TuniSER: Toward a Tunisian Speech Emotion Recognition System

Abir Messaoudi
iCompass
abir@icompass.digital

Hatem Haddad
iCompass
hatem@icompass.digital

Moez Benhaj Hmida
iCompass
moez@icompass.digital

Mohamed Graiet
ISIMM
mohamed.graiet1@gmail.com

Abstract

Speech Emotion Recognition (SER) has become an important component for Human-Computer interaction. It is generally used in job interviews, caller-agent calls and streaming videos, etc. In the speech emotion recognition literature, many languages have tackled this topic to extract emotions from signals. The purpose of this work is to build a Speech Emotion Recognition model that predicts the emotional state of Tunisian speakers. We explore different pre-trained acoustic models, we detail the process of building the first Tunisian Speech Emotion Recognition dataset (TuniSER) and we describe the training and testing phases. Our experiments' results show that fine-tuning the pretrained multilingual wav2vec 2.0 model on the Automatic Speech Recognition downstream task then building a classifier on top of fit outperformed all the tested models achieving an Accuracy of 60.6%.

1 Introduction

In a conversation, non-verbal communication contains important information such as the speaker's intentions or emotions. This information needs to be processed and recognised. Indeed, speech systems should be able to understand this non-linguistic information. In the recent years, the interest in Speech Emotion Recognition (SER) has increased due to the role that this task plays in improving both the naturalness and efficiency of human-machine interactions. Voice assistants and conversational interfaces have become omnipresent through technology devices such as smartphones and smart home interaction systems. Once these systems capture the emotional content of speech aside from semantics, their capability will increase.

SER is a non-trivial task on account of many reasons such as the ambiguity of defining emotions itself and the non-obvious ability of detecting the natural from the acted emotions. Also, it requires large annotated emotional datasets. Yet, creating such data is cost prohibitive because of the large human efforts involved.

Moreover, SER becomes extremely a complicated task with an under-resourced language like the Tunisian dialect (Fourati et al., 2020) because of the lack of resources and the non-existence of the emotional Tunisian dataset. There are various existing researches in the field of Arabic Speech Emotion Recognition. But, they are basically restricted to MSA (Hifny and Ali, 2019).

The primary research questions we wish to investigate in the paper are:

- Question 1: How to build a labeled speech emotional dataset when it comes to under-resourced dialect?
- Question 2: Can the SER model identify the emotional state of a person regardless of the language or dialect used?
- Question 3: Can we use large multilingual pre-trained acoustic models to classify Tunisian dialectal emotional states?
- Question 4: Which mathematical model for creating reliable recognizers in the case of Tunisian dialect?

The paper structure is described as follows. In Sections 2 we introduce the related work on English and Arabic Sentiment Emotion Recognition. In Section 3 we describe the different steps to build a SER Tunisian dataset. In Section 4, we present the proposed methods to build a SER model. In Section 5 we detail the different experiments and present the outcomes, and finally, the paper is concluded in Section 6.

2 Relation to prior work

For Speech Emotion Recognition problems, different methods have been used such as SVM (Seehapoch and Wongthanavasu, 2013), HMM

(Schuller et al., 2003), decision-trees (Lee et al., 2011) etc. These methods explored different features that detect the emotion held in speech including pitch, shimmer, jitter and MFCCs (Mel-Frequency Cepstral Coefficients) (Ghosh et al., 2016) (Liu et al., 2018). However, one of the major drawbacks of these techniques is that they require a previous knowledge of all the mandatory features that had a direct impact on the emotion recognition task like energy and fundamental frequency (F0). A work done (Sahu, 2019) has revealed that traditional machine learning methods still can reach a high performance as the latest Deep learning models (such as LSTM) in this field.

Deep neural networks represent the latest models in Speech Emotion Recognition. They were used to automatically extract high-level features from audio and have successfully achieved high performances (Han et al., 2014). Since then, different neural network architectures have been used for Speech Emotion Recognition. An innovative study (Zheng et al., 2015) was done by applying CNN for speaker independent emotion recognition systems. They came up with the conclusion that deep learning methods outperform machine learning methods for SER.

Other successful methods were deployed such as RNN and bidirectional long-short term memory (BLSTM) (Lee and Tashev, 2015). Another attempt was to combine a CNN model with a RNN model (Trigeorgis et al., 2016) and it has shown a success as well and an efficient speech emotion recognition. Artificial Neural Networks (Shaw et al., 2016), Deep Convolutional Neural Networks (Zhang et al., 2017) and other deep learning approaches (Abbaschian et al., 2021) were used to bring out the best result for the SER task.

Recently, wav2vec 2.0 (Baevski et al., 2020) has been used in emotion classification task (Pepino et al., 2021) and results in state-of-the-art results for both IEMOCAP and RAVDESS datasets with a recall of 0.67 and 0.84 respectively.

When it comes to Arabic Speech Emotion Recognition (ASER), there are few available datasets for Arabic language. The Basic Arabic Vocal Emotions Dataset (BAVED) (Aouf, 2019) contains Arabic words spelled in three levels of emotions recorded in an audio format, low emotion (tired or exhausted), neutral emotion, and high emotion positive or negative emotions (happiness, joy, sadness, anger). The dataset contains 1935 recordings that are recorded by 61 speakers (45 males and 16 females).

(Mohamed and Aly, 2021) introduced a recognition model for Arabic speech dialogues based on deep learning . The developed model employs the state of the art audio representations including wav2vec2.0 and HuBERT (Hsu et al., 2021). They reached an accuracy of 89% and 87% respectively for wav2vec 2.0 and HuBERT.

(Meddeb et al., 2016) present the main steps to extract and recognize basic emotions (Neutral, Happiness, Sadness, Anger and Fear) in the Arabic speech. They created an Emotional speech database called REGIM_TES (Meddeb et al., 2014) containing 720 speech samples. The length of speech samples is up to 5 Seconds. The selected features in the study are: Pitch of voice, Energy, MFCCs, Formant, LPC and the spectrogram. Results showed that pooling together features extracted at different sites improved classification performances.

As far as we know, there is no Speech Emotion Recognition dataset for the Tunisian Dialect. In the next section, we present our methodology to build a SER dataset for an under-represented language.

3 Tunisian Speech Emotion Recognition Dataset

Speech datasets used for building Speech Emotion Recognition systems are divided into three types namely:

- **Simulated**: Simulated databases are created by reading the same text by different trained speakers with different emotions. The numbers of distinct emotions are important, as they have synthesized emotions, they are disposed to have over-fitted models around emotions a little bit different than what is happening in real life and day-to-day conversations. Those databases make comparing results very easy due to the standardized collections of emotions. Berlin Database of Emotional Speech[1] and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo, 2018)[2], are some simulated datasets used for SER (Abbaschian et al., 2021).

- **Induced**: Semi-natural databases are very

[1] http://emodb.bilderbar.info/start.html
[2] https://smartlaboratory.org/ravdess/

similar to the natural utterances of speech, even if they are made based on scenarios with a contextual setting. The emotions are artificial because speakers know that they are recorded. Unfortunately, those data sets have a limited number of emotions due to the limited cases of the scenarios compared to other types of data sets (Abbaschian et al., 2021)(Zheng et al., 2015).

- **Natural**: The natural datasets are made of fully natural emotions which eliminate the problem of being artificially made. They are very effective because they perfectly represent our daily life due to the contentiousness of emotions and the existence of the background noise and concurrent emotions and the dynamic variation of the speech. However, those characteristics make the detection and the modeling of the emotions more complicated. The number of emotions is limited due to the limited sources (Abbaschian et al., 2021). VAM[3] is one of the most famous natural databases used for SER tasks.

Universal emotions are defined as six categories: happiness, sadness, disgust, fear, surprise, and anger (Collet et al., 1997). Nevertheless, in this study, we focus only on four categories: Happy, Sad and Angry in addition to Neutral because it is largely used in the state of the art (Han et al., 2014) (Zheng et al., 2015).

3.1 Tunisian dialect

Tunisian Arabic (Tunisian), is the set of dialects of Maghrebi Arabic spoken in Tunisia known locally as Derja (Sayahi, 2014). It is used in the daily life and turn out to be the language of online communication since the 1990 like the social media, SMS, and emails etc. Considering Tunisia as a multilingual country, code-switch and mixing Tunisian with other languages as French, English and Modern Standard Arabic in daily speech is a common thing for the Tunisian people (Daoud, 2007). Tunisian dialect contains many varieties differing from a region to another.

Emotions does not feel the Same across different Cultures, they can differ across cultures through our use of language to understand and express our emotions. Tunisian emotional states are quite different to the other non Tunisian speakers, being angry as a Tunisian is not the same as a German or as an Algerian.

3.2 TuniSER Dataset

In order to build the first Tunisian Speech Emotion Recognition dataset (TuniSER), we divide emotions into two categories:

- Primary emotions: These emotions occur the most and they are the most used for the emotion recognition task. These emotions are: Happy, Sad, Angry, Neutral.
- Secondary emotions: Due to the rarity of these emotions and for long term purposes, we chose to keep them for annotation, the emotions are: Fear, Disgust and Surprised.

The first step is to choose the data sources. We focus on Tunisian series and TV programs publicly available online. We are seeking suitable, useful audios that are rich of emotions. Due to the non effectiveness of natural databases on emotion speech recognition, our dataset is then semi-natural (Induced). We select different audios for actors playing their roles in a sequence and manually extract the ones that contain explicit emotional states. We focus on the quality and diversity of the sources. Audios have to be quite clear and contain both male and females, different ages, situations and contexts.

Once audio sequences containing emotions are collected, we convert them to the required format. To take out the best emotional scenes from the audio, we choose to apply segmentation which separates the input audio sequence into small utterances of a range from 0.41 to 15.31 seconds with an average around 1 seconds (figure 1), with a clear expression (the whole word or sentence) that contains a significant emotional state without ignoring parts of the words and without overlap between the speakers.

We do a manual validation step before the annotation process. We only keep audios based on the following criteria:

- Each utterance should contain only one subject talking and contain only one scene.
- Sound should be clear without noise so that you can hear the speaker clearly talking.
- utterances should not contain music, silence or be with a bad quality.

[3] $https://sail.usc.edu/VAM/vam_release.htm$

The annotation process is divided on three Tunisian native speakers. Two female annotators are at a higher education level (Master/PhD), aged 23, and one female, aged 27, working as research Engineer. Labels are divided into the two emotions categories (Primary and Secondary emotions), according this guideline:

- Happiness: an upbeat, pleasant way of speaking/laughing.
- Sadness: Crying/Dampened mood/lack of energy and enthusiasm.
- Anger: such as speaking harshly or yelling.
- Neutral: lack of emotional state/ nothing in particular.

If an annotator finds an utterance that have another emotional state, it can be annotated as one of these secondary emotions: Fear (Such as rapid breathing and trembling voice.), Disgust (Voice expression that shows disgust.) and Surprise (Such as yelling, screaming or gasping).

We also mark the gender of the speaker in a segment (Male or Female) for long term purposes.

In Table 1, we present the number of clips per label.

Emotion	Female	Male	Total
Happy	153	172	325
Angry	230	387	617
Sad	247	127	374
Neutral	426	896	1322
Fear	27	7	34
Surprised	46	42	88
Disgust	2	2	4
Total	1136	1635	2771

Table 1: Statistics of the Tunisian SER dataset.

The obtained dataset is composed of 2771 utterances, with 1136 female utterances and 1635 male utterances. Our data carries semi-natural emotions that contain contextual and situational information. Every utterance holds a single emotion.

Using an unbalanced dataset will generate an over-classification problem for the larger classes. To solve this, we manually balance the data by extracting the same number of samples for each class. Data preparation was done to facilitate the use of our dataset to train the models. Finally, we obtain a final balanced dataset with 1300 samples

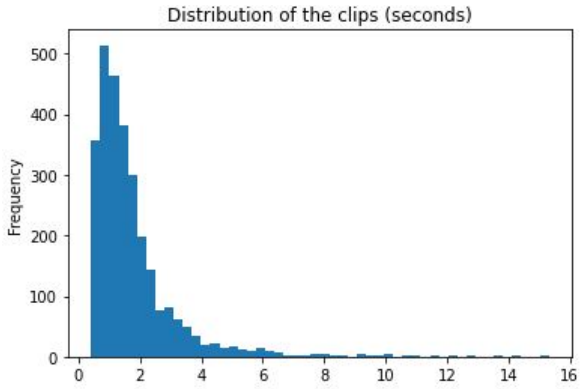

Figure 1: Distribution of the clips length

distributed equally between the four main classes (happy, angry, sad and neutral).

4 Training models

In this section, we will present the architectures of the models used to build our Tunisian SER. In fact, we explored different Machine Learning and Deep learning models for both features extraction and classification steps.

First, we wanted to investigate the influence of languages in SER. We trained using an SVM architecture a SER model on the RAVDESS dataset (Collet et al., 1997) which is an English SER dataset. It includes 1440 audio files of 24 speakers which includes half of male and half female actors and each of them has 60 recordings. The speech is in North-American accent. Then, we used the model to test our in-house Tunisian dialectal dataset. Due to the unsatisfactory results, we deduce that the interpretation of the sense of a word and the emotional states change from a language to another and this is due to multiple factors, such as culture.

Second, we experiment the Long short term memory model (LSTM) with different techniques of feature extraction (Chroma, Mel spectrogram, and Mel-frequency ceptral coefficients). We wanted to investigate which feature works best for the Tunisian SER. We noticed that even if the combination of the 3 features give good results, using only MFCC gives us the best results.

Finally, we introduced two large pre-trained models (The VGGish and wav2vec 2.0 models) as feature extractors followed by classifiers to predict emotions.

4.1 VGGish Model

The google VGGish is a variant of the VGG model (Sahoo et al., 2019) and they have a very similar

architecture. The VGGish model was pre-trained on the AudioSet (Gemmeke et al., 2017) database, which is a collection of more than 2 Million human labeled audio clips collected from Youtube videos with 10 second length each. This database contains over 600 sound classes: Music, Speech, Vehicle, etc. Before feeding the audio clip to the VGGish model, the following steps are applied:

- All audios are resampled to 16 kHz.
- Spectrogram is extracted using magnitudes of the Short-Time Fourier Transform, with three windows: a window size of 25ms, a window hop of 10ms and a periodic Hann window.
- By mapping the spectrogram to 64 mel bins covering the range 125-7500 Hz, a mel spectrogram is computed.
- A stabilized log mel spectrogram is obtained by applying log using the offset to avoid taking a logarithm of zero.
- These features are then framed into non-overlapping examples of 0.96 seconds length, where each example covers 64 mel bands and 96 frames of 10ms each.

Speech segments that are longer than 0.25 carry enough information about the emotional state of the speaker. Therefore, detecting emotion from consecutive segments of the same audio clip will be more efficient (Sahoo et al., 2019). So, we apply overlapping segmentation to catch better correlation between the segments of the same clip and at the same time it is considered as a data augmentation because it increases the number of data points. We extract one-second duration of overlapping segments and fill the last segment with silence to make it one second long. For all segments the overlapping duration is 0.5. This is the first component of our model representing the feature extractor, in which it takes a 96×64 dimensional mel spectrogram as input. VGGish network architecture is constructed of four blocks, two convolution and max-pooling layers followed by 2 fully connected (FC) layers of 4096 units each and finally a FC layer of 128 units that gives the embedding vector. VGGish gives a high-level 128-D embedding from audio input features. Those embedding could be fed to the downstream classification model as input. The VGGish embedding is semantically meaningful and compact than classic and raw audio features

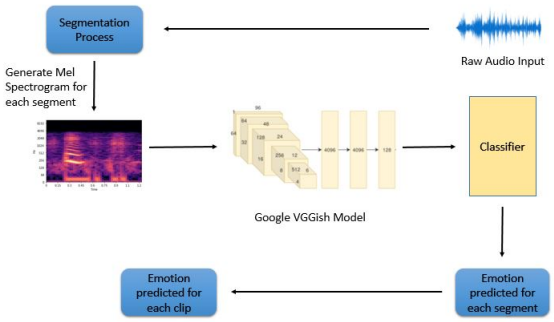

Figure 2: Visual Representation of the VGGish-based model.

so they allow downstream models to be shallower than usual.

For the classification, we applied different approaches based on feeding the 128-dimensional embedding vector to n fully connected layers with N hidden units, with n=1,2 and N =[100,200,400]. Finally, the logit layer is used to predict the emotion for each segment of the same utterance. The full architecture is presented in Figure 2.

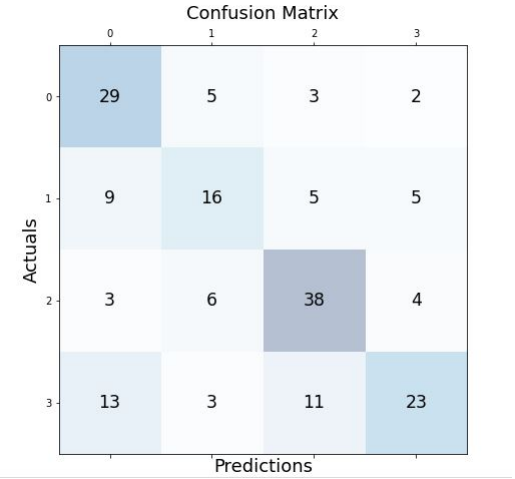

Figure 3: Confusion Matrix with 1 FC layers and number of units=400

4.2 Multilingual wav2vec 2.0

Wav2vec 2.0 (Baevski et al., 2020) is a framework for self-supervised learning of representations from raw audio. It has 2 stages: pre-training and fine-tuning. In pre-training, the speech input is masked in the latent space and a contrastive task with predictions from the transformer and quantized latent speech representations is solved to learn contextualized information. This enables learning powerful representations from speech audio alone. The architecture of wav2vec 2.0 represents three stages,

a local encoder, which contains several convolutional blocks, a contextualized encoder, and a quantization module. To pre-train the acoustic model, we use a multi-layer convolutional feature encoder which takes raw audio as input and outputs latent speech representations. They are then fed to a Transformer to build representations capturing information from the entire sequence. The output of the feature encoder is discretized with a quantization module to represent the targets in the self-supervised objective. The model builds context representations over continuous speech representations and self-attention captures dependencies over the entire sequence of latent representations. We masked a certain proportion of time steps in the latent feature encoder space similar to masked language modeling in BERT (Devlin et al., 2019).

Pre-trained models are then fine-tuned for downstream tasks like Automatic Speech Recognition by adding a classifier with C classes representing the output vocabulary of the respective downstream task on top of the model and training on the labeled data with a Connectionist Temporal Classification (CTC) loss (Graves et al., 2006).

To train our SER model, we used fine-tuned models on both MSA and Tunisian dialect Automatic Speech Recognition downstream task to better determine the context representations for the input audios, since ASR datasets are more available and larger than the SER ones. Finally, we built a neural network as classifier (Table 2) on top of it to predict the class of each audio. The workflow is presented in Figure 4.

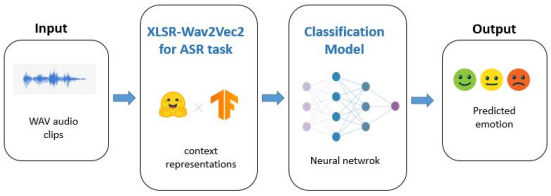

Figure 4: Multilingual Wav2vec 2.0-based model workflow.

5 Experimental Setup and Results

For the VGGish model, we start training our model with the following configuration: 16 as batch size, Adam optimizer with epsilon equal to 10^{-8} and a Learning rate of 10^{-6}. We start our training by applying the segmentation process after splitting our data into 70% train set, 15% validation set and 15% test set. We trained the model for 261 epochs, with three different numbers of hidden units N=[100,200,400] and different numbers of FC layers [1,2]. The best result is obtained with 400 numbers of hidden units and 1 Fully connected layer.

Architecture
Dropout
Single layer feed forward network
Tanh
Dropout
Single layer feed forward network

Table 2: Neural Network classifier architecture.

For the second approach, we used the multilingual wav2vec 2.0 model. As a first step, we fine-tuned it for the Automatic Speech Recognition downstream task on two languages: Modern Standard Arabic (MSA) and Tunisian dialect. For MSA, the model is trained using the Mozilla Common Voice dataset. For the Tunisian dialect, we used the STAC (Zribi et al., 2015) which is a small Tunisian ASR dataset. Table 3 presents the results of fine-tuning wav2vec on the ASR downstream task on both MSA and Tunisian dialect datasets.

Language	**Dataset**	**WER (%)**
MSA	Common voice	52.53
Tunisian Dialect	STAC	62

Table 3: Results of fine-tuning multilingual wav2vec 2.0 on the Automatic Speech Recognition downstream task with MSA and Tunisian Dialect data.

Finally, we built a classification layer on top of the two fine-tuned models to perform the emotions classification. For both models, we adjust the number of epochs and the value of learning rate intuitively to find the best results. We split our balanced data into 80% as training and 20% validation set. 1035 samples for the training and 256 samples for the validation.

6 Discussion

Using the VGGish model as a feature extractor followed by a classifier trained on our constructed Tunisian dialect dataset, we achieved 58,2% as frame accuracy and 60.05% as Average logits clip accuracy, with 1 Fully Connected layer and number

Model	Acc. (%)
LSTM	52.63
VGGish	60.05
multilingual wav2vec MSA	52.10
multilingual wav2vec TD	60.60

Table 4: Comparing different models.

of hidden units = 400. We noticed that changing the number of units gives closer results. The training loss decreased continuously over epochs and the validation loss decreased until the epoch number 151 and started to increase so an early stopping was applied to avoid over-fitting and keep the best result of the model. The confusion matrix, Figure 3, represents the performance of the model for each label of the test dataset. We noticed that adding two Fully Connected layers instead of one layer decreased the results from 58,2% to 54% for the frame accuracy and from 60.05% to 53.1% for the Average logits clip accuracy.

For the second approach, fine-tuning the multilingual wav2vec 2.0 model with MSA and Tunisian dialect gives satisfactory results. In the first try, intuitively running the model with a different number of epochs and learning rate gives 52,1% accuracy with MSA language and 60,6% accuracy with the Tunisian dialect. The training and validation losses for both attempts were decreasing with remarkable fluctuations. These are explained by the fact of using a large neural network with a lot of parameters with small datasets such as our case. This could be solved by increasing the batch size or reducing the parameters of the model.

7 Conclusion and future work

In this paper, we described our methodology to build a Tunisian Speech Emotion Recognition dataset. We detailed the process of using LSTM, and two large pre-trained models: the VGGish and the multilingual wav2vec 2.0 as feature extractors for the Speech Emotion Recognition task. We explained the implementation part for each approach and the different steps followed to train the models. The best result was obtained by fine-tuning the multilingual wav2vec 2.0, which reaches a WER of 60.6%. Our work is an important step for the SER task on the Tunisian dialect, since our satisfactory results could be improved in a future work by augmenting the datasets size and applying enhancement techniques.

References

Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. 2021. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4):1249.

A Aouf. 2019. Basic arabic vocal emotions dataset (baved)–github.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Christian Collet, Evelyne Vernet-Maury, Georges Delhomme, and André Dittmar. 1997. Autonomic nervous system response patterns specificity to basic emotions. *Journal of the autonomic nervous system*, 62(1-2):45–57.

Mohamed Daoud. 2007. The language situation in tunisia. In *Language Planning and Policy in Africa, Vol. 2*, pages 256–307. Multilingual Matters.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Chayma Fourati, Abir Messaoudi, and Hatem Haddad. 2020. Tunizi: a tunisian arabizi sentiment analysis dataset. *arXiv preprint arXiv:2004.14303*.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE.

Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2016. Representation learning for speech emotion recognition. In *Interspeech*, pages 3603–3607.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks. volume 2006, pages 369–376.

Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the international speech communication association*.

Yasser Hifny and Ahmed Ali. 2019. Efficient arabic emotion recognition using deep neural networks.

In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6710–6714. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *CoRR*, abs/2106.07447.

Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2011. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9-10):1162–1171.

Jinkyu Lee and Ivan Tashev. 2015. High-level feature representation using recurrent neural network for speech emotion recognition. In *Sixteenth annual conference of the international speech communication association*.

Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, and Guan-Zheng Tan. 2018. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, 273:271–280.

Steven R Livingstone and Frank A Russo. 2018. 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.

Mohamed Meddeb, K Hichem, and A Alimi. 2016. Automated extraction of features from arabic emotional speech corpus. *International Journal of Computer Information Systems and Industrial Management Applications*, 8:184–194.

Mohamed Meddeb, Hichem Karray, and Adel M Alimi. 2014. Intelligent remote control for tv program based on emotion in arabic speech. *arXiv preprint arXiv:1404.5248*.

Omar Mohamed and Salah A Aly. 2021. Arabic speech emotion recognition employing wav2vec2. 0 and hubert based on baved dataset. *arXiv preprint arXiv:2110.04425*.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

Sourav Sahoo, Puneet Kumar, Balasubramanian Raman, and Partha Pratim Roy. 2019. A segment level approach to speech emotion recognition using transfer learning. In *Asian Conference on Pattern Recognition*, pages 435–448. Springer.

Gaurav Sahu. 2019. Multimodal speech emotion recognition and ambiguity resolution. *arXiv preprint arXiv:1904.06022*.

L. Sayahi. 2014. *Diglossia and Language Contact: Language Variation and Change in North Africa*. Cambridge Approaches to Language Contact. Cambridge University Press.

Björn Schuller, Gerhard Rigoll, and Manfred Lang. 2003. Hidden markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 2, pages II–1. Ieee.

Thapanee Seehapoch and Sartra Wongthanavasu. 2013. Speech emotion recognition using support vector machines. In *2013 5th international conference on Knowledge and smart technology (KST)*, pages 86–91. IEEE.

Akash Shaw, Rohan Kumar Vardhan, and Siddharth Saxena. 2016. Emotion recognition and classification in speech using artificial neural networks. *International Journal of Computer Applications*, 145(8):5–9.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE.

Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. 2017. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590.

WQ Zheng, JS Yu, and YX Zou. 2015. An experimental study of speech emotion recognition based on deep convolutional neural networks. In *2015 international conference on affective computing and intelligent interaction (ACII)*, pages 827–831. IEEE.

Inès Zribi, Mariem Ellouze, Lamia Hadrich Belguith, and Philippe Blache. 2015. Spoken tunisian arabic corpus "stac": transcription and annotation. *Research in computing science*, 90:123–135.