current most brilliant applications of SER are [7]: i) robotics: to design intelligent collaborative or service robots which can interact with humans,

ii) marketing: to create specialised adverts, based on the emotional state of the potential customer, iii) in education: used for improving learning processes, knowledge transfer, and perception methodologies, iv)entertainment industries: to propose the most appropriate entertainment for the target audience and v) health, to gather real-time emotional information from the patients in order to make decisions to improve their lives.

Between the most important issues to consider in an SER system are: Definition of the kind of emotions to classify Look for the suitable datasets Selection of the suitable input features Definition of the strategy of the proposal

1.1 Affective state taxonomy

According to the latter classification, the affective states can be clustered in four main concepts [7]:

- "emotion" is a response of the organism to a particular stimulus (person, situation or event). Usually, it is an intense, short-duration experience, and the person is typically well aware of it;
- "affect" is a result of the effect caused by emotion and includes their dynamic interaction;
- "feeling" is always experienced with a particular object of which the person is aware; its duration depends on the length of time that the representation of the object remains active in the person's mind;
- "mood" tends to be subtler, longer-lasting, less intensive, and more in the background, but it can affect a person's affective state in a positive or negative direction.

Usually, the applications of SER systems are focused on "emotions" recognition since his concept provides instant and a piece of valuable information for the decision module. Besides, it is easier to obtain affective state "emotion" from the voice speech than the remaining affective states, so most SER datasets available in the literature label the data with "emotion" classes. Hence, current work will be focused on **emotions** identification.

1.2 The election of the Emotion Speech dataset

Since the classification process is dependent on labelled data, databases are an integral component of SER. Furthermore, the success of the recognition process is influenced by the size and quality of the data. Data that is incomplete, low-quality, or faulty may result in incorrect predictions; thus, data should be carefully designed and collected[2]. So, there are three types of Speech Emotion Datasets (SED) for Speech Emotion Recognition: a) Simulated SEDs, b) Elicited (Induced) SEDs, and c) Natural SEDs. Thus, SED Utterances in simulated SEDs are played and recorded in soundproof studios by professional or semi-professional actors; elicited SEDs are produced in a simulated emotional environment that can induce various emotions. In this case, emotions are not

entirely evoked, similar to actual emotions; Natural language datasets are compiled chiefly from talk shows, call centre interviews, radio conversations, and related media. This kind of data is more challenging to get because collecting and distributing private data entails ethical and legal challenges.

1.3 The Emotion Speech features

Speech is a variable-length signal that carries both information and emotion, so global or local features can be extracted depending on the required goal. Global features represent the gross statistics such as mean, minimum and maximum values, and standard deviation. Local features represent the temporal dynamics, where the purpose is to approximate a stationary state. These stationary states are essential because emotional features are not uniformly distributed over all positions of the speech signal [**?**]. For example, emotions such as anger are predominant at the beginning of utterances, whereas surprise is overwhelmingly conveyed at the end of it. Hence, to capture the temporal information from the speech, local features are used. These local and global features of SER systems can be categorised mainly under four groups[2]: prosodic, spectral, voice quality, and features based on Teager energy operator. Nevertheless, classifiers can be improved by incorporating additional features from other modalities, such as visual or linguistic depending on the application and availability. Commonly, prosodic and spectral are the typical features used in SER, although in practice are combined to obtain better performance[2]. Prosodic features are those that can be perceived by humans, such as intonation and rhythm[**?**], while spectral features capture the vocal tract shape of the person[**?**]. Characteristics of the vocal tract are well represented in the frequency domain, so usually, spectral features are obtained by transforming the time domain signal into the frequency domain signal using the Fourier transform. One of the more useful spectral feature used in SER is MFCC (Mel Frequency Cepstral Coefficient).

1.4 The goal

Thus, this study focuses on designing a low-power SER algorithm based on a combination of prosodic and spectral features, using low-power Classical classification Machine Learning techniques (CML). The main goal will be to optimise two main parameters of the MFCC spectral feature using different evolutionary meta-heuristic techniques. The resulting algorithm should be deployed on low-cost embedded systems with limited computational power like a smart speakers or smartwatches. Finally, the proposed SER algorithm will be validated with four of the most well-known SEDs. The structure of the paper is as follows. The following section deals with the description of the proposal, together with the combination and description of the input features. Section 3 will include the details of the selected SED, the experimental setup and discussion of the results. Finally, conclusions and future work are depicted.

2 The proposal

Typically, the stages that comprise an SER system are: Pre-processing, Feature computation, Feature selection and Classification[20, 2]. However, it is not common in this field to optimise each feature individually after or before the feature selection stage. So this paper proposes including a Feature Optimisation stage (in Green) where the key features will be optimised (see Figure 1).

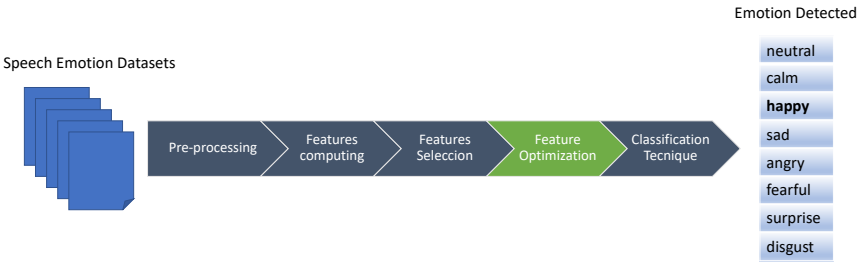

Fig. 1: Overall steps of the proposed SER algorithm

Concerning this new stage, there are two essential issues: including this step after or before calculating the features and which features are suitable to be optimised. Concerning the former issue, the optimisation stage is included after the feature selection to not compute the optimisation on non-significative features, and the features that will be optimised will be the ones with more sensible parameters.

Algorithm 1 details the stages defined in figure 1 for one generic speech dataset D composed of NF records:

1. L1-7: The sampling rate of all the files in dataset D, have been set to 16kHz, and the stereo channels have been unified in mono. As the range of values of all the sound wave files is typical, it was not updated.
2. L9-14: As this is a preliminary study of an optimal SER algorithm, segmental transformations have been considered, avoiding spectrographic ones. Thus, two groups of prosodic and spectral features have been selected to capture the rhythm and the frequency respectively. :
   a) Prosodic features[22]
      - Root-mean-square (RMS): value for each frame, either from the audio samples or from a spectrogram S. It represents the energy of the signal and can be calculated by taking the root average of the square of the amplitude. The global energy of the signal x:

b) Spectral features.
- Mel-Frequency Cepstral Coefficients (MFCCs): are a compact representation of the spectrum(When a waveform is represented by a summation of a possibly infinite number of sinusoids) of an audio signal.
- Chroma_stft: Compute a chromagram from a waveform or power spectrogram.
- Spectral_centroid (spec_cent): Compute the spectral centroid.
- Spectral_bandwidth (spec_bw): Compute p'th-order spectral bandwidth.
- Spectral_rolloff (rolloff): Compute roll-off frequency.
- Zero_crossing_rate (zcr): It is the calculation of how many times a signal is crossing its zero axis. Due to a change in peoples' biological and psychological behaviour with a change in their emotion, it also changed how many times a signal crossed its zero axis.

The seven features are calculated for each Window of WS secs.

3. L16-18: After computing the features a PCA analysis is carried out discarding the features above the 90% of representativity.
4. L20-28: Each problem and feature has to be analysed carefully to decide which features are suitable to be optimised. In the case of our problem, MFCC feature will be the one.
5. L30-34: Once the features have been optimized, the best model can be obtained through a cross-validation process to be deployed in the proper device.
6. L33-34: The Model obtained in previous step has to be adapted to the embedded device in order to be deployed in a real context.

2.1 MFCC optimization

As one of the most common features and powerful[18, 3], in the absence of noise, in the SER field is the MFCC feature, it has been selected to be optimised. The steps to calculate MFCCs are[21]:

1. Frame the signal into short frames (25ms is standard). This means the frame length for a 16kHz signal is 0.025*16000 = 400 samples with a sample hop length of 160 samples.
2. Take the Fourier transform of each frame.
3. Map the powers of the spectrum obtained above onto the mel scale, using overlapping triangular windows or alternatively, cosine overlapping windows.
4. Take the logs of the powers at each of the mel-frequencies.
5. Take the discrete cosine transform of the list of mel log powers as if it were a signal.
6. The MFCCs are the amplitudes of the resulting spectrum.

Hence, we have two relevant issues when calculating MFCC: the number of coefficients that determines the precision of the representation of the original signal (n_mfcc) and the overlapping size (hop_length) between frames(see Fig.2).

Algorithm 1 SER_SYSTEM(D: Dataset, NF: Number of Files in D, WS: Window Size, WO: Window Overlapping, S: Significance value for PCA, Features: Features, FP: Features Parameters, NE: Number of Emotions, NP: Number of Participants, SR: SamplingRate)

```
1: Step1: Preprocessing
2: W ← []
3: for f in 1:NF do
4:    D[f]← SetStandardSamplingRate(D[f], SR)
5:    D[f]← ConvertToMono(D[f])
6:    W← W + WindowFraming(D[f], WS, WO)
7: end for
8:
9: Step2: Feature calculation
10: for w in 1:Size(W) do
11:    for ft in 1:Size(Features) do
12:       FEATURES[w, ft] ← ComputeFeature(W, ft, FP[ft])
13:    end for
14: end for
15:
16: Step3: Feature selection
17: PCAContributions← PCA(D, FEATURES, S, FP
18: FEATURES'[] ← DiscardFeatures(FEATURES, PCAContributions)
19:
20: Step4: Feature optimization
21: for ft in 1:Size(FEATURES') do
22:    if ft should be OPTIMIZED then
23:       OPTIMISED_FP ← Meta-Heuristic(FEATURES')
24:       for w in 1:Size(W) do
25:          OPTIMISED_FEATURES[w, ft] ← ComputeFeature(W, ft, OPTI-
          MISED_FP)
26:       end for
27:    end if
28: end for
29:
30: Step5: Train Clasification Model
31: MODEL ← CrossValidationRun(OPTIMISED_FEATURES, NF)
32:
33: Step6: Test Clasification Model
34: OUTPUT_TEST ← ModelDeployment(MODEL, OPTIMISED_FEATURES, NF)
```

The default window size has been set to 2048 tics and the n_mfcc and hop_length have taken values in the ranges 20-128 and 4-2048 respectively. Between the meta-heuristics available in the literature, it has been selected one the simplest and effective, Simmulated Anneling (SA), [10]. The parameters used have been: i) Stopping criteria: FunctionTolerance under 1.0e-4 and individuals with two vaariables: number of parameters of MFCC and hop_length.

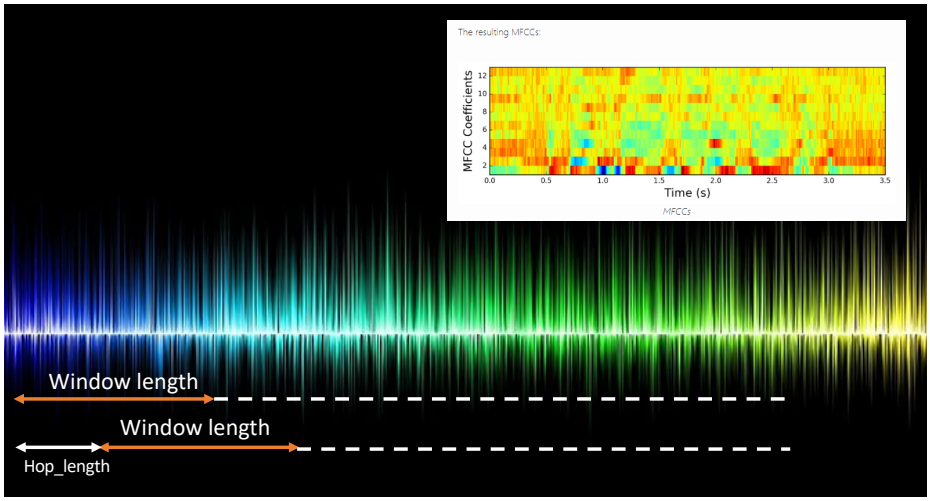

Fig. 2: MFCC coefficients example and Window framing (hop_length)

3 Numerical results

3.1 Materials and methods

In order to validate the proposal presented in this work three public Simulated SEDs, and one Elicited SED have been selected: EmoDB[5], TESS[19], SAVEE[12, 11, 13] and RAVDESS[16] (see table 1).

Eleven CML techniques has been selected in order to compare the efect of the MFCC optimization: BernoulliNB [17], DecisionTree[14, 1], RandomForestClassifier[4], ExtraTrees[9], KNeighbors, RidgeClassifierCV, SVC[6], AdaBoost [23], GradientBoosting[14], MLP (Multi Layer Perceptron) [15] and XGB[8].

3.2 The results

The accuracy for each model after running SA 500 epocs, optimising MFCC for the eleven chosen models and for the four studied datasets as well as the fusion of all the datasets (ALL), can be seen in table 2. It can be stated that ExtraTrees

Table 1: Summary of the Speech Emotion Dataset details

Dataset	Participants	Language	Type	Emotions	#Utterances
EmoDB	Actors	German	Simulated	anger, disgust, fear, happiness,sadness, surprise and neutra (7)	535
TESS	Nonprofessional actors (Mixed gender)	English	Simulated	anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral (7)	2800
SAVEE	Actors (Male)	English	Simulated	anger, disgust, fear, happiness, sadness, surprise and neutral (8)	480
RAVDESS	Professional Actors (Mixed gender)	English	Elicited	neutral, calm, happy, sad, angry, fearful, surprise, and disgust (8)	1440

model ourperforms clearly the remaining models. Moreover, the dataset TESS is one with the best results, since it comprises just two female participants.

	Accuracy				
Models/Datasets	ALL	RAVNESS	SAVEE	TESS	EMO-DB
BernoulliNB	49.31%	38.13%	46.46%	88.71%	49.56%
DecisionTree	66.44%	49.67%	58.33%	92.11%	57.58%
ExtraTrees	84.51%	75.29%	70.00%	99.96%	77.39%
RandomForest	82.49%	71.70%	69.37%	99.89%	73.66%
MLP	62.23%	44.25%	58.33%	98.96%	63.75%
KNeighbors	48.76%	30.51%	37.92%	69.29%	44.88%
RidgeCV	67.42%	60.24%	70.83%	99.68%	72.35%
SVC	24.21%	23.25%	25.42%	26.43%	29.18%
AdaBoost	35.20%	36.79%	45.42%	67.61%	33.09%
XGB	84.11%	73.57%	70.42%	99.57%	74.58%
GradientBoosting	77.37%	65,09%	70.83%	99.54%	72.35%

Table 2: Accuracy for eleven CML method carried out on datasets RAVNESS, SAVEE, TESS and EMO-DB

Fig. 3 shows the evolution of the Accuracy for the metaheuristic SA carring the ExtraTree CML technique out. For the sake of space just the curve corresponding to the ExtraTrees technique has been included. It can be observed that the accuracy of all the datasets has raised aproximately the same.

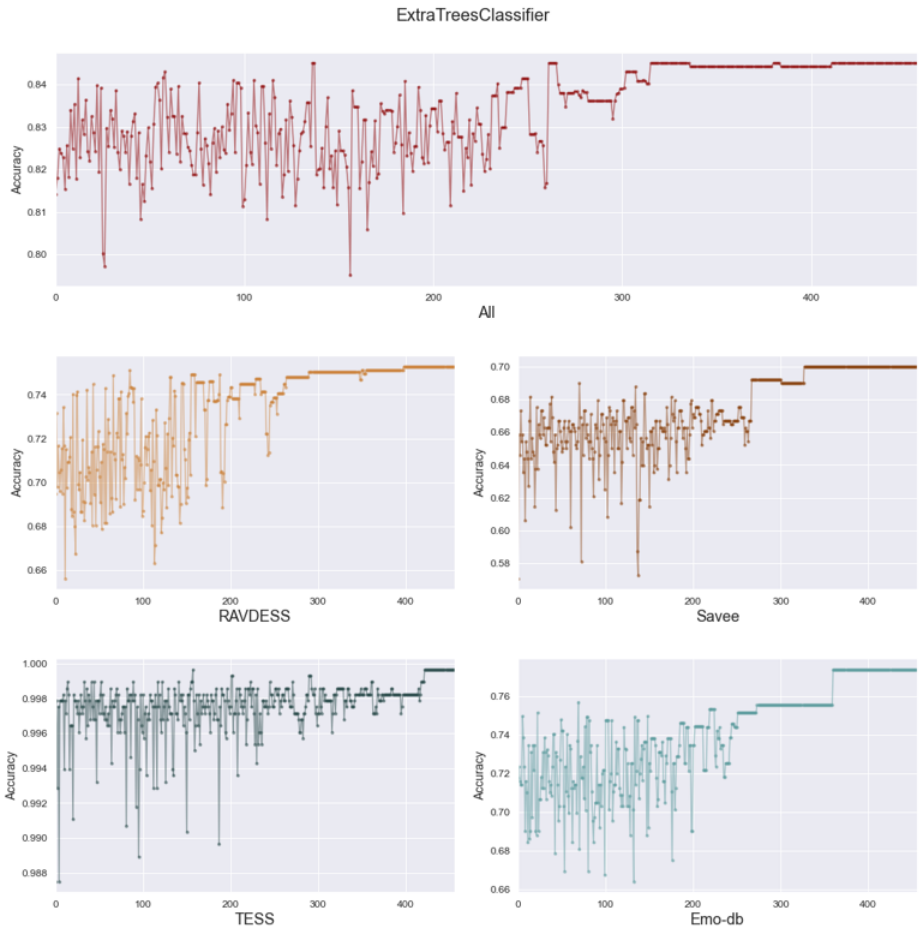

Fig. 3: Evolution of the accuracy for the metaheristic SA for the model Extra-Trees (RAVDESS, Savee, TESS and EmoDB train-datasets

Figure 4 includes the boxplot for the baseline models obtained with the parameters by default for MFCC (Base) compared with the models trained with optimized MFCC parameters (Best). It can be stated that the optimized models outperforms all the baseline models.

4 Conclusion and future work

A simple SER method is presented, including SER algorithm based on a optimising two of the parameters of the MFCC transform. Seven prosodic and spectral features has been studied including the MFCC. The best model obtained, ExtraTress, enhances all the baseline models accurancy and specially the dataset

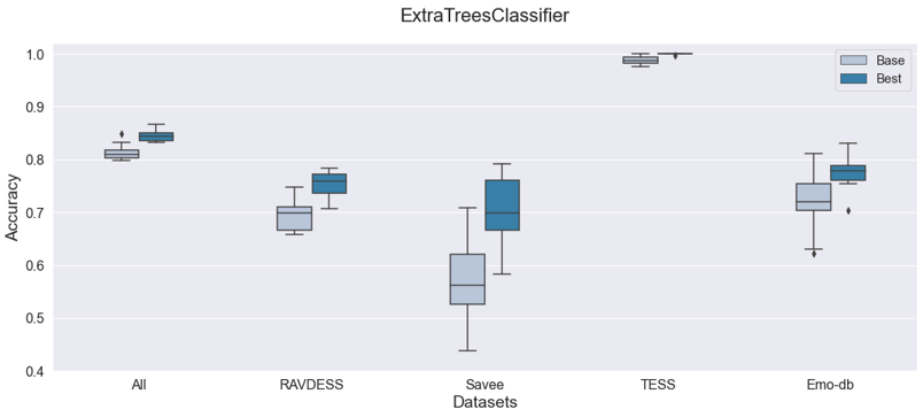

Fig. 4: Comparison of 10 k-fold boxplot with base and optimized parameters

SAVEE with an improvement of a 10% . Attending that one dataset corresponds to german speakers and the three to english native speakers, it can be seen that the results are invariant to this fact obtaining good results for all the models and dataset event the fusion of datasets (All).

Moreover, we think that the kind of records of the dataset are gather, simulated, electided or natural, should be analysed in order to study how affect this informatin in the transfer learning process. More datasets corresponding to other languages must be included in the study.

Acknowledgement

This research has been funded partially by Spanish Ministry of Economy, Industry and Competitiveness (MINECO) under grant TIN2017-84804-R/PID2020-112726RB-I00.

References

1. Random forests - classification description, `https://www.stat.berkeley.edu/$\sim$breiman/RandomForests/cc_home.htm`
2. Akçay, M.B., Oğuz, K.: Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication **116**(October 2019), 56–76 (2020). https://doi.org/10.1016/j.specom.2019.12.001
3. Anagnostopoulos, C.N., Iliou, T., Giannoukos, I.: Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artificial Intelligence Review **43**(2), 155–177 (2012). https://doi.org/10.1007/s10462-012-9368-5
4. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (oct 2001). https://doi.org/10.1023/A:1010933404324, `https://link.springer.com/article/10.1023/A:1010933404324`

5. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of German emotional speech. 9th European Conference on Speech Communication and Technology pp. 1517–1520 (2005)
6. Chang, C.C., Lin, C.J.: LIBSVM: A Library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(3) (2011). https://doi.org/10.1145/1961189.1961199, www.csie.ntu.edu.tw/
7. Dzedzickis, A., Kaklauskas, A., Bucinskas, V.: Human emotion recognition: Review of sensors and methods. Sensors (Switzerland) **20**(3) (2020). https://doi.org/10.3390/s20030592
8. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Annals of Statistics (5), 1189–1232 (oct). https://doi.org/10.1214/aos/1013203451
9. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning **63**(1), 3–42 (apr 2006). https://doi.org/10.1007/s10994-006-6226-1, https://link.springer.com/article/10.1007/s10994-006-6226-1
10. Greenwood, C.T.: A n Overview (x), 102–104 (1986)
11. Haq, S., Jackson, P.J.B.: Speaker-dependent audio-visual emotion recognition. In: Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Norwich, UK (2009)
12. Haq, S., Jackson, P.J.B.: Machine Audition: Principles, Algorithms and Systems, chap. Multimodal, pp. 398–423. IGI Global, Hershey PA (2010)
13. Haq, S., Jackson, P., Edge, J.: Audio-visual feature selection and reduction for emotion classification. In: Expert Systems with Applications. vol. 39, pp. 7420–7431 (2008), http://epubs.surrey.ac.uk/7738/4/licence.txt
14. Hastie, T., Tibshirani, R., Friedman, J.: Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction. Tech. rep.
15. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR (2015)
16. Livingstone, S.R., Russo, F.A.: 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 **13**(5), e0196391 (may 2018). https://doi.org/10.1371/journal.pone.0196391, https://dx.plos.org/10.1371/journal.pone.0196391
17. Manning, C.D., Raghavan, P., Schuetze, H.: The Bernoulli model. In: Introduction to Information Retrieval, pp. 234–265 (2009), https://nlp.stanford.edu/IR-book/html/htmledition/the-bernoulli-model-1.html
18. Pandey, S.K., Shekhawat, H.S., Prasanna, S.R.: Deep learning techniques for speech emotion recognition: A review. 2019 29th International Conference Radioelektronika, RADIOELEKTRONIKA 2019 - Microwave and Radio Electronics Week, MAREW 2019 (2019). https://doi.org/10.1109/RADIOELEK.2019.8733432
19. Pichora-Fuller, M. Kathleen; Dupuis, K.: Toronto emotional speech set (TESS) Collection. Tech. rep. (2015), http://www.alz.org/research/funding/global%7B_%7Dbiomarker%7B_%7Dconsortium.asp
20. Rahi, P.K.: Speech Emotion Recognition Systems: Review. International Journal for Research in Applied Science and Engineering Technology **8**(1), 45–50 (2020). https://doi.org/10.22214/ijraset.2020.1007
21. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Com-

munication **54**(4), 543–565 (2012). https://doi.org/10.1016/j.specom.2011.11.004, http://dx.doi.org/10.1016/j.specom.2011.11.004
22. Väyrynen, E.: Emotion recognition from speech using prosodic features. Ph.D. thesis (2014), http://www.oulu.fi/cse/personnel/eero-v{\"{a}}yrynen/homepage
23. Zhu, J., Zou, H., Rosset, S., Hastie, T.: Multi-class AdaBoost *. Tech. rep. (2009)