# SPR Final report

Vlad Tabusca and Sara Frusone

June 18, 2021

## Abstract

In the last years Speech Emotion Recognition (SER) has become one of the most important topics in the field of Speech Recognition.

In this study an approach that utilizes Deep Neural Networks is proposed. Firstly, a theoretical overview of the state-of-the-art methodologies for SER is presented. Then, a description of the procedure adopted is given, followed by a presentation and comparison of the results obtained. Lastly, brief suggestions for further works are offered.

# Contents

**Introduction** **1**

**1 State of the art** **2**

**2 Methods** **4**

2.1 Datasets . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.2 Overfitting problems . . . . . . . . . . . . . . . . . . . . . . . . 4

2.2.1 Data augmentation . . . . . . . . . . . . . . . . . . . . . 5

2.3 Features extraction . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.3.1 openSMILE . . . . . . . . . . . . . . . . . . . . . . . . . 6

2.4 Feature dimension reduction . . . . . . . . . . . . . . . . . . . . 7

2.5 Classification model . . . . . . . . . . . . . . . . . . . . . . . . 7

2.6 End-to-End learning for speech emotion recognition . . . . . . . . 8

**3 Experiments and Results** **9**

3.1 Experiments conducted . . . . . . . . . . . . . . . . . . . . . . . 10

3.1.1 Live demonstration . . . . . . . . . . . . . . . . . . . . . 10

3.2 Presentation of the results . . . . . . . . . . . . . . . . . . . . . 10

**4 Further works** **11**

4.1 Emotion changes . . . . . . . . . . . . . . . . . . . . . . . . . . 11

4.2 Adding Visual data . . . . . . . . . . . . . . . . . . . . . . . . . 12

**5 Conclusions** **13**

**Bibliography** **17**

# Introduction

Human communication is based on speech as a way of transmitting information, as a main aspect of every society since the dawn of our species.
Speech contains a lot of hidden information, in the emotions of the speaker: it can reveal a wide variety of notions about the events that surround the speaker and influence his way of expressing.
Analyzing the human voice-pattern and speech-pattern has become an important subject of research in the last few years, together with the analysis of facial expressions. In fact, those results can be applied in many fields, such as human-machine interactions, education, disability-aid and communication.
This project presents a way to detect emotions elicited by the speaker while talking. For example, Speech produced in a state of fear, anger, or joy becomes loud and fast, with a higher and wider range in pitch, whereas emotions such as sadness or tiredness generate slow and low-pitched speech.
In particular, the classification model presented is based on deep neural networks (DNN), used to automatically detect emotions present in speech signals. The model has been trained using the a combination of 4 English voiced data-sets: SAVEE, TESS, CREMA_d and RAVDESS, to classify six different emotions (namely: anger, disgust, fear, happiness, sadness, surprise) and a neutral state, starting from human-generated audio signals.

# Chapter 1

# State of the art

Understanding the feeling of the collocutor who is speaking is complementary to the understanding of the meaning of the speak, and it is related to a better comprehension of the message.
Several studies have been conducted on recognizing feelings from auditory data. In recent years, SPR studies focused on building means and methods to create such an understanding for computers. The process of recognizing emotions from speech involves extracting the characteristics from a corpus of emotional speech selected or implemented and, after that, the classification of emotions is done on the basis of the extracted characteristics, such as the spectral characteristics. To do that a mathematical algorithm model is needed.
The performance of the classification of emotions strongly depends on the good extraction of those characteristics. There are two major approaches in SER, either recognizing based on the three dimensions of emotions or recognition based on statistical pattern recognition techniques for the qualitative named emotions. Infact, whereas psychologists based their approach on three measures that describes all the emotions. They are: pleasure, arousal, and dominance. A combination of these qualities will create a vector that will be in one of the defined emotion territories, and based on that, we can report the most relevant emotion. We calculate degrees of correlation between the given signal and passion, arousal, and dominance, and then using a hierarchical classifier, the complex emotion is determined. Using pleasure, arousal, and dominance, we can describe almost any emotion, but such a deterministic system will be very complex to be implemented with machine learning.
An other approach is to use statistical pattern recognition methods, such as the Gaussian mixture model (GMM) , support vector machine (SVM) , hidden Markov model (HMM) , artificial neural network (ANN) , deep neural network (DNN) , and genetic algorithm (GA).

First, to classify and cluster emotions, It is necessary to model them using features extracted from the speech. Many authors agree that the most important audio characteristics to recognize emotions are spectral energy distribution,Teager Energy Operator (TEO), MFCC, zero crossings Rate (ZCR) and the energy parameters of the filter bank Energies (FBE). Any of these categories have benefits in classifying some emotions and weaknesses in detecting others.

For example, the prosody features usually focused on fundamental frequency (F0), speaking rate, duration, and intensity, are not able to confidently differentiate angry and happy emotions from each other.

Voice quality features are usually dominant in the detection of emotions of the same speaker. They are hard to be used in a speaker-independent setting because they differ from speaker to speaker. Spectral features have been extensively analyzed to derive emotions from speech. The immediate advantage that they have compared to prosody features is that they can confidently distinguish angry from happy. However, an area of concern is that the magnitude and shift of the formants for the same emotions vary across different vowels, and this would add more complexity to an emotion recognition system, and it needs to be speech content-aware.

For each feature category, there are various standard feature representations. Prosody features are usually being shown by F0 and measures related to speaking rate, and spectral features are generally being described using one of the cepstrum-based representations available.

# Chapter 2

# Methods

Traditional speech emotion recognition can be divided into two steps. First performs feature extraction from the speech signal, and secondly applies a classifier to determine the emotional category. Some approaches add gender information to improve accuracy.

The classification model of emotion recognition can be based on many strategies, as we said before, such as convolutional neural networks (CNN), Support Vector Machine (SVM) classifier, MLP Classifier. The method we are going to consider is based on DNN, because it has become a flagship in the fields of artificial intelligence. Deep learning has surpassed state- of-the-art results in many domains.

The audio file is divided into frames, usually using a fixed window size, to obtain statistically stationary waves. The amplitude spectrum is normalized with a reduction of the ”Mel” frequency scale. This operation is performed for empathizing the frequency more meaningful for a significant reconstruction of the wave as the human auditory system can perceive.

For each audio file, five features have been extracted. The features have been generated by converting each audio file to a floating-point time series.

## 2.1 Datasets

You can read a presentation of each dataset we considered in our analysis in the appendix section. Most of them are in english, while one is in italian, but it has been considered in a separate experiment which has been reported.

## 2.2 Overfitting problems

Data sparsity could cause a machine learning model not able to learn the true data distribution, which leads to the overfitting problem. For example, overfitting

would occur when training a deep model with only hundreds of samples, but each sample has thousands of features. To solve this problem, regularization can be used to impose constraints on the model. Another general solution is dimension reduction with sparsity constraint. This approach will be effective when redundant features exist; otherwise, it will eliminate useful information and result in performance degradation.
To solve the data sparsity problem, the training set can be enlarged by data augmentation. Traditional data-augmentation methods typically transform (e.g., adding noise and reverberation to speech signals and cropping, rotating, and flipping of images) the original data, followed by augmenting the transformed data to the original data. More advanced methods augment the data based on GANs or variants of GANs, such as conditional GANs (cGANs) or adversarial autoencoders (AAEs).

### 2.2.1 Data augmentation

Data augmentation is the process by which we create new synthetic data samples by adding small perturbations on our initial training set. To generate syntactic data for audio, we can apply noise injection, shifting time, changing pitch and speed. The objective is to make our model invariant to those perturbations and enhace its ability to generalize. In order to this to work adding the perturbations must conserve the same label as the original training sample.
To have a bigger dataset, It can be usefull to use 4 dataframes togeter. In this way the audio files came from different speakes, and different situations, so we will create a sort of new dataset, excluding some emotions which were not important or useful. The huge dataset obtained is not balanced, and this is an information we should consider in the analysis.

## 2.3 Features extraction

Audio data cannot be understood by the models directly so we need to convert them into an understandable format for which feature extraction is used. There exist many methods to extract features, this part is very important for the analysis. The audio signal is a three-dimensional signal in which three axes represent time, amplitude and frequency. According to the theoretical method presentation, we have decided to consider:
1. Zero Crossing Rate : The rate of sign-changes of the signal during the duration of a particular frame.

2. Energy : The sum of squares of the signal values, normalized by the respective frame length.
3. Entropy of Energy : The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4. Spectral Centroid : The center of gravity of the spectrum.
5. Spectral Spread : The second central moment of the spectrum.
6. Spectral Entropy : Entropy of the normalized spectral energies for a set of sub-frames.
7. Spectral Flux : The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8. Spectral Rolloff : The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9. MFCCs Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
10. Chroma Vector : A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
11. Chroma Deviation : The standard deviation of the 12 chroma coefficients.
In this project we will extract 5 features by hand, using the appropriate Python commands: Zero Crossing Rate, Chroma stft, MFCC, RMS(root mean square) value MelSpectogram to train our model.
An other option is to use OpenSMILE to extract the features, see section below for more information.

### 2.3.1 openSMILE

`openSMILE` (open-source Speech and Music Interpretation by Large-space Extraction) is an open-source toolkit for audio feature extraction and classification of speech and music signals. `openSMILE` is widely applied in automatic emotion recognition for effective computing.
The configuration used is based on the `emobase.conf` and it allows the computation of 988 acoustic features, which included the following low-level descriptors (LLD): Intensity, Loudness, 12 MFCC, Pitch (F0), Probability of voicing, F0 envelope, 8 LSF (Line Spectral Frequencies) and Zero-Crossing Rate.
They include also statistical functionals as min/max, delta regression, linear and quadratic error.

## 2.4 Feature dimension reduction

In many practical applications it is of interest to reduce the dimensionality of the data. In particular, this is useful for data visualization, or for investigating the "effective" dimensionality of the data. This problem is often referred to as dimensionality reduction and can be seen as the problem of defining a map

$$M : X = \mathbb{R}^D \to \mathbb{R}^k, \quad k \ll D$$

according to some suitable criterion.
PCA is arguably the most popular dimensionality reduction procedure. It is a data driven procedure that given an (unsupervised) sample $S = (x_1, \ldots, x_n)$ derives a dimensionality reduction defined by a linear map $M$. PCA can be derived from several perspectives. Here we provide a geometric/analytical derivation.
We begin by considering the case where $k = 1$. We are interested in finding the single most relevant dimension according to some suitable criterion. Recall that, if $w \in \mathbb{R}^D$ with $\|w\| = 1$, then the (orthogonal) projection of a point $x$ on $w$ is given by $\left(w^T x\right) w$. Consider the problem of finding the direction $p$ which allows the best possible average reconstruction of the training set, that is the solution of the problem

$$\min_{w \in S^{D-1}} \frac{1}{n} \sum_{i=1}^{n} \left\| x_i - \left(w^T x_i\right) w \right\|^2$$

## 2.5 Classification model

The deep neural network(CNN) is designed in order to achieve the best performance and the lower computational cost. To do that, we built differents neural networks and we selected the most appropriate. The best one, a sequential model, has been reported in the notebook, his structure is composed by conv1d, maxpooling, dropout, flatten and dense layers, A detail overwiew of the models structure is showed in the notebook where a summary scheme has been reported. In order to fix the correct number of epoches and the batchsize, again we made many experiments up to the final version. We built three different models based on the same neural network, but trained as follow, one on TESS dataset, one on EMOVO and one in a self-built augumented Dataset.

## 2.6 End-to-End learning for speech emotion recognition

With the rise of deep learning techniques, recent studies have proposed emerging end-to-end recognition structures. In this section, we review these related works.

The term “end-to-end” in the deep learning field means a complex learning system by applying gradient-based learning to the system. The network can directly convert the input signal into corresponding mapped output, bypassing the intermediate step in traditional algorithms, such as feature extraction in some emotion recognition systems. Therefore, we can consider the network a black box trained by the global objective function. The main idea is that the network automatically learns a representation of the raw input signal that better suits the task at hand, leading to improved performance.

The input of the raw speech signal passes through several convolution layers, followed by LSTM layers to capture the temporal structure.

All current end-to-end discrete speech emotion recognition algorithms use the speech spectrogram as an input. That is, although current algorithms are an end-to-end structure, the algorithms model the speech emotion recognition task as an image classification problem.

Looking at the SER problems as a CV image classification, data augumentation can be performed by shifting the image, zooming, rotating,...

It is also possible to give an alternative approach to use an end-to-end deep learning algorithm to capture the information clues from emotional speech. Therefore, the input of the algorithm is the raw speech data, rather than a spectrogram.

# Chapter 3

# Experiments and Results

You will find a more accurate description of each step in the attached Notebook.

To have a more accurate model we thought it could be interesting to built a dataset from the existing ones, since the fact that they were all in English but built in different situations, and with different kind of recording (Read in the appendix further information about each dataset). After with a data augumentation technique, we augumented the data in order to make the model more robust to noise for example or other type of changes.

During that part of the procedure in the notebook we also show many plots about the signal in order to be coherent with what we studied during the SPR course.

The pre processing of the data is very important. We also want to highlight that the dataset classes are unbalanced, this is an interesting point because we expect the model to be more robust to classify the emtions in the most numerous class, and less in the smallest one which is "surprise". The extraction of features is a relevant part, as discussed in the theoretical introduction the choice of features to be extracted is huge, after many attempts we decided to considering the following features:.......... We did the extraction procedure by hand, an other option could have been using Opensmile.

We split the datset in training and validation in a $75-25\%$ percentage division. We use the training to make the model learn, and the validation to test it. It is important to do correctly that part in order to preserve the unbalancing among classes.

Then we build a model using neural networks as prevoiusly expained.

## 3.1 Experiments conducted

We tried the experiment with many different epoches and batchsize in order to make it performing and as fast as possible.
We did three experiments: one with the italian dataset EMOVO, one wih Tess and an other one with the augumented dataset. We wanted also to point out how changes the behavour of the classifier in balanced or unbalanced situation. We achieved a good accuracy on the training and on the test set in all the experiments, more details are in the attached Notebooks.
We save each model, coming from each one of the experiments in order to try each one also with some data which do not came from the dataset, in the Notebook you will find an interactive part where you can record your voice and see if the model classify it correctly.

### 3.1.1 Live demonstration

An example of new file to try the model on, could be an audio live recording. We built a part of the notebook which takes in input an audio live file that can been recorded live from the user, to make it more interactive.

## 3.2 Presentation of the results

You will find a deep comment about the results in the attached notebooks. To sum up we got better results on the two small dataset EMOVO and TESS, but we think them to be less robust, if compared to the augumented dataset, which contains more speaker, more accents, more different situations, despite of the worst accuracy.

# Chapter 4

# Further works

## 4.1 Emotion changes

For a deeper understanding of emotion in speech, an investigation into emotion change detection is very interesting. Detecting emotion changes is somewhat analogous to speaker change detection.

Most literature has focused on detecting emotion change points in time, which we refer to as Emotion Change Detection (ECD). It is found in Böck and Siegert (2015) that emotional evolution both inter- and intra-speaker is detectable using per-file emotion recognition methods. Also, there have been some studies explicitly attempting to localize the time when emotion changes occur, among different emotion categories using audio features (Xu and Xu, 2009; Pao et al., 2010; Fan et al., 2014) or psychological measures (Leon et al., 2004).

Specifically, the presence of emotion changes was detected via a large residual between measured emotion and estimated emotion. Recognizing emotions using pre-segmented speech utterances results in a loss in continuity of emotions and does not provide insights into emotion changes. An investigation into emotion change detection is also possible from the perspective of exchangeability of data points observed sequentially using a martingale framework. Within the framework, a per-frame GMM likelihood based approach is a measure of strangeness from a particular emotion class.

The problem of localizing emotion change points in time can be investigated from the perspective of testing exchangeability using a martingale framework, where data points (frame-based features) from speech are observed one by one. This method potentially offers higher temporal resolution than using large sliding windows. Moreover, emotional models may be helpful to reduce effects of phonetic variability compared with methods that require no prior knowledge of emotions.

However, applying these methods into an emotion change detection task remains problematic because of the phonetic and speaker variability embedded in emotional speech and the complex nature of emotion (e.g. a person might experience more than one emotion at a time).
To do experiments the IEMOCAP database is very useful, since it shows how the martingale framework offers significant improvements over the baseline GLR method for detecting emotion changes not only between neutral and emotional speech, but also between positive and negative classes along the arousal and valence emotion dimensions.

## 4.2 Adding Visual data

To recognise emotions it would be interesting not to consider only the audio but also the visual part.
There exist two versions of the IEMOCAP dataset, one for speech and an other one for both speech and motion. For each session, one actor wears the Motion Capture (MoCap) camera data which records the facial expression, head and hand movements of the actor. The Mocap data contains column tuples, for facial expressions the tuples are contained in 165 dimensions, 18 for hand positions and 6 for head rotations. As this Mocap data is very extensive we use it instead of the video recording in the dataset. These three modes (Speech, Text, Mocap) of data form the basis of our multi-modal emotion detection pipeline. For a more complete analysis it could be interesting to consider that multi modal emotion detection, where the SPR part encounter the CV part. Unfortunately, here there is not enough time to go further.

# Chapter 5

# Conclusions

In this work, we presented an architecture based on deep neural networks for the classification of emotions using some audio Database of Emotional Speech. Three models have been trained to classify seven different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised) and obtained an overall performance which goes from the 70% of accuracy to the 90% on the train and which is around the 60% on the validation set, according to the dataset we are considering. To obtain such a result, we extracted manually the features from the audio files used for the training.
At the end we performed the models on a live demonstration.

# Appendix

## IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is an acted, multimodal and multispeaker database.
It consists of 12 hours of audio-visual data of dyadic sessions where 10 actors, male and female, perform improvisations or scripted scenarios, specifically selected to elicit emotional expressions.
After the audio-visual data has been collected it is divided into small utterances of length between 3 to 15 seconds which are then labelled by evaluators. Each utterance is evaluated by 3-4 assessors. IEMOCAP database is annotated by multiple annotators into categorical labels and dimensional labels (neutral, happiness, sadness, anger, surprise, fear, disgust,frustration,excited, other). We consider only 6 of them.
Along with the .wav file for the dialogue we also have the transcript for each the utterance. Next we preprocess the IEMOCAP data for these modes.

## EMOVO

EMOVO is a database in Italian built from the voices of up to 6 actors who played 14 sentences simulating 6 emotional states (disgust, fear, anger, joy, surprise, sadness) plus the neutral state. These six emotions are found in most of the literature related to emotional speech. The recordings were made with professional equipment in the Fondazione Ugo Bordoni laboratories. The paper also describes a subjective validation test of the corpus, based on emotion-discrimination of two sentences carried out by two different groups of 24 listeners. The test was successful because it yielded an overall recognition accuracy of 80%. It is observed that emotions less easy to recognize are joy and disgust, whereas the most easy to detect are anger, sadness and the neutral state.

## TESS

Toronto emotional speech set (TESS) Collection is a dataset of auditory data. These stimuli were modeled on the Northwestern University Auditory Test No. 6 (NU-6; Tillman and Carhart, 1966). A set of 200 target words were spoken in the carrier phrase "Say the word ... " by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 stimuli in total.

Two actresses were recruited from the Toronto area. Both actresses speak English as their first language, are university educated, and have musical training. Audiometric testing indicated that both actresses have thresholds within the normal range.

## CREMA- D

CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

In our specific situation, CREMA-D dataset is the sheer variety of data which helps train a model that can be generalised across new datasets. Many audio datasets use a limited number of speakers which leads to a lot of information leakage but CREMA-D has many speakers. For this fact, the CREMA-D is a very good dataset to use to ensure the model does not overfit.

## RAVDESS

Speech audio-only files (16bit, 48kHz .wav) from the RAVDESS. Full dataset of speech and song, audio and video (24.8 GB) available from Zenodo. Construction and perceptual validation of the RAVDESS is described in our Open Access paper in PLoS ONE.

This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions includes neutral, calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional in-

tensity (normal, strong), with an additional neutral expression.
File naming convention : Each of the 1440 files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav). These identifiers define the stimulus characteristics.

## SAVEE

Surrey Audio-Visual Expressed Emotion (SAVEE) IS an audio-visual database of expressed emotions. It has been recorded as a pre-requisite for the development of an automatic emotion recognition system. The SAVEE database was recorded from four native English male speakers (identified as DC, JE, JK, KL), postgraduate students and researchers at the University of Surrey aged from 27 to 31 years. It consists of recordings from 4 male actors in 7 different emotions, six basic emotions and neutral.
The text material consisted of 15 TIMIT sentences per emotion: 3 common, 2 emotion-specific and 10 generic sentences that were different for each emotion and phonetically-balanced. The 3 common and $2 \times 6 = 12$ emotion-specific sentences were recorded as neutral to give 30 neutral sentences. The database was evaluated by 10 subjects with respect to recognizability for each of the audio, visual and audio-visual data.
Emotion has been described psychologically in discrete categories: anger, disgust, fear, happiness, sadness and surprise. .
Human evaluation and machine learning experimental results show the usefulness of this database for research in the field of emotion recognition.

# Bibliography

[1] https://github.com/naxingyu/opensmile/blob/master/config/emobase.conf

[2] https://www.mdpi.com/2073-8994/11/8/1018

[3] https://towardsdatascience.com/anova-for-feature-selection-in-machine-learnin

[4] https://medium.com/heuristics/audio-signal-feature-extraction-and-clustering-

[5] Speech Emotion Recognition with deep learning : https://www.sciencedirect.com/science/article/pii/S1877050920318512

[6] https://www.mdpi.com/journal/sensors

[7] www.sciencedirect.com

[8] https://sail.usc.edu/iemocap/

[9] Detecting the instant of emotion change from speech using a martingale framework: https://ieeexplore.ieee.org/document/7472668

[10] Emotional recognition from the speech signal for a virtual education agent: https://iopscience.iop.org/article/10.1088/1742-6596/450/1/012053/pdf

[11] Ting-Wei Sun, *End-to-End Speech Emotion Recognition with Gender Information Ting-Wei Sun Graduate Institute of Electrical Engineering, National Taiwan University, Taipei, Taiwan.*

[12] Samarth Tripathi, *MULTI-MODAL EMOTION RECOGNITION ON IEMOCAP WITH NEURAL NETWORKS. Samarth Tripathi‡ Sarthak Tripathi? Homayoon Beigi†* https://arxiv.org/pdf/1804.05788.pdf

[13] Lorenzo Rosasco, *Introductory Machine Learning Notes1 Lorenzo Rosasco,DIBRIS, Universita' degli Studi di Genova LCSL, Massachusetts Institute of Technology and Istituto Italiano di Tecnologia.*

[14] Yu, *Yu, Yeonguk; Kim, Yoon-Joong. 2020. "Attention-LSTM-Attention Model for Speech Emotion Recognition and Analysis of IEMOCAP Database" Electronics 9, no. 5: 713.* `https://doi.org/10.3390/electronics9050713`

[15] Kjersti Engan, *Frame Based Signal Representation and Compression*, Departement of Electrical and Computer Engineering Stavanger University College Norway.

[16] Ting-Wei Sun *End-to-End Speech Emotion Recognition with Gender Information, Graduate Institute of Electrical Engineering, National Taiwan University, Taipei, Taiwan.*

[17] Zou Cairong, *A Novel DBN Feature Fusion Model for Cross-Corpus Speech Emotion Recognition Zou Cairong,1,2 Zhang Xinran,2 Zha Cheng,2 and Zhao Li2*

[18] L. Yi, *L. Yi and M. -W. Mak, "Improving Speech Emotion Recognition With Adversarial Data Augmentation Network," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2020.3027600.*