Natural Arabic Language Resources for Emotion Recognition in Algerian Dialect

Habiba Dahmani[1](✉), Hussein Hussein[2], Burkhard Meyer-Sickendiek[2], and Oliver Jokisch[3]

[1] Department of Electrical Engineering,
University of Mohamed Boudiaf, M'Sila, Algeria
habiba.dahmani@univ-msila.dz

[2] Department of Literary Studies, Free University of Berlin, Berlin, Germany
{hussein,bumesi}@zedat.fu-berlin.de

[3] Institute of Communications Engineering,
Leipzig University of Telecommunications (HfTL), Leipzig, Germany
jokisch@hft-leipzig.de

Abstract. In this paper, we present and describe our first work to design and build a natural Arabic visual-audio database for the computational processing of emotions and affect in speech and language which will be made available to the research community. It is high time to have spontaneous data representative of the Modern Standard Arabic (MSA) and its dialects. The database consists of audio-visual recordings of some Arabic TV talk shows. Our choice comes down on the different dialects with the MSA. As a first step, we present a sample data of Algerian dialect. It contains two hours of audio-visual recordings of the Algerian TV talk show "Red line". The data consists of 14 speakers with 1, 443 utterances which are complete sentences. 15 emotions investigated with five that are dominants: enthusiasm, admiration, disapproval, neutral, and joy. The emotion corpus serves in classification experiments using a variety of acoustic features extracted by *openSMILE*. Some algorithms of classification are implemented with the *WEKA* toolkit. Low-level audio features and the corresponding delta features are utilized. Statistical functionals are applied to each of the features and delta features. The best classification results - measured by a weighted average of f-measure - is 0.48 for the five emotions.

Keywords: Emotion recognition · Arabic language resources · Algerian dialect

1 Introduction

Emotions are omnipresent whether we speak or not. It is a continuous state of mind. Emotions are the mind of our verbal and non-verbal reactions. Without emotions, our reactions are incomprehensible. Without even partial or minimal presence of emotions, we are somehow sick or less intelligent as we should

K. Smaïli (Ed.): ICALP 2019, CCIS 1108, pp. 18–33, 2019.
https://doi.org/10.1007/978-3-030-32959-4_2

be normally. Therefore, the importance of emotions in our daily lives is self-evident. Studying emotions to improve the quality of our interactions between humans or between humans and machines is essential.

The progress in automatic processing includes the understanding of natural language. This make emotions detection and classification within the speech signal possible. Speech includes all linguistic features such as the phonological, lexical, semantic information, prosodic information. The information expressed through speech can be divided into three categories: linguistic (such as accent, phrase and sentence type), paralinguistic (for example: intention, attitude and speaking style), and nonlinguistic information (such as age, gender, physical and emotional states of speakers). The acoustic parameters correlating with prosodic properties are fundamental frequency, duration, and intensity. Nonlinguistic information is concerned with information about age, gender, physical and emotional states of speakers [19]. In human interactions, the listener interprets and responds to the emotive state of the speaker and adjusts the reaction depending on the emotions that the speaker communicates. Consequently, it is extensively argued that artificial intelligence needs to recognize human emotions and understand them in order to achieve natural Human-Machine communication. Recognizing human emotions mainly from the speech signal is a challenging process for many reasons. In general, studies mention two principal difficulties: audio-video databases and recognition algorithms [45].

Indeed developing a database is a requirement for building emotion recognition. Emotional speech recognition, based on recorded and annotated databases, has received much attention from many researchers [10,11,13,24,25,31,33,47,49,52]. However, there is a growing need for real-time and offline emotion recognition systems that are based on an analysis of the speech signal or both visual and speech signal. A significant emotion recognition obstacle is the quality of the recorded speech samples; more specifically: the quality, the size and the type of the database. The speech corpora are either acted, induced or natural. Acted or simulated corpora are collected from professional television or radio actors who are widely used in research work. Natural databases can be used for real-world emotion modelling. These databases are created by collecting the real world conversations such as Call center conversation, a conversation between patient and doctors, or TV show programs, etc. [12,20].

Arabic language and its dialects are still considered a relatively resource-poor language when compared to other languages such as English [53]. In its dialectal forms, Arabic is the mother tongue of more than 250 million speakers. The Arabic language has three forms: classical Arabic or literary Arabic language, Modern Standard Arabic (MSA), and Colloquial Arabic. Classical Arabic is essentially the form of the language found in the Quran, the MSA is a modern form of Arabic used in news media and formal speech, and by definition dialects are spoken. These Dialects have no written standards.

In recent years, concerning Arabic affective computing, there has been a considerable amount of works on the collection of emotional speech. However, most databases built up to now are induced, small, and just constructed for particular purposes. The corpora are mostly recorded by unprofessional actors, which affect the quality of the generated speech. Another problem is the lack of enough recorded multimodal data of emotions. Most of the developed databases are not available for public use. Thus, the ultimate consequence of this domain is the absence of coordination and collaboration among researchers in this field. Hence the need to think and proceed to build a large database for this language and its dialects.

This paper reports progress in constructing an Arabic natural database that presents speech in the context of natural interactions where emotion is conveyed via multiple modalities. This database will be available online for sharing with the scientific community. Algerian dialect data is used as a sample data segmented and annotated to 15 emotions. There are five dominant emotions that are: enthusiasm, admiration, disapproval neutral, and joy. We classify these emotions with several machine learning algorithms to experiment variety of acoustic features.

The remainder of this paper is organized as follows: The state-of-the-art for emotion recognition in Arabic is presented in Sect. 2. Section 3 provides an overview about the database (spontaneous Algerian data) as well as collection and annotation steps. Section 4 illustrates the experiment by feature extraction and classification of emotions. The results are described in Sect. 5 and finally, conclusions and future works are presented in Sect. 6.

2 Emotion Recognition in Arabic

The need and the motivation for such a work for the Arabic natural language processing (NLP) are imperative. Then, our primary goal is to show the scarcity of the Arabic linguistic resources and to confirm too that little work has been devoted for the analysis of emotional speech in Arabic by presenting a state-of-the-art overview of Arabic studies that have been carried out in this domain.

Going through the literature of Arabic speech emotion processing, we find some studies and a few more or less significant emotional databases. Between 2005 and 2006, we find three Syrian works about the introduction of the emotion parameters for Arabic text-to-speech synthesis [2,4]. Al-Dakkak et al. tried to improve the Arabic synthetic speech (MSA Arabic) in order to sound as natural. They incorporated different prosodic features with five emotions: anger, joy, sadness, fear, and surprise in an educational Arabic text-to-speech system. The authors elected three sentences for each emotion. Each sentence is recorded twice, one emotionless and the other with the intended emotion. They did not specify the number of speakers or any further details about their database. To generate prosody automatically, they considered the most crucial acoustic parameters: pitch, duration, and intensity. In 2011, Khalil reported in [27,28] that he constructed a well-annotated corpus for anger and neutral emotion states from real-world Arabic speech dialogues for his experiments. It consists of a set of recorded episodes of a live Arabic political debate show: The Opposite Direction program of Al-Jazeera Satellite Channel. The number of samples extracted from the selected episodes are more than 400 samples for anger emotion state, varied from one second to 9 s with different speaker gender in different Arabic cultures. The second source of his data is an angry customer's call: This is a phone call published on YouTube that features a customer who was very angry while speaking with the agent. The call duration was four minutes long. He was able to extract 45 samples of data which reflect anger states from this clip. He used prosodic and spectral features and several classifiers are evaluated: Support Vector Machine (SVM), Probabilistic Neural Networks (PNN), simple decision tree, and decision tree forest.

Azmy et al. in [6,7] built an Arabic unit selection voice that could carry emotional information. Three emotional states are covered: normal, sad, and questions. They used the text-to-speech from RDI 'the Engineering Company for the Development of Digital Systems' (RDITTS) for Saudi speaker database. This database consists of 10 h of recording with neutral emotion and one hour of recordings with four different emotions that are: sadness, happiness, surprise, and enquiring. However, they did not give any reference or either further details about this database. The authors used an automatic emotion classification system: Emovoice. The system comes with a predefined two classification models; probabilistic Naïve Bayesian (NB) and SVM classifiers [51]. Meddeb et al. in [34] propose the architecture of Automatic Emotion Recognition from Speech in order to recommend a system for TV programs based on human behavior. The emotional recognition for the remote control includes unimodal and multimodal approach and shows the hierarchical recognition steps of emotions. They used the six following databases: two publicly available ones, the Danish Emotional Speech corpus (DES) and Berlin Emotional Database (EMO-DB), and four databases from the interface project with Spanish, Slovenian, French, and English emotional speech. In 2014, they created their own Tunisian dialect sound database for automatic emotion speech recognition always to achieve intelligent remote control. They called it the Tunisian Emotional Speech database (TUES). The database composed primarily of sound passages and isolated word recorded by actors where the ages, sex, and region are different. The sentences are designed to use for recording the seven emotions: neutral, anger, surprise, disgust, fear, happiness, and sadness. This database contains 720 speech samples. The length of speech samples is up to five seconds [36]. Hammami reported in [22] that his search resulted in finding a single emotional speech database called Emotional speech database of Research Groups on Intelligent Machines (REGIM_TES). The database is from the National Engineering School of Sfax in Tunisia and developed by Meddeb et al. in [35,37–39]. Due to undisclosed reasons, the database has been made private. The REGIM_TES database is composed primarily of isolated words and very short, semantically neutral phrases made of few collocate words not exceeding four. The research used 12 actors, six of each gender to generate five emotion categories that include: anger, fear, happiness, sadness, and neutral. The selected descriptors in the study are the pitch of voice, energy, Mel Frequency Cepstral Coefficients MFCCs, Formant, Linear Predictive Coding (LPC) and the spectrogram. They experienced a different type of classifiers and they opted for the SVM multiclass classifier.

Meftah et al. [40] describe an emotional speech corpus recorded for MSA (KSUEmotions: the King Saud University Emotions). The KSUEmotions corpus recordings contain five hours of emotional MSA speech for five emotions: neutral, sadness, happiness, surprised, and angry as well as 16 sentences, and 20 speakers from three countries: Saudi Arabia, Yemen, and Syria. From Damascus University, Al-Faham and Ghneim [5] built an Arabic emotional speech corpus, covering five emotions: happiness, anger, sadness, surprise, and neutrality to recognize the Arabic user's emotional state by analyzing the speech signal. The speech data used in the experiment contains 24 emotional Arabic sentences recorded by six performers (3 male and three female) and every sentence recorded twice for every emotion. The classification results in these works are carried out using the *WEKA* software [21] and by using neural network classifier based on Multilayer Perceptron (MLP) with rhythm metrics as new descriptors.

Another research about the spontaneous emotional Arabic database is from Klaylat's et al. in [29,30]. They confirmed that no natural emotional Arabic corpus was found to date. A realistic speech corpus from Arabic TV shows is collected. The videos are labeled by their perceived emotions, i.e. happy, angry, or surprised. The corpus composed of 1,384 records with 505 happy, 137 surprised and 741 angry units. The unit is one second of speech. Low-Level Descriptors (LLDs) are extracted using the open source *openSMILE* feature extractor [14] that is developed at the Technische Universität München (TUM). Thirty-five classification models are applied to the Sequential Minimal Optimization (SMO) classifier. Finally, Abdo et al. [1] built an MSA audio-visual corpus. The corpus is annotated both phonetically and visually and dedicated to emotional speech processing studies. 500 sentences are critically selected based on their phonemic distribution with six emotions (happiness, sadness, fear, anger, inquiry, and neutral). The recorded audio-visual corpus is contributing to the field of speech processing specifically Text-to-Speech (TTS) applications. The authors intend to let the corpus publically available for general research purposes.

It becomes clear from this short overview of the Arabic emotion processing studies and achievements that textcolorredthe Arabic language and its dialects initially require the construction of important databases to be shared freely online for the scientific community. The availability of these resources will strengthen and encourage useful research in the automatic processing of the Arabic language in general and coordination between researchers in this field. This overview is condensed in Table 1. The Table gives the most important databases for the recognition of emotions. It is ordered by publication date.

3 Database

The collection of new emotional speech databases that tries to overcome the limitations of the existing corpora is a crucial necessity. These efforts lead to widespread knowledge of language technology.

Table 1. List of Arabic speech emotion databases.

Name	Type	MSA or dialect	Speakers	Linguistic material (Nr. of sentences)	Emotions	References
KSUEmotions	Simulated	MSA	20	16	5 emotions: neutral, sadness, happiness, surprised, questioning	[40, 41]
Egypt	Acted	MSA	07	500	6 emotions: happiness, sadness, fear, anger, inquiry, neutral	[1]
REGIM_TES	Acted	Tunisian	-	-	-	[39]
Arabic-Natural-Audio-Dataset	Natural	Dialectal (Egyptian, Jordan, Gulf, Lebanese)	06	-	3 emotions: happiness, anger, surprised	[29, 30]

3.1 Data Acquisition

The undergoing database consists of a collected data of spontaneous emotional speech in Arabic language, including MSA and colloquial Algerian, Tunisian, Lebanese, Jordanian, Syrian, and Egyptian which will be made available to the research community. The database consists of audio-visual recordings of some Arabic TV talk shows, segmented into broadcasts. The corpus contains spontaneous and very emotional speech recorded from discussions between the guests of the talk shows. We decided to select some TV talk shows for this data collection because the spontaneous discussions between the talk show guests are often somewhat affective. Such interpersonal communication leads to a wide variety of emotional states, depending on the topics discussed. These topics were mainly personal issues such as friendship crises, questions about paternity or romantic affairs: emotionally intense and discussing hot social phenomena. So far, We collected about 50 broadcasts of the talk show for Arabic (including MSA, and Algerian, Tunisian, Lebanese, Jordanian, Syrian, and Egyptian dialects). In the present work, only the Algerian data is processed. The number of programs and dialects can be expanded further to balance the database concerning the number and gender of speakers and also the categories of emotions.

3.2 Case Study: Algerian Emotional Speech

"Red Line", a weekly social program on Al-Shorouk Algerian channel. A talk show where guests are invited to talk. The show is presented by three permanent hosts; anchorwomen who appear on each program, introduces and interacts with the guests. Religious and psychological opinions are present. Social program sometimes deals with sensitive subjects that are difficult to discuss. The program begins in a studio on a red background, with shapes and lines that reflect the orientation of the program [50].

Both MSA and dialects are used for communication. Dialectal Arabic is a term that covers Arabic dialects, resulting from linguistic interference between the Arabic language and local or neighboring languages as a result of the process of Arabization or any cultural influence due mainly to colonization, migration, trade, and more recently the media. Algerian dialect is generally described as an Arabic idiom attached to the Maghrebian Arabic group (Algerian, Moroccan, Tunisian and Libyan). However, its morphology, syntax, pronunciation, and vocabulary are quite different from other Arabic dialects. Algerian Arabic is established on a substrate that was initially Berber, Latin (African Romance Language), and to a lesser extent Punic. It has also been enriched by the languages of the powers that influenced this region including Ottoman Turkish, Spanish, and French. This dialect is characterized by the multitude including sub-dialects that are clearly variants of Arabic, and other sub-dialects that are non-Arabic which we call the Amazigh dialect. Therefore, the Algerian and the other North African dialects are considered a little distant from the MSA. While the Arabic dialects of the East: Egypt, Sudan, Levantine, Gulf countries are dialects closer to the MSA. We focus in this study on the Algerian dialect which is classified by the Algerians themselves as a mixture of three languages: Arabic, Berber, and French.

3.3 Segmentation and Emotional Labelling

The corpus contains spontaneous and very emotional speech recorded from unscripted, authentic discussions between the guests of the talk show. These first records consist of two hours which are segmented in the first step into smaller units or clips containing the whole dialogue between a limited number of talk show guests. In general, such discussions are extracted as videos containing the audio and video signals. In a second step, the dialogues are segmented into turns: the turn is when the one partner finishes speaking, and the other partner takes over. The third step is to segment these turns to utterances. These utterances were mainly complete sentences, but sometimes also grammatically incomplete sentences which were due to the spontaneous nature of the interactions. Many utterances had to be discarded because of background music, applause from the audience, and overlaps between speakers or other interruptions. The audio signal was stored separately for each sentence. An identification is given to each speaker and each utterance. We have four categories of speakers and each category is designed by four characters. The anchor women appointed by Mode (moderator), for the other speakers, we used these letters G, H, P, S to indicate respectively, Guest, Host, Principal, Speaker. GH is for the permanent speakers with the moderator in the talk show. GP is for the principal Guest in each episode. Guest Speaker (GS) is a guest among several or among the spectators of the show. Concerning the two digits: the first number concerns the episode number and the second number is for the speaker's order in the show (for example GS11 is the first guest speaker in episode one).

The collected data contains 1,443 utterances from 14 different speakers among them 5 females (two little girls and 3 adult females). But it should be noted that there are actually three dominant speakers (see Fig. 1): moderator of the show (Mode), the psychologist (GH01), and the religious expert (GH02).

The next step should be to define the emotions and their data segments or emotional analysis units. Indeed, the most crucial problem for the analysis of emotional data is to determine what an emotional utterance is, where it starts and where it ends [8]. The data were evaluated by three human listeners. Each listener assessed the emotional content for the whole of utterances in terms of the emotion categories. Categorical rating involved applying labels from a list of terms (a total of 28 emotions until now). However, only 15 emotions were detected in the collected database and rated for the present data: Anger, Joy, Happiness, Sadness, Disapproval, Admiration, Surprise, Enthusiasm, Adoration, Calm, Gratitude, Reproach, Neutral, Sympathy, and Satisfaction. The raters watch the videos and the corresponding audio files and then they try to follow the emotion and define exactly the audio part related to the emotion. However, we insist on the fact that our method will be significantly reduced for the rest of the data and we will use a higher number of raters to be able to complete the evaluation of the whole database. There are about five emotions that are more present than others (Enthusiasm, Admiration, Disapproval, Neutral, and Joy) as shown in Fig. 2.

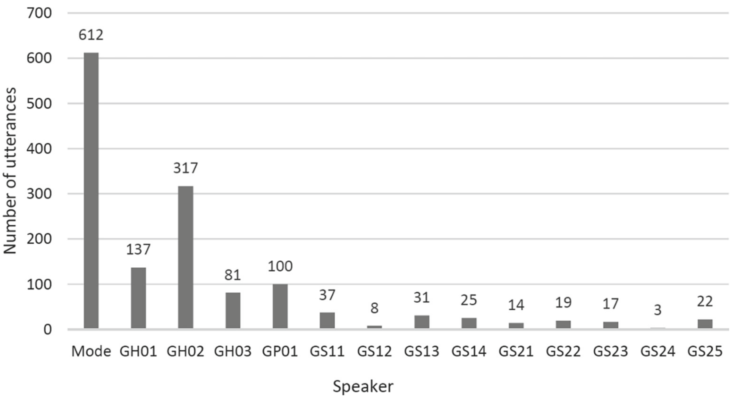

Fig. 1. The distribution of utterances per speaker.

3.4 Recording Quality

The video files are MPEG-coded image sequences of 352 × 288 pixels with a frame rate of 25 fps. A constant code rate of 1.15 Mbit/s was used. Recordings were taken with a sampling frequency of 48 kHz and later downsampled to 16 kHz (16 bit). These criteria are commonly used for speech databases.

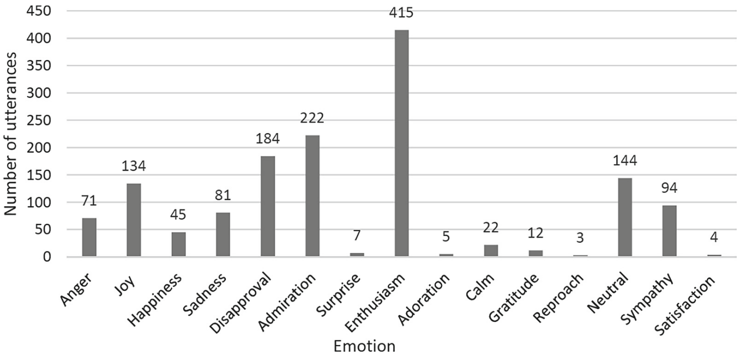

Fig. 2. The distribution of utterances per emotion.

4 Experiment

The performance of a variety of audio features and several classifiers is investigated to determine the best-suited features and classifiers for the classification of emotions.

4.1 Features

Feature extraction can be split into two categories according to the processing domain: the time-domain features and frequency-domain features. The time-domain features are Zero-Crossing Rate (ZCR) and short-time average energy. The frequency-domain features are pitch or fundamental frequency (F_0), spectral features (band energy, spectral roll-off, spectral flux, and spectral centroid), cepstral features (MFCCs), and linear prediction features (LPC).

We perform feature extraction for emotions by using the *openSMILE* feature extraction tool [14]. The feature set contains features which result from LLDs with the corresponding delta coefficients (ΔLLD) and statistical functionals applied to each of the LLD and ΔLLD. The main LLD used in the experiment are the features described above. The features are energy, pitch, ZCR, spectral features, MFCCs, and Line Spectral Frequencies (LSP) which are computed from LPC coefficients. The default values of coefficients in *openSMILE* are used (12 cepstral coefficients for MFCCs and 8 linear predictive coding coefficients for LPC). The following statistical functionals are used for every feature: min, max, range, standard deviation and mean.

Many feature vector sets are used:

* A (120 features): MFCCs
* B (50 features): Energy, pitch, ZCR
* C (230 features): Energy, pitch, ZCR, spectral features
* D (350 features): Energy, pitch, ZCR, spectral features, MFCCs
* E (430 features): Energy, pitch, ZCR, spectral features, MFCCs, LSP

* F (384 features): Baseline feature set of the Emotion Challenge in the Interspeech 2009 [46]
* G (6373 features): Baseline feature set of the Computational Paralinguistics Challenge (ComParE) in the Interspeech 2013 [44]
* H (6552 features): The large *openSMILE* emotion feature set with more functionals and more LLD [15].

We used the feature sets A-E in a previous experiment of acoustic event classification for the utilization in the sector of healthcare.

4.2 Classification

A series of classifiers are selected in order to determine the best-suited classifier for the evaluation. The following machine learning algorithms with default values using the *WEKA* data mining toolkit [21] are applied for the recognition of emotions:

* **IBk**: the Instance-Based (IB) classifier with a number of (k) neighbors is the K-Nearest Neighbours (KNN) classifier using the euclidean distance and 1-nearest neighbour [3].
* **AdaBoostM1**: the boosting algorithm uses the Adaboost M1 method [17].
* **LogitBoost**: The classifier performs additive logistic regression [18].
* **SimpleLogistic**: a classifier for building linear logistic regression models [32,48].
* **RandomTree**: Random trees is a collection of decision trees that considers K randomly chosen attributes at each node [16].
* **RandomForest**: The classifier of random forest consists of several uncorrelated decision trees [9].
* **SMO**: The Sequential Minimal Optimisation (SMO) for training a Support Vector Machines (SVM) classifier [23,26,42].
* **J48**: The J48 algorithm used to generate a pruned or unpruned decision tree [43].

We implement our experiment for the recognition of emotions only on the five emotions that are more present than others: Enthusiasm, Admiration, Disapproval, Neutral, and Joy.

5 Results

The performance is measured using the f-measure which is the harmonic mean between precision and recall. Table 2 shows the classification results by applying eight kinds of feature sets and eight classifiers. We used the extracted features and classifiers to classify the five emotions: Enthusiasm, Admiration, Disapproval, Neutral, and Joy. The Table shows that the SMO classifier yielded better results (0.48) than other classifiers for different feature sets. The best results according to the features are for the feature sets D and E (with the SMO classifier) (0.47 and 0.48, respectively). Increasing the number of feature engineering from feature set (B) to (E) leads to a slight improvement in the results.

Adding MFCCs to the feature set C yielded better improvement of results (see the columns C and D). The results by using baseline features (feature sets F and G) and large *openSMILE* emotion feature set (H) are in general not better than feature engineering sets (A till E).

Table 2. Experimental results (weighted average of f-measure) obtained with the 10 fold cross-validation by applying different feature sets on several classification algorithms.

	A	B	C	D	E	F	G	H
SimpleLogistic	0.42	0.33	0.39	0.46	0.46	0.42	0.42	0.42
SMO	0.42	0.20	0.39	0.47	0.48	0.40	0.42	0.44
IBk	0.42	0.33	0.36	0.42	0.42	0.41	0.34	0.46
AdaBoostM1	0.20	0.23	0.20	0.20	0.25	0.20	0.23	0.20
LogitBoost	0.38	0.35	0.39	0.44	0.44	0.40	0.42	0.40
J48	0.33	0.30	0.34	0.34	0.37	0.32	0.34	0.40
RandomForest	0.40	0.34	0.40	0.40	0.44	0.35	0.32	0.45
RandomTree	0.30	0.32	0.34	0.35	0.34	0.29	0.28	0.34

Analyzing the confusion matrix in the Table 3 for the five emotions by using the SMO classifier and the feature set E, we find that the most misclassifications are between "Admiration" and "Enthusiasm", "Disapproval" and "Enthusiasm", and "Joy" and "Enthusiasm". This means that the recognition of Enthusiasm between the among five emotions is the most difficult task. We think it is understandable because enthusiasm is a little bit of all these emotions. thanks to this example, we realized the importance of the annotation phase and also the use of a more precise method: dimensional labeling approach.

Table 3. The confusion matrix for the five emotions with the SMO classifier and the feature set E.

	Admiration	Disapproval	Enthusiasm	Joy	Neutral	Sum
Admiration	**96**	17	83	12	14	222
Disapproval	14	**99**	60	2	9	184
Enthusiasm	82	44	**219**	38	32	415
Joy	7	8	50	**64**	5	134
Neutral	21	19	49	3	**52**	144

6 Conclusion and Future Works

In this paper, we presented the first steps of our work whose objective is to design and build a new Natural Arabic multimodal spontaneous emotion database for the research community in order to facilitate the research in the field. We have tried to highlight some points concerning the importance of the Arabic language and the fact that, there is insufficient interest in this language regarding the recognition of emotions or regarding all disciplines of artificial intelligence in overall. Arabic is one of the oldest languages in the world. It is one of the first widely used languages nowadays. So it is really vital to give a lot of importance to further study this language and its variants which are its dialects. Basically, high quality and large speech corpora are required for the emotion recognition task. However, the existing MSA and dialectal Arabic speech corpora are very sparse and are of low quality. For some Arabic dialects, speech resources do not exist at all. That is why it is difficult to initiate research and studies in this field. With our research, we mainly pursued two goals. The first one is to collect a high-quality MSA and dialectal speech corpus. The second goal is to rapidly develop phonetic transcriptions for dialectal speech data. We have started with an overview of the Arabic language from an emotion recognition point of view. This overview shows that it is not easy to access extensive and up-to-date freely available Arabic corpora. it should be noted that the use of corpora has been a major factor in the recent advance in artificial intelligence in general and in the natural language processing development and evaluation particularly.

We measured the performance of emotion classification by using a variety of audio features and several classifiers for five emotions in the Algerian dialect.

However, our current work has certain limitations, which give rise to our future work as follows: (1) Considering the Arabic language and its dialects (besides the different varieties of the natural language in its self) is a problem for speech recognition before being a problem for the recognition of emotions. Simply, these varieties cannot be modelled in an appropriate way. Firstly, for the transliteration of what has been said, we propose to use Romanization method for transcription of the speech corpus. (2) Continuing to perform different annotations like prosodic, Part-of-Speech tags, and syntactic labels and segmentation in word and chunk levels. (3) We will investigate both methods: the dimensional and the category labels for the emotional annotation and use the two classic criteria for assessing the quality of such labels: validity and reliability. (4) For naturalistic data, both acoustic and linguistic features should be employed, both for a deeper understanding and a better classification performance. We will establish different measures of impact and discuss the mutual influence of acoustics and linguistics. (5) The classification will be experimented using different levels, word, chunk, and utterances. Classification performance relies on deep learning and pattern recognition techniques. One defies to address in emotion classification is how to prune into this depth of methods and find a good one for this specific task. And of course, we have to deal with the curse of dimensionality and class skewness or the sparse data problem in the output space. As a result, the multimodal spontaneous corpus will be designed for general analysis of human behavior of emotional as well as for automatic emotion classification purposes.

References

1. Abdo, O., Abdou, S.M., Fashal, M.: Building audio-visual phonetically annotated Arabic corpus for expressive text to speech. In: INTERSPEECH, pp. 3767–3771 (2017)
2. Abou Zliekha, M., Al Moubayed, S., Al Dakkak, O., Ghneim, N.: Emotional audio-visual Arabic text to speech. In: The XIV European Signal Processing Conference (EUSIPCO) (2006)
3. Aha, D., Kibler, D., Albert, M.: Instance-based learning algorithms. Mach. Learn. **6**, 37–66 (1991)
4. Al-Dakkak, O., Ghneim, N., Zliekha, M.A., Al-Moubayed, S.: Emotion inclusion in an Arabic text-to-speech. In: 2005 13th European Signal Processing Conference, pp. 1–4. IEEE (2005)
5. Al-Faham, A., Ghneim, N.: Towards enhanced arabic speech emotion recognition: comparison between three methodologies. Asian J. Sci. Technol. **07**(03), 2665–2669 (2016)
6. Azmy, W.M., Abdou, S., Shoman, M.: The creation of emotional effects for an Arabic speech synthesis system. In: The Egyptian Society of Language Engineering, International Workshop, ESOLE (2013)
7. Azmy, W.M., Abdou, S., Shoman, M.: Arabic unit selection emotional speech synthesis using blending data approach. Int. J. Comput. Appl. **81**(8), 22–28 (2013)
8. Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E.: How to find trouble in communication. Speech Commun. **40**(1–2), 117–143 (2003)
9. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
10. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of German emotional speech. In: Ninth European Conference on Speech Communication and Technology (2005)
11. Calix, R.A., Khazaeli, M.A., Javadpour, L., Knapp, G.M.: Dimensionality reduction and classification analysis on the audio section of the SEMAINE database. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011. LNCS, vol. 6975, pp. 323–331. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24571-8_43
12. Douglas-Cowie, E., et al.: The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 488–500. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74889-2_43
13. Engberg, I.S., Hansen, A.V., Andersen, O., Dalsgaard, P.: Design, recording and verification of a Danish emotional speech database. In: Fifth European Conference on Speech Communication and Technology (1997)
14. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia, MM 2013, Barcelona, Spain, pp. 835–838 (2013). https://doi.org/10.1145/2502081.2502224
15. Eyben, F., Weninger, F., Wöllmer, M., Schuller, B.: openSMILE: open-Source Media Interpretation by Large feature-space Extraction. audEERING GmbH (2016)

16. Frank, E., Kirkby, R.: Weka Classifiers Trees: Random Tree. http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/RandomTree.html (2018). Accessed 22 Feb 2018
17. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Thirteenth International Conference on Machine Learning, pp. 148–156. Morgan Kaufmann, San Francisco (1996)
18. Friedman, J., Hastie, T., Tibshirani, R.: Additive Logistic Regression: a Statistical View of Boosting. Stanford University, Technical report (1998)
19. Fujisaki, H.: The Interplay Between Physiology, Physics and Phonetics in the Production of Tonal Features of Speech of Various Languages. In: Proceedings of the 10th International Workshop Speech and Computer (SPECOM), Patras, Greece, pp. 39–48, October 2005
20. Grimm, M., Kroschel, K., Narayanan, S.: The vera am mittag German audio-visual emotional speech database. In: 2008 IEEE International Conference on Multimedia and Expo, pp. 865–868. IEEE (2008)
21. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009). https://doi.org/10.1145/1656274.1656278
22. Hammami, A.: Towards developing a speech emotion database for Tunisian Arabic (2018)
23. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (eds.) Advances in Neural Information Processing Systems, vol. 10. MIT Press, Cambridge (1998)
24. Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., Nogueiras, A.: Interface databases: design and collection of a multilingual emotional speech database. In: LREC (2002)
25. Jovicic, S.T., Kasic, Z., Dordevic, M., Rajkovic, M.: Serbian emotional speech database: design, processing and evaluation. In: 9th Conference Speech and Computer (2004)
26. Keerthi, S., Shevade, S., Bhattacharyya, C., Murthy, K.: Improvements to platt's SMO algorithm for SVM classifier design. Neural Comput. **13**(3), 637–649 (2001)
27. Khalil, A., Al-Khatib, W., El-Alfy, E.S., Cheded, L.: Anger detection in Arabic speech dialogs. In: 2018 International Conference on Computing Sciences and Engineering (ICCSE), pp. 1–6. IEEE (2018)
28. Khalil, A.A.A.S.: Real-Time Anger Detection In Arabic Speech Dialogs. Ph.D. thesis, King Fahd University of Petroleum and Minerals (Saudi Arabia) (2011)
29. Klaylat, S., Osman, Z., Hamandi, L., Zantout, R.: Emotion recognition in Arabic speech. Analog. Integr. Circuits Signal Process. **96**(2), 337–351 (2018)
30. Klaylat, S., Osman, Z., Hamandi, L., Zantout, R.: Enhancement of an Arabic speech emotion recognition system. Int. J. Appl. Eng. Res. **13**(5), 2380–2389 (2018)
31. Kostoulas, T., Ganchev, T., Mporas, I., Fakotakis, N.: A real-world emotional speech corpus for modern Greek. In: LREC (2008)
32. Landwehr, N., Hall, M., Frank, E.: Logist. Model Trees **95**(1–2), 161–205 (2005)
33. Liberman, M.: Emotional prosody speech and transcripts (2002). http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28
34. Meddeb, M., BenAmmar, M., Alimi, A.: Towards a recommendation system for TV programs based on human behavior. In: The International Conference on Control, Engineering Information Technology, CEIT, pp. 180–182 (2013)
35. Meddeb, M., Hichem, K., Alimi, A.: Automated extraction of features from Arabic emotional speech corpus. Int. J. Comput. Inf. Syst. Ind. Manag. Appl. IJCISIM **8**, 184–194 (2016)

36. Meddeb, M., Karray, H., Alimi, A.M.: Intelligent remote control for TV program based on emotion in Arabic speech. arXiv preprint arXiv:1404.5248 (2014)
37. Meddeb, M., Karray, H., Alimi, A.M.: Speech emotion recognition based on Arabic features. In: 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 46–51. IEEE (2015)
38. Meddeb, M., Karray, H., Alimi, A.M.: Content-based arabic speech similarity search and emotion detection. In: Hassanien, A.E., Shaalan, K., Gaber, T., Azar, A.T., Tolba, M.F. (eds.) AISI 2016. AISC, vol. 533, pp. 530–539. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-48308-5_51
39. Meddeb, M., Karray, H., Alimi, A.M.: Building and analysing emotion corpus of the Arabic speech. In: 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), pp. 134–139. IEEE (2017)
40. Meftah, A., Alotaibi, Y., Selouani, S.A.: Designing, building, and analyzing an Arabic speech emotional corpus. In: Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme, p. 22 (2014)
41. Meftah, A.H., Alotaibi, Y.A., Selouani, S.A.: Evaluation of an Arabic speech corpus of emotions: a perceptual and statistical analysis. IEEE Access **6**, 72845–72861 (2018)
42. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press, Cambridge (1998). http://research.microsoft.com/ jplatt/smo.html
43. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)
44. Schuller, B.,et al.: The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In: Proceedings of Interspeech, Lyon, France, pp. 148–152 (2013)
45. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. **53**(9–10), 1062–1087 (2011). https://doi.org/10.1016/j.specom.2011.01.011
46. Schuller, B.W., Steidl, S., Batliner, A.: The INTERSPEECH 2009 emotion challenge. In: Proceedings of Interspeech, Brighton UK, pp. 312–315 (2009)
47. Staroniewicz, P., Majewski, W.: Polish emotional speech database – recording and preliminary validation. In: Esposito, A., Vích, R. (eds.) Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions. LNCS (LNAI), vol. 5641, pp. 42–49. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-03320-9_5
48. Sumner, M., Frank, E., Hall, M.: Speeding up logistic model tree induction. In: Jorge, A.M., Torgo, L., Brazdil, P., Camacho, R., Gama, J. (eds.) PKDD 2005. LNCS (LNAI), vol. 3721, pp. 675–683. Springer, Heidelberg (2005). https://doi.org/10.1007/11564126_72
49. Sun, X., Lichtenauer, J., Valstar, M., Nijholt, A., Pantic, M.: A multimodal database for mimicry analysis. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011. LNCS, vol. 6974, pp. 367–376. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24600-5_40
50. TV, E.: RED line (in Arabic: Khat Ahmar - social talk show (2019). https://tv.echoroukonline.com
51. Vogt, T., André, E., Bee, N.: EmoVoice—a framework for online recognition of emotions from voice. In: André, E., Dybkjær, L., Minker, W., Neumann, H., Pieraccini, R., Weber, M. (eds.) PIT 2008. LNCS (LNAI), vol. 5078, pp. 188–199. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-69369-7_21

52. Yilmazyildiz, S., Henderickx, D., Vanderborght, B., Verhelst, W., Soetens, E., Lefeber, D.: EMOGIB: emotional gibberish speech database for affective human-robot interaction. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011. LNCS, vol. 6975, pp. 163–172. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24571-8_17
53. Zaghouani, W.: Critical survey of the freely available Arabic corpora. arXiv preprint arXiv:1702.07835 (2017)