



Emotion recognition in Arabic speech

Samira Klaylat¹ · Ziad Osman² · Lama Hamandi³ · Rached Zantout⁴

Received: 25 October 2017 / Revised: 14 February 2018 / Accepted: 23 February 2018 / Published online: 14 March 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

Automatic emotion recognition from speech signals without linguistic cues has been an important emerging research area. Integrating emotions in human–computer interaction is of great importance to effectively simulate real life scenarios. Research has been focusing on recognizing emotions from acted speech while little work was done on natural real life utterances. English, French, German and Chinese corpora were used for that purpose while no natural Arabic corpus was found to date. In this paper, emotion recognition in Arabic spoken data is studied for the first time. A realistic speech corpus from Arabic TV shows is collected. The videos are labeled by their perceived emotions; namely happy, angry or surprised. Prosodic features are extracted and thirty-five classification methods are applied. Results are analyzed in this paper and conclusions and future recommendations are identified.

Keywords Emotional recognition · Arabic speech · Natural corpus · Prosodic features

1 Introduction

Emotion expression is a fundamental factor in human to human communication. The speaker's voice pitch, intonation and rate affect the emotions perceived by the listener and hence can change the semantics of the utterance. Installing an effective emotion recognition system in human–machine interaction is of great benefit. One example is call centers, where it can be used to manage

customers' disputes. Recognizing the underlying emotions of the customers leads to better customer service and increases business profit [1, 2].

Psychology, behavioral science, and psychiatry are some research areas that can benefit from automatic emotion recognition systems. Such studies can rely on more reliable measurements to analyze human behavior [3]. Lie detection and schizophrenia diagnosis are examples of applications where the objectivity and accuracy in measurements is required. The Automatic detection of moods like depression, fatigue, and anxiety can lead to better results in achieving human well-being [4] and recognizing psychiatric disorders [5]. Medical doctors may use emotional contents of a patient's speech as a diagnosing tool for various disorders [6], while [7] used emotion recognition to avoid car accidents by tracking the emotional state of the driver.

Automatic training and education scenarios can also benefit from emotion detection. The system might suggest a break or change in the tutoring speed if boredom, frustration or fatigue emotions are perceived [8]. Computer games industry may install effective emotion detection models in their systems to achieve a more natural interaction between the player and the computer [9–11].

Another important area for speech emotion recognition is mobile applications, specifically for deaf and hearing-impaired humans. IP-Relay [12] and SKC Interpret [13] are

✉ Samira Klaylat
samiraklaylat@gmail.com

Ziad Osman
zosman@bau.edu.lb

Lama Hamandi
lh13@aub.edu.lb

Rached Zantout
zantoutrn@rhu.edu.lb

¹ Department of Computer Science, Beirut Arab University, Beirut, Lebanon

² Electrical and Computer Engineering Department, Beirut Arab University, Beirut, Lebanon

³ Electrical and Computer Engineering Department, American University of Beirut, Beirut, Lebanon

⁴ Electrical and Computer Engineering Department, Rafik Hariri University, Mechref, Lebanon

examples of applications that allow a hearing-impaired person to make and receive phone calls. The hearing-impaired individual can type a message and the person on the other side hears the words spoken. When the person at the end of the line speaks to the hearing-impaired person, the words are received as text on the mobile phone of the hearing-impaired. However, since the emotion of both parties is missing, this reduces the reliability and usefulness of those systems. Effective speech-to-text and text-to-speech systems which can be used by the hearing-impaired in their everyday life, starting from a very young age, are very useful. Such systems will aid deaf people to enroll in normal schools at very young age and will help them adapt better in classrooms and with their classmates. It will help them experience a more normal childhood and hence grow up to be able to integrate within the society without external help.

In this paper, the first natural Arabic speech to recognize emotions is built, and three emotions, happy, angry and surprised are recognized. Acoustic low level descriptors are extracted and thirty-five classifiers are applied. The Sequential minimal optimization (SMO) classifier gave the best result with 95.52% accuracy.

This paper is organized as follows: a review of existing work related to emotion recognition in speech is done in Sect. 2. In Sect. 3, the process of building a natural Arabic corpus is described. In Sect. 4 the feature extraction process is presented while in Sect. 5 the classification models are proposed. Classification models with accuracy above 90% are analyzed in Sect. 6. Finally, in Sect. 7 the contributions of this paper are summarized and future work is presented.

2 Related work

Emotions do not have a clear theoretical definition. Two common strategies are used to characterize emotions discrete and continuous. Examples of discrete labels are happiness, sadness, and anger [14–16] while continuous labeling uses measures such as valence (how much is the speaker being positive or negative) and activation (how much is the speaker active or passive) [17–19]. The main advantage of using discrete labeling is that most people use this approach to describe emotions in their daily life; hence the labeling scheme in the emotion recognition model matches human experience. On the other hand, continuous labeling describes a range of emotions like measuring the scale of whether a person feels positive, negative, powerful, controlling, passive or active. This type of measuring emotions is not intuitive and requires special training to be able to assign a number indicating the level of the emotion [20].

Most emotion recognition systems are trained and tested with data available from corpora. Speech corpora used for emotion recognition are either acted, induced or natural. Acted or simulated corpora are collected from professional television or radio actors. The subjects are usually asked to read a specific sentence with a specific emotion and hence the emotions are fully blown. More than 60% of the expressive speech databases are of this kind where the aspects of recording are more controllable and acted emotions are more expressive than real ones [21]; hence process of labeling is more accurate. Table 1 shows some popular acted databases.

Based on the appraisal theory [22, 23], emotion are expressed as a reaction to events, thus to perceive more realistic emotions researchers should create the appropriate environment to trigger targeted emotions. This leads us to the induced/elicited speech corpora that are collected by creating artificial emotional situations. The speakers are involved with emotional conversation with an anchor without knowing that they are recorded. Through the conversation the anchor tries to trigger the targeted emotions of the speaker to achieve natural speech data. Such databases are semi-normal where the reacted emotions are spontaneous but the recording process is created. Table 2 shows some popular induced databases.

Recent efforts have been focusing on collecting natural speech databases, since acted data cannot model realistic data sufficiently as demonstrated by [24, 25]. One way of collecting data for such databases is from audio or television programs however this is limited by copyright issues and prevents free distribution of collected corpora in some cases. Other alternatives involve live recording of conversations. Such conversations can be between doctors and patients, parents and children, employee and employers. Other options include the use of speech collected from call centers where telephone calls between humans and machine are recorded. Unlike simulated corpora, emotions in natural speech are not highly prototyped and may convey a mixture of emotions [26, 27]. Table 3 shows a listing of some popular natural databases.

The main advantage of acted databases is that most emotions are available as well as most languages and hence results can be compared easily, however it doesn't represent real life scenarios. Induced databases are nearer to natural but if the speakers knew they are recorded then the emotions expressed will be artificial. Natural databases reflects real life situations, however not all emotions are available and the recording environment may not be suitable for modeling.

In general, the emotion recognition problem is simply a mapping between the feature space and the label space of the corresponding unit of study. Thus choosing the suitable set of features is an important step in recognizing

Table 1 Example of acted speech corpora

Name	Language	Task	References
DES	Danish	Neutral, angry, happy, sad and surprised emotions	[28]
aGender	Dutch	Age groups (child, youth, adult, and senior) and gender (female, male, child)	[27, 29]
Emo-DB	German	Neutral, anger, happiness, sadness, fear, boredom and disgust	[30]
LDC emotional prosody speech and transcripts	American English	Neutral, panic, anxiety, hot anger, cold anger, despair, sadness, elation, joy, interest, boredom, shame, pride, contempt	[31]
Serbian database of acted emotions	Serbian	Neutral, anger, happiness, sadness and fear	[32]
Emotional Speech of Mandarin and Burmese Speakers (ESMBS)	Mandarin and Burmese	Anger, disgust, fear, joy, sadness, and surprise	[33]
KISMET	American English	Approval, attention, prohibition, soothing, neutral	[34]
CLDC	Chinese	Joy, anger, surprise, fear, neutral, sadness	[35]
REGIM_TES	Arabic-Tunisian	Happiness, sadness, fear, anger, neutral	[36]
–	Herbew	Anger, disgust, fear, joy, neutral, sadness	[37]
IITKGP-SEHSC	Indian	anger, disgust, fear, happy, neutral, sadness, sarcastic and surprise	[38]
–	Mexican Spanish	Anger, happiness, neutral, sadness	[39]
–	German and English	Anger, boredom, disgust, fear, happiness, sadness and neutral	[40]
–	Tamil, Malayalam and Indian English	Anger, happy, sad	[41]

Table 2 Example of induced speech corpora

Name	Language	Task	References
Interactive emotional dyadic motion capture database (IEMOCA)	English	Discrete: anger, happiness, sadness, neutrality Continuous: valence, activation, dominance	[42]
SmartKom	German	Joy/gratification, anger/irritation, helplessness, pondering/reflecting, surprise, neutral, unidentifiable	[43]
FAU Aibo	German	Anger, emphatic, neutral, positive, rest negative idle	[44]
Belfast naturalistic database	English	Frustration, amusement, fear, disgust, surprise, sad	[45]
–	Odiya	Anger, sadness, astonish, fear, happiness, neutral	[46]

Table 3 Example of natural speech corpora

Name	Language	Task	References
Vera am Mittag (VAM)	German	Valence, activation, dominant	[47]
NATURAL	Mandarin	Anger and neutral	[48]
Speaker Personality Corpus (SPC)	French	Big five OCEAN dimensions: Openness, Conscientiousness, Conscientiousness, Extraversion, Agreeableness, Neuroticism	[49]
TUM AVIC	English	Disinterest, indifference, neutrality, interest, curiosity	[50]
Call Center Database (CCD)	English	Negative and non-negative	[51]
CEMO	French	Fear, anger, sadness, neutral, relief	[52]
–	Latin American, Japanese	Negative, positive	[53]

emotion from speech. Speech features are divided into three types: spectral, excitation, and acoustic features [54]. Spectral features, also known as vocal features, were used in [55] to distinguish speech from music and in [56, 57] to recognize emotions from Mandarin corpora. The Spectral features include:

1. MFCC (Mel frequency cepstral) coefficients [58]
2. LPCC (Linear prediction cepstral) coefficients,
3. LFPC (log frequency power) coefficients.
4. (MFB) Mel filter bank.
5. Spectral centroid.
6. Formant: F1, F2 and their bandwidth BW1, BW2.

Excitation speech features are derived from linear prediction residual of the source signal [54]. The linear prediction residual of the source signal is obtained by first predicting the vocal tract information using linear prediction coefficients (LPCs) from speech signal, and then separating it by inverse filter formulation [59]. Few studies have been conducted to recognize emotions using LP residual signals. Glottal excitation signals were used in [60] to analyze the relation between the emotional state of the speaker and emotional disorders. Few studies have been done in recognizing emotion from speech using excitation speech features [41, 61–64].

Acoustic features are the most widely used to recognize emotions from spoken data [65–69]. According to [70] these features proved to be the most effective in recognizing emotions from speech signals. Using acoustic features, five emotions; fear, anger, happiness, sadness, and neutrality were classified in [71] while negative and non-negative emotions were detected in [47]. Acoustic features represent the prosodic properties of human speech which include:

1. Pitch: measured by the fundamental frequency F0.
2. Intensity/energy: models the loudness of the sound signal.
3. Duration: measured by zero-crossing rate.
4. Voice quality: measured by Noise-to-Harmonic Ratio, jitter, shimmer.

The number of speech features used in emotion recognition systems depends on the training and testing conditions like the corpus size and type, emotions to be recognized, and recording environment. Choosing a large number of features may increase the complexity of the problem. One approach is to use a small feature set based on expert knowledge which is computationally efficient, but unfortunately, there is no general agreement on what are the effective features to recognize emotions. Since not all features make sense in some models, one standard approach is to compose a large feature vector and then

eliminate ineffective features to achieve better performance. This is called the Brute-forcing technique [72].

Combining several feature types has proved to improve performance of emotion recognition systems [73]. These features are naturally related to each other and hence the correct combination of features is expected to improve the emotion recognition performance. A performance of 88% and 75% was attained by [74, 75] respectively when combining prosodic and acoustic features. An accuracy of 80% was achieved by [76] when using Spectral, prosodic, disfluent and linguistic features. In [77], 50% accuracy is accomplished when combining LPPC and pitch related features to recognize eight emotions. A two phase approach was proposed by [78], where spectral, prosodic, and voice quality features were extracted and then in the first phase, the best features within each group were selected. Then, in the second phase, the final feature set was composed from the candidate features. Spectral and prosodic features were combined by [79, 80] to recognize emotions from the Danish database DES with a result of 52 and 83% respectively.

Emotion recognition from spoken language can be considered as a classification problem. Hence machine learning methods can be used to design a system that will classify emotions by learning about the classes from a corpus. Classifiers used in emotion recognition from speech can be categorized in two categories linear classifiers and non-linear classifiers.

1. Linear classifiers: These are classifiers that perform classification based on a linear combination of the features. They are simple and computationally efficient. These classifiers perform faster if the features are linearly separable. Linear classifiers include Naïve Bayes, logistic regression and Support Vector Machines. SVMs are the most widely used classifiers [81–83] and have proved to produce high accuracy results especially for small data sets. SVM was applied by [84, 85] to classify emotions from Mandarin language with a result of 88% and 76% respectively.
2. Non-linear classifiers: for nonlinearly separable data, linear classifiers might lead to inaccurate results. In such cases, non-linear classifiers are more efficient. Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Artificial Neural Networks (ANN), Polynomial classifiers, K-nearest neighbours (KNN) and Decision trees are all examples of non-linear classifiers employed to recognize emotion from speech. GMM achieved a 92% and 74.6% accuracy when used by [86, 87] to recognize emotions from a Basque and Berlin corpora respectively. In [88] seven emotions were recognized from Berlin corpus using ANN and a 51% performance is reported. Negative

and non-negative emotions were recognized in [89] by classifying acoustic features using KNN classifier, a result of 75.81% for male and 80% for female is reported. In [90], decision tree classifier accomplished a 66% performance when recognizing positive, negative and neutral emotions from computer tutor dialogues.

Several approaches to combine classifiers were proposed such as the GMM-SVM approach in [91] to recognize idle and negative emotions and the GMM-HMM approach in [92] to recognize neutral, anger, happiness, sadness, fear, boredom and disgust emotions. A SVM-KNN model was proposed by [93] and got 87% accuracy when applied on the KISMET database [34]. In literature, choosing a specific classifier was rarely justified. Usually they are chosen based on past references or some experimental evaluation. No classifier or combination of classifiers has been proved to give the best accuracy in emotion recognition systems.

3 Corpus engineering

The majority of speech databases are in English followed by German and Chinese. Little work was done on Dutch, Swedish, Russian, Japanese, Slovenian and Spanish languages. An Arabic acted corpus composed of Tunisian dialect isolated words was created by [36]. In this study the first Arabic natural corpus of different dialects is proposed to recognize discrete emotions.

The characteristics of the database depend on the purpose of the research. The motivation of this study is based on helping hearing-impaired and deaf people to improve their daily life communication. Integrating an effective emotion recognition system with a reliable speech-to-text system enables deaf and hearing impaired individuals to make successful phone calls with normal people. Therefore, we target to collect natural phone call recordings to build our corpus.

Eight videos of live calls between an anchor and a human outside the studio are downloaded from online Arabic talk shows [94–101]. Since the videos are available online for public, no copyright issues exist. Eighteen human labelers are asked to listen to the videos and label each one of them as happy, angry or surprised. The average result is used to label each video.

Each video is then divided into turns: callers and receivers. Silence, laughs and noisy chunks were removed. Every chunk was then automatically divided into 1 s speech units forming our final corpus composed of 1384 records with 505 happy, 137 surprised and 741 angry units. No global properties are specified for building a speech database for recognizing emotions. In Table 4 the properties of popular corpora used in different emotion recognition research are listed. Notice that the size, number of chunks, and number of labelers vary a lot between the databases. Table 5 summarizes the properties of the videos used to build the corpus [102].

4 Feature extraction

As mentioned in Sect. 2, acoustic features have proved to be very effective in recognizing emotions [69]. A combination of acoustic and spectral features, known as low-level descriptors, was provided for participants by the first international research challenge INTERSPEECH conference in 2009 [107]. This challenge was initiated to provide a good benchmark for speech processing tasks and to enable more accurate comparison between models proposed by participants. More similar challenges took place later like the Interspeech 2010 Paralinguistic Challenge [108], the Interspeech 2011 Speaker State Challenge [109], the Interspeech 2012 Speaker Trait Challenge [110] and second Audio/Visual Emotion Challenge (AVEC 2011 [111] and AVEC 2012 [112]).

These Low-level descriptors (LLDs) are extracted using the open source OpenSMILE feature extractor [113] that is developed by Technische Universität München's (TUM's).

Table 4 Properties of existing corpora for emotion recognition in speech

Corpus	Length (h)	Language	# of chunks	# of labelers	References
FAU AEC	8.9	German	18,216	5	[40]
TUM AVIC	2.3	English	3880	4	[50]
aGender	50.6	German	65,364	–	[27]
ALC	43.8	German	12,360	–	[103]
SLC	21.3	German	9089	3	[104]
SPC	1.7	French	640	11	[49]
SLD	0.7	German	800	32	[105]
TIMIT	4.4	English	6300	–	[106]

Table 5 Videos used to build the corpus

Id	Dialect	Gender	Length (s)	#of chunks	Emotion perceived
1	Egyptian	Male	114	9	Happy
2	Egyptian	Male	78	6	Surprised
3	Gulf	Female	73.8	6	Happy
4	Jordan	Male	210	17	Angry
5	Gulf	Male	198	34	Angry
6	Egyptian	Female	23.4	2	Surprised
7	Lebanese	Female	504	24	Angry
8	Egyptian	Female	430.8	87	Happy

LLDs extracted from every speech unit are listed in Table 6.

On each of the above features the following statistical functions are calculated:

1. Maximum
2. Minimum
3. Range: Max–Min
4. Absolute position of maximum
5. Absolute position of minimum
6. Arithmetic Mean
7. Linear Regression 1: slope of linear approximation of contour
8. Linear Regression 2: offset of linear approximation of the contour.
9. Linear Regression A: difference of linear approximation and the contour.
10. Linear Regression Q: quadratic error between the linear approximation and the contour.
11. Standard deviation
12. Kurtosis
13. Skewness
14. Quartiles 1, 2, 3
15. Inter-quartile ranges 1–2, 2–3, 1–3.

Next, the delta coefficient for every LLD is also computed as an estimate of the first derivative hence leading to a total of 950 features as detailed in Table 7.

Finally, to remove ineffective features, Kruskal–Wallis non-parametric test [114] is applied. We considered a

significance level of 0.05 i.e. a level of 95% confidence; hence we removed the features with p-values less than 0.05 resulting in a new database [102] of 1384 records with 845 features.

5 Classification

Thirty-five classifiers belonging to six classification groups are applied separately on the collected speech corpus. The classification groups are Trees, Rules, Bayes, Misc., Lazy, Functions and Meta. The Trees group includes LMT, REPTree, Random Forest, Decision Stump, Random Tree, HoefdingTree, and J48 classifiers, while the Rules group includes Jrip, Decision Table, PART Decision List, OneR, and ZeroR. The K-Nearest Neighbour classifier belongs to the Lazy group, while the Bayes group includes Naïve Bayes, Naïve Bayes Updatable, Bayes Net, and Naive Bayes Multinomial classifiers. The Logistic, Simple Logistic, and the Sequential Minimal Optimization classifiers belong to the Functions group. Finally, the Meta group includes the RandomSubspace, Filtered Classifier, Multi-class Classifier, Weighted Instances Handler Wrapper, Iterative Classifier Optimizer, Classification via Regression, Attribute Selected, MultiClass Classifier Updatable, Multi scheme, Logit Boost, Random Committee, Randomizable Filtered Classifier, Adaptive Boosting, and Bagging classifiers.

A 10-fold cross validation [115] was applied to all classifiers. In 10-fold cross-validation, the database is randomly partitioned into 10 equal size subsamples. Of the 10 subsamples, 1 subsample which is 10% of the database is considered as the testing data to validate the classification model, and the remaining 9 subsamples are used as

Table 6 Low-level descriptors

Group	LLDs
1	Loudness Intensity
2	WaveForm Zero crossing rates
3	Cepstral MFCC 1–12 (Mel-frequency cepstral coefficients)
4	Pitch F0 (Fundamental frequency)
5	F0 envelope
6	Probability of voicing
7	LSP (Line Spectral frequency)

Table 7 Calculating the number of features used

LLDS	25
Functionals	19
Total =	25 * 19 = 475
Delta regression for each feature	475 * 2 = 950

training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the testing data. This leads to a lower variance than a single hold-out set estimator, which can be very important if the amount of data available is limited. If a single hold out set is considered, where 90% of data are used for training and 10% used for testing, or even 80% for training and 20% testing, the test set is very small, and hence a lot of variation in the performance estimate for different partitions of the data to form training and test sets. 10-fold validation reduces this variance by averaging over 10 different partitions, so the performance estimate is less sensitive to the partitioning of the data. Another advantage of this method is that all subsamples are used for both training and testing, and each subsample is used for validation only once.

The highest accuracy of 95.52% was achieved by Sequential Mean Optimization (SMO) classifier and the lowest result was 53.58% by four different methods. For every classification the rate of true positive, rate of false positive, precision and recall are computed for happy, surprised and angry emotion classes. Tables 8, 9, 10, 11 and 12 shows the confusion matrix and accuracy details of all classification models. In these tables, H corresponds to Happy, S corresponds to Surprised, and A corresponds to Angry.

The classifiers with accuracy between 96 and 90% are:

1. SMO: 95.52%
2. Simple Logistic: 95.45%
3. Logistic Model Trees: 95.45%
4. Random Sub Space with RepTree: 94.14%
5. Random Committee with RandomTree: 93.06%
6. Bagging with RepTree: 92.99%
7. Logit Boost performs additive logistic regression to Decision Stump: 92.91%
8. Iterative Classifier Optimizer optimizes the number of iterations of the given iterative classifier using cross-validation with the LogitBoost: 92.91%
9. Random Forest: 92.84%
10. Multiclass Classifier Updatable: 92.70%
11. Classification via regression: 92.62%
12. K-Nearest Neighbour: 91.11%
13. Filtered Classifier with J48: 90.74%

The classifiers with accuracy between 89.7 and 80% are:

14. PART: 89.66%
15. RepTree: 89.37%
16. Attribute Selected with J48: 89.29%
17. JRIP: 88.72%
18. J48: 88.65%
19. Multiclass Classifier: 84.53%
20. Decision Table: 84.09%

21. Random Tree: 82.36%
22. Logistic: 80.69%

The classifiers with accuracy between 79.7 and 70% are:

23. Decision Stump: 79.10%
24. Adaptive Boosting for Decision Stump: 79.61%
25. OneR: 78.02%
26. Bayes Net: 73.17%

The classifiers with accuracy between 69.9 and 60% are:

27. Hoeffding Tree: 69.99%
28. Naive Bayes: 69.05%
29. Naive Bayes Updatable: 69.05%
30. Naive Bayes Multinomial: 68.25%

The classifiers with accuracy below 60% are:

31. Randomizable Filtered Classifier with K-nearest neighbour: 56.98%
32. ZeroR: 53.58%
33. CVParameterSelection with ZeroR: 53.58%
34. Weighted Instances Handler Wrapper with ZeroR: 53.58%
35. InputMapped Classifier with ZeroR: 53.58%

6 Analysis

In this section the classification models with performance above 90% are analyzed. The highest accuracy of 95.52% is achieved by Sequential Mean Optimization (SMO) classifier. SMO was invented by [116] to solve quadratic problems of training large SVM [117] models. SVM takes a geometric optimization approach to the classification problem, seeking to construct a hyperplane or a set of hyperplanes in a high dimensional space [118]. Many studies have shown highest classification accuracy for SVM compared to KNN [119, 120], and to Linear Least Squares Fit, Naïve Bayes and Neural Networks as well [120]. SMO uses heuristics to partition the training problem into smaller problems that can be solved analytically. The values of the kernel and calibrator parameters of the SMO classifier are polykernel and logistics respectively.

Next, we observe from Table 8 that tree-based models and regression-based models have promising results. Simple Logistic and Logistic Model Trees (LMT) have equal performance of 95.45% and very similar accuracy details. Both methods use linear regression. The LMT [121] combines trees model and linear models; trees with linear regression functions at the leaves. The RandomSubspace [122] model constructs a decision tree classifier that maintains highest accuracy on training data and improves on generalization accuracy as it grows in complexity. The

Table 8 Confusion matrix and accuracy details of classifiers with accuracy more than 90%

Classification method	Class	Classified as			TP rate	FP rate	Precision	Recall
		H	S	A				
SMO	H	489	8	8	0.97	0.03	0.96	0.97
	S	13	111	13	0.81	0.02	0.85	0.81
	A	9	11	721	0.97	0.03	0.97	0.97
Simple logistic	H	484	10	11	0.96	0.02	0.97	0.96
	S	11	112	14	0.82	0.02	0.84	0.82
	A	6	11	724	0.98	0.04	0.97	0.98
Logistic model trees	H	485	9	11	0.96	0.02	0.96	0.96
	S	11	112	14	0.82	0.02	0.85	0.82
	A	7	11	723	0.98	0.04	0.97	0.98
K-Nearest neighbor	H	471	21	13	0.93	0.06	0.90	0.93
	S	20	103	14	0.75	0.04	0.70	0.75
	A	31	24	686	0.93	0.04	0.96	0.93
Random subspace	H	485	8	12	0.95	0.02	0.96	0.95
	S	19	84	34	0.67	0.01	0.94	0.67
	A	8	0	733	0.99	0.08	0.93	0.99
Random committee	H	486	7	12	0.96	0.03	0.94	0.96
	S	20	78	39	0.66	0.00	0.95	0.66
	A	16	2	723	0.98	0.08	0.94	0.98
Bagging	H	472	9	24	0.94	0.03	0.94	0.94
	S	20	82	35	0.60	0.01	0.90	0.60
	A	9	0	723	0.99	0.09	0.93	0.99
Iterative classifier optimizer	H	477	17	11	0.95	0.03	0.95	0.95
	S	19	82	36	0.60	0.02	0.77	0.60
	A	8	7	726	0.98	0.07	0.94	0.98
Logit boost	H	477	17	11	0.95	0.03	0.95	0.95
	S	19	82	36	0.60	0.02	0.77	0.60
	A	8	7	726	0.98	0.07	0.94	0.98
Random forest	H	485	4	16	0.96	0.03	0.95	0.96
	S	22	65	50	0.55	0.00	0.95	0.55
	A	6	1	724	0.99	0.09	0.93	0.99
MultiClass classifier updateable	H	474	8	23	0.94	0.02	0.96	0.94
	S	10	102	25	0.72	0.02	0.78	0.72
	A	11	24	706	0.96	0.08	0.93	0.96
Classification via regression	H	470	12	23	0.93	0.02	0.96	0.93
	S	13	87	37	0.64	0.02	0.80	0.64
	A	7	10	724	0.98	0.09	0.92	0.98
Filtered classifier	H	468	21	16	0.93	0.04	0.93	0.93
	S	14	94	29	0.69	0.04	0.66	0.69
	A	20	28	693	0.94	0.07	0.94	0.94

model consists of multiple trees constructed systematically by randomly selecting subsets of components of the feature vector, that is, trees constructed in randomly chosen subspaces. When used with RepTree, a high result of 93.06% is obtained. The RepTree [123] classifier is a fast decision tree learner. It builds a decision tree using information gain/variance and prunes it using reduced-error pruning. When applied alone, the RepTree model gave a result

of 89.37% however when applied with the Bagging [123] model, the performance is enhanced to 92.99%.

The Random Tree classifier [123] has 82.36% accuracy; however the performance increases to 93.06% when applied with the Random Committee model [123]. The Random Committee model is an ensemble of randomizable base classifiers. Each base classifier is built using a different random number seed based on the same data. The

Table 9 Confusion matrix and accuracy details of classifiers with accuracy between 89.9 and 80%

Classification method	Class	Classified as			TP rate	FP rate	Precision	Recall
		H	S	A				
PART	H	460	23	22	0.92	0.03	0.94	0.92
	S	24	81	32	0.67	0.04	0.67	0.67
	A	19	23	699	0.95	0.07	0.94	0.95
REPTree	H	464	21	20	0.92	0.05	0.91	0.92
	S	26	76	35	0.56	0.04	0.63	0.56
	A	21	24	696	0.94	0.09	0.93	0.94
Attribute selected	H	468	15	22	0.94	0.05	0.91	0.94
	S	21	79	37	0.60	0.03	0.67	0.60
	A	28	25	688	0.93	0.09	0.93	0.93
Jrip	H	459	20	26	0.91	0.06	0.91	0.91
	S	18	80	39	0.63	0.04	0.66	0.63
	A	26	27	688	0.93	0.09	0.93	0.93
J48	H	457	24	24	0.91	0.05	0.92	0.91
	S	19	85	33	0.64	0.05	0.60	0.64
	A	22	35	684	0.92	0.08	0.93	0.92
MultiClass classifier	H	415	26	64	0.83	0.07	0.87	0.83
	S	17	96	24	0.71	0.05	0.62	0.71
	A	49	34	685	0.89	0.13	0.89	0.89
Decision table	H	437	20	48	0.87	0.09	0.84	0.87
	S	36	45	56	0.33	0.03	0.57	0.33
	A	46	14	681	0.92	0.16	0.87	0.92
Random tree	H	422	27	56	0.87	0.10	0.83	0.87
	S	23	69	45	0.43	0.06	0.46	0.43
	A	57	36	648	0.86	0.14	0.88	0.86
Logistic	H	408	57	40	0.80	0.10	0.83	0.80
	S	33	83	21	0.61	0.10	0.41	0.61
	A	54	62	625	0.84	0.10	0.91	0.84

Table 10 Confusion matrix and accuracy details of classifiers with accuracy between 79.9 and 70%

Classification method	Class	Classified as			TP rate	FP rate	Precision	Recall
		H	S	A				
Adaptive boosting	H	366	0	139	0.73	0.01	0.98	0.73
	S	0	0	137	0.00	0.00	0.00	0.00
	A	6	0	735	0.99	0.43	0.73	0.99
Decision stump	H	359	0	146	0.71	0.01	0.98	0.71
	S	0	0	137	0.00	0.00	0.00	0.00
	A	6	0	735	0.99	0.44	0.72	0.99
OneR	H	407	3	95	0.81	0.09	0.84	0.81
	S	20	0	117	0.00	0.01	0.00	0.00
	A	59	10	672	0.91	0.33	0.76	0.91
BayesNet	H	398	93	14	0.79	0.07	0.87	0.79
	S	11	89	37	0.65	0.21	0.25	0.65
	A	48	168	525	0.71	0.08	0.91	0.71

final prediction is a straight average of the predictions generated by the individual base classifiers. The Decision stump [123] model consists of one level decision tree,

where the tree contains one internal root node connected directly to the leaves. When applied to the corpus, a result of 79.10% is obtained but an enhancement up to 92.91% is

Table 11 Confusion matrix and accuracy details of classifiers with accuracy between 69.9 and 60%

Classification method	Class	Classified as			TP rate	FP rate	Precision	Recall
		H	S	A				
Hoeffding tree	H	332	123	50	0.66	0.05	0.88	0.66
	S	6	106	25	0.85	0.26	0.26	0.85
	A	41	170	530	0.69	0.09	0.90	0.69
NaiveBayes	H	333	127	45	0.66	0.05	0.88	0.66
	S	6	108	23	0.79	0.25	0.26	0.79
	A	40	187	514	0.69	0.11	0.88	0.69
Naïve Bayes updateable	H	333	127	45	0.66	0.05	0.88	0.66
	S	6	108	23	0.79	0.25	0.26	0.79
	A	40	187	514	0.69	0.11	0.88	0.69
Naïve Bayes multinomial	H	339	106	60	0.67	0.07	0.85	0.67
	S	5	94	38	0.69	0.23	0.25	0.69
	A	57	174	510	0.69	0.15	0.84	0.69

Table 12 Confusion matrix and accuracy details of classifiers with accuracy below 60%

Classification method	Class	Classified as			TP rate	FP rate	Precision	Recall
		H	S	A				
Randomizable filtered classifier	H	291	40	174	0.65	0.23	0.62	0.65
	S	47	32	58	0.28	0.09	0.26	0.28
	A	201	75	465	0.69	0.31	0.72	0.69
Weighted instances handler wrapper ZeroR CV parameter InputMapped classifier	H	0	0	505	0.00	0.00	0.00	0.00
	S	0	0	137	0.00	0.00	0.00	0.00
	A	0	0	741	1.00	1.00	0.54	1.00

achieved when performing additive logistic regression using Logit Boost [123] model. On the other hand, the Classifier Optimizer [123], which optimizes the number of iterations of the given iterative classifier, doesn't improve the result obtained by Logit Boost. The Random Forest model [123] which constructs a forest of random trees gives 92.84% accuracy.

The Multiclass Classifier Updatable [123] model is used when a multi-class dataset is classified by two-class classifiers. In this work a 3-class dataset (happy, angry, and surprised) is classified by the SGD [123] 2-class classifier giving 92.70% accuracy. The Classification-via-regression model also applies regression methods. The dataset is binarized and one regression model is built for each class

value [124]. When applying the KNN classifier [125] on the proposed dataset, the K (number of neighbors to use) value is equal to 1 and the nearest neighbor search algorithm applied is the brute force algorithm search. Finally, the J48 classifier [123] gives 88.65% accuracy result, but the result is boosted to 90.74% when applying the filtered classifier model [123]. With the filtered classifier; the training data are not affected by the testing data which leads to better performance and may be applied to other classifiers for enhancement.

In Table 13 the proposed work is compared with similar studies that used SVM classification model to recognize emotions using acoustic features. Acted and natural corpora of different languages (German, English, French and

Table 13 Some studies that applied SMO/SMV to recognize emotions from speech

Language	Type	Emotions	Result (%)
Arabic (this work)	Natural	Happiness, angry, surprise	95.5
Arabic [126]	Acted	Happiness, angry, fear, neutral, sadness	93
German [127]	Acted	Anger, sadness boredom, disgust, fear, joy, neutral	85.2
English [92]	Mixed	Stress, neutral	83
French [76]	Natural	Positive, negative	83.16
Mandarin [48]	Natural	Anger, neutral	76.93

Arabic) are listed. Notice that the proposed system gives high accuracy 95.5% compared when compared to natural corpora (French [76], Mandarin [48]) where emotions are not prototyped. No work was found that recognize happy, angry and surprised emotions using prosodic features and SVM model.

For the Arabic corpus, only one study was found [118] in recognizing emotions from Arabic spoken language, however, in that study, professional Tunisian actors were employed to express isolated Arabic words in Tunisian dialect to recognize happy, anger, fear, sadness and neutral emotions and the overall accuracy of their proposed model was not mentioned. In this study, different Arabic dialects were considered (Egyptian, Gulf and Lebanese) see Table 5, the utterance was a complete natural speech and not isolated words, resulting in a more reliable speech corpus [102].

7 Contributions and future work

In this paper the first system to recognize sentiment in natural Arabic speech is built. This is a contribution not only to Arabic speech but to speech in general since there are few corpora for natural speech in existence. Another contribution of this work is the number of features extracted which is 950. This number is considered to be high compared to literature, for example, [128] used 17 pitch features, [129] used 27 prosodic features, [130] used 276 features, and [107] used 384 features while [131] reduced 375 features to 32. One more contribution is the use of thirty-five classification methods to classify happy, angry and surprised emotions where the SMO method gives the best classification performance of 95.5%.

In future research excitation features could be extracted and other classification models like hidden Markov and Gaussian model could be used. An intelligent 2-phase model to recognize emotions is built based on this work, with a 3% improvement in the performance of all the classifiers. The proposed feature set may be deduced such that correlated features are removed. Also combining more than one classifier to improve performance could be studied. More emotions may be recognized from the proposed corpus and results can be compared with those obtained using acted Arabic speech.

References

- Petrushin, V. (2000). Emotion recognition in speech signal: Experimental study, development, and application. In *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China.
- Liscombe, J., Riccardi, G., & Hakkani-Tnr, D. (2005). Using context to improve emotion detection in spoken dialog systems. In *Interspeech*, pp. 1845–1848.
- Roisman, G. I., Tsai, J. L., & Chiang, K. S. (2004). The emotional integration of childhood experience: Physiological, facial expressive, and self-reported emotional response during the adult attachment interview. *Developmental Psychology*, *40*(5), 776–789.
- Pantic, M., Pentland, A., Nijholt, A., & Huang, T. S. (2006). Human computing and machine understanding of human behavior: A survey. In *Proceedings Eighth ACM Int'l Conf. Multimodal Interfaces*, 2006, pp. 239–248.
- Chevrie-Muller, C., Segurier, N., Spira, A., & Dordain, M. (1978). Recognition of psychiatric disorders from voice quality. *Language and Speech*, *21*, 87–111.
- France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Eng.*, *47*(7), 829–837.
- Ji, Q., Lan, P., & Looney, C. (2006). A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Systems, Man, and Cybernetics Part A*, *36*(5), 862–875.
- Ai, H., Litman, D. J., Forbes-Riley, K., Rotaru, M., Tetreault, J., & Purandare, A. (2006). Using system and user performance features to improve emotion detection in spoken tutoring systems. In *Proceedings of Interspeech*, 2006, pp. 797–800.
- Klein, J., Moon, Y., & Picard, R. W. (2002). This computer responds to user frustration: Theory, design and results. *Interacting with Computers*, *14*(2), 119–140.
- Kuncheva, L., Bezdek, J., & Duin, R. (2006). Decision templates for multiple classifier fusion: An experimental comparison. *Pattern Recognition*, *34*(2), 299–314.
- Scheirer, J., Fernandez, R., Klein, J., & Picard, R. W. (2002). Frustrating the user on purpose: A step toward building an affective computer. *Interactive Comput.*, *34*(2), 93–118.
- <http://android-apps.com/apps/skc-interpret/>, Android Apps website.
- <http://appcrawlr.com/android/sprint-mobile-ip>, Sprint Mobile IP, App Crawl website.
- Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. In *Proceedings of Nebraska Symp. Motivation*, pp. 207–283.
- Ekman, P. (1982). *Emotion in the human face* (2nd ed.). Cambridge: Cambridge University Press.
- Ekman, P., & Oster, H. (1979). Facial expressions of emotion. *Annual Review of Psychology*, *30*, 527–554.
- Clavel, C., Vasilescu, I., Devillers, L., Richard, G., & Ehrette, T. (2008). Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, *50*(6), 487–503.
- Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*(1–2), 5–32.
- Kehrein, R. (2002). The prosody of authentic emotions. In *Proceedings of Speech Prosody*, Aix-en-Provence, 2002, pp. 423–426.
- Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). ‘Feeltrace’: An instrument for recording perceived emotion in real time. In *Proceedings ISCA Workshop Speech and Emotion*, 2000, pp. 19–24.
- Ayadi, M. E., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*, 572–587.
- Marsella, S. C., & Gratch, J. (2009). EMA: A process model of appraisal dynamics. *Journal of Cognitive Systems Research*, *10*(1), 70–90.

23. Gratch, J., Marsella, S., & Petta, P. (2009). Modeling the cognitive antecedents and consequences of emotion. *Journal of Cognitive Systems Research*, 10(1), 1–5.
24. Batliner, A., Fischer, K., Huber, R., Spilker, J., & E. Nöth. (2000). Desperately seeking emotions: Actors, wizards, and human beings. In *Proceedings of the ISCA workshop on speech and emotion*, Newcastle, Northern Ireland, 2000, pp. 195–200.
25. Wilting, J., Kraemer, E., & Swerts, M. (2006). Real vs. acted emotional speech. In *Proceedings of Interspeech*, Pittsburgh, PA, 2006, pp. 805–808.
26. Douglas, E., Devillers, L., Martin, J. C., Cowie, R., Savvidou, S., & Abrilian, S. (2005). Multimodal databases of everyday emotion: Facing up to complexity. In *9th European Conference on Speech Communication and Technology* Lisbon, Portugal, 2005, pp. 813–816.
27. Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4), 407–422.
28. Engberg, I., & Hansen, A. (1996). *Documentation of the Danish emotional speech database DES*. Center for Person Communication, Institute of Electronic Systems, Aalborg University, Aalborg, Denmark.
29. Burkhardt, F., Eckert, M., Johannsen, W., & Stegmann, J. (2010). A database of age and gender annotated telephone speech. In *Proceedings of LREC*, Valletta, Malta, 2010, pp. 1562–1565.
30. Schuller, B., Eyben, F., Can, S., & Feussner, H. (2010). Speech in minimal invasive surgery—Towards an affective language resource of real-life medical operations. In *Proceedings of the 3rd International Workshop on emotion: Corpora for Research on Emotion and Affect, satellite of LREC*, Valletta, Malta, 2010, pp. 5–9.
31. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005) A database of German emotional speech. In *Proceedings of Interspeech*, Lisbon, 2005, pp. 1517–1520.
32. <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28S>, University of Pennsylvania Linguistic Data Consortium, Emotional prosody speech and transcripts, July, 2002.
33. Jovicic, S. T., Kacic, Z., Dordevic, M., & Rajkovic, M. (2004). Serbian emotional speech database: Design, processing and evaluation. In *Proceedings of 9th Conference on Speech and Computer*, St. Petersburg, Russia, 2004, pp. 77–81.
34. Nwe, T. L. (2003). *Analysis and detection of human emotion and stress from speech signals*. Ph.D. thesis, Department of Electrical and Computer Engineering, National University of Singapore, 2003.
35. Breazeal, C., & Aryananda, L. (2002). Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots*, 12(2002), 83–104.
36. Meddeb, M., & Alimi, A. (2017). Building and analyzing emotion corpus of the arabic speech. *International Workshop on Arabic Script Analysis and Recognition (ASAR)*, IEEE, 2017.
37. Zhou, J., Wang, G., Yang, Y., & Chen, P. (2006). Speech emotion recognition based on rough set and SVM. In *5th IEEE International Conference on Cognitive Informatics*, 2006, Vol. 1, pp. 53–61.
38. Rao, K. S., & Koolagudi, S. G. (2011). Identification of Hindi dialects and emotions using spectral and prosodic features of speech. *Systemics, Cybernetics, and Informatics*, 9(4), 24–33.
39. Caballero-Morales, S. O. (2013). Recognition of emotions in Mexican Spanish speech: An approach based on acoustic modeling of emotion-specific vowels. *The Scientific World Journal*, 1–13.
40. Song, P., Ou, S., Zheng, W., Jin, Y., & Zhao, L. (2016). Speech emotion recognition using transfer non-negative matrix factorization. In *Proceedings of IEEE international conference ICASSP*, 2016, pp. 5180–5184.
41. Pravena, D., & Govind, D. (2017). Development of simulated emotion speech database for excitation source analysis. *International Journal of Speech Technology*, 20, 327–338.
42. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., & Kim, S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4), 335–359.
43. Schiel, F., Steininger, S., & Turk, U. (2002). The SmartKom multimodal corpus at BAS. In *Proceedings of the 3rd Language Resources and Evaluation Conference*, 2002, Canary Islands, Spain, pp. 200–206.
44. Batliner, A., Hacker, C., Steidl, S., Noth, E., D’Arcy, S., & Russell, M. (2004). ‘You stupid tin box’—Children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proceedings of 4th Language Resources and Evaluation Conference*, 2004, Lisbon, Portugal, pp. 171–174.
45. Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1–2), 33–60.
46. Mohanty, S., & Swain, B. K. (2010). Emotion recognition using fuzzy K-means from Oriya speech. In *International Conference [ACCTA-2010] on Special Issue of IJCCCT*, 2010, Vol. 1, Issue 2–4.
47. Grimm, M., Kroschel, K., & Narayanan, S. (2008). The Vera am Mittag German audio–visual emotional speech database. In *Proceedings IEEE International Conference on Multimedia and Expo*, 2008, Hannover, Germany, pp. 865–868.
48. Morrison, D., Wang, R., & De Silva, L. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2), 98–112.
49. Mohammadi, G., Vinciarelli, A., & Mortillaro, M. (2010). The voice of personality: Mapping nonverbal vocal behavior into trait attributions. In *Proceedings of second international workshop on social signal processing*, 2010, Florence, pp. 17–20.
50. Schuller, B., Muller, R., Eyben, F., Gast, J., Hornler, B., Wöllmer, M., et al. (2009). Being bored? recognizing natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, 27, 1760–1774.
51. Lee, C. M., & Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2), 293–303.
52. Vidrascu, L., & Devillers, L. (2006). Real-life emotions in naturalistic data recorded in a medical call center. In *1st International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation)*, Genoa, Italy, 2006, pp. 20–24.
53. Quiros-Ramirez, M. A., Polikovskiy, S., Kameda, Y., & Onisawa, T. (2014). A spontaneous cross-cultural emotion database: Latin-America vs. Japan. In *International conference on Kansei Engineering and emotion research*, 2014, pp. 1127–1134.
54. Koolagudi, S., & Sreenivasa Rao, K. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology*, 15(2), 99–117.
55. Mubarak, O. M., Ambikairajah, E., & Epps, J. (2005). Analysis of an MFCC-based audio indexing system for efficient coding of multimedia sources. In *The 8th International Symposium on Signal Processing and its Applications*, Sydney, Australia, 2005, pp. 28–31.
56. Pao, T. L., Chen, Y. T., Yeh, J. H., & Liao, W. Y. (2005). Combining acoustic features for improved emotion recognition in mandarin speech. In *Lecture Notes in Computer Science 3784*, ACII 2005 (pp. 279–285). Berlin, Heidelberg: Springer.
57. Pao, T. L., Chen, Y. T., Yeh, J. H., Cheng, Y. M., & Chien, C. S. (2007). Feature combination for better differentiating anger

- from neutral in mandarin emotional speech. In *LNCS 4738, ACII 2007*. Berlin, Heidelberg: Springer.
58. Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28, 357–366.
 59. Makhou, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4), 561–580.
 60. Cummings, K. E., & Clements, M. A. (1995). Analysis of the glottal excitation of emotionally styled and stressed speech. *Journal of Acoustic Society of America*, 98, 88–98.
 61. Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice-Hall.
 62. Chauhan, A., Koolagudi, S. G., Kafley, S., & Rao, K. S. (2010). Emotion recognition using LP residual. In *IEEE TechSym West Bengal, India*.
 63. Koolagudi, S. G., Reddy, R., & Rao, K. S. (2010). Emotion recognition from speech signal using epoch parameters. In *International conference on signal processing and communications, IISc, Bangalore, India* (pp. 1–5). New York: IEEE Press.
 64. Iliev, A., & Scordilis, M. S. (2001). Spoken emotion recognition using glottal symmetry. *EURASIP Journal on Advances in Signal Processing*, 1, 11.
 65. Li, Y., & Zhao, Y. (1999). *Recognizing emotions in speech using short-term and long term features*. Budapest: Eurospeech.
 66. Wang, Y., Li, B., Meng, Q., & Li, P. (2009). *Emotional feature analysis and recognition in multilingual speech signal*. Beijing: Electronic Measurement and Instruments (ICEMI).
 67. Vidrascu, L., & Devillers, L. (2007). *Five emotion classes detection in real-world call center data: The use of various types of paralinguistic features*. Orsay: LIMSI-CNRS.
 68. Ayadi, M. E., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3), 572–587.
 69. Xie, B., Chen, L., Chen, G. C., & Chen, C. (2007). Feature selection for emotion recognition of mandarin speech. *Journal of Zhejiang University (Engineering Science)*, 41(11), 1816–1822.
 70. Murray, I. R., & Arnott, J. L. (2008). Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. *Computer Speech & Language*, 22(2), 107–129.
 71. McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., & Strcve, S. (2000). Approaching automatic recognition of emotion from voice: A rough benchmark. In *ISCA Workshop on Speech and Emotion*, Belfast.
 72. Schuller, B., Wimmer, M., Mosenlechner, L., Kern, C., Arsić, D., & Rigoll, G. (2008). Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space. In *Proceedings of international conference in acoustic, speech, signal processing*, Las Vegas, NV, 2008, pp. 4501–4504.
 73. Schuller, B., Rigoll, G., & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Proceedings of IEEE international conference in acoustic, speech, signal processing*, New York, 2004, pp. 577–580.
 74. Wu, S., Falk, T. H., & Chan, W. Y. (2009). Automatic recognition of speech emotion using long-term spectro-temporal features. In *16th international conference on digital signal processing*, Santorini-Hellas, 2009, pp. 1–6.
 75. Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., & Deng, Z. (2004). An acoustic study of emotions expressed in speech. In *International conference on spoken language processing*, Jeju Island, Korea.
 76. Vidrascu, L., & Devillers, L. (2005). Real-life emotion representation and detection in call centers data. In J. Tao, T. Tan, & R. Picard (Eds.), *LNCS: ACII* (Vol. 3784, pp. 739–746). Berlin: Springer.
 77. Nakatsu, R., Nicholson, J., & Tosa, N. (2000). Emotion recognition and its application to computer agents with spontaneous interactive capabilities. *Knowledge-Based Systems*, 13, 497–504.
 78. Clavel, C., Vasilescu, I., Devillers, L., Richard, G., & Ehrette, T. (2008). Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6), 487–503.
 79. Ververidis, D., Kotropoulos, C., & Pitas, I. (2004). Automatic emotional speech classification. In *International conference on digital signal processing* (pp. I593–I596). New York: IEEE Press.
 80. Hoque, M. E., Yeasin, M., & Louwerse, M. M. (2006). Robust recognition of emotion from speech. In *Intelligent virtual agents. Lecture Notes in Computer Science* (pp. 42–53). Berlin: Springer.
 81. Chuang, Z. J., & Wu, C. H. (2004). Emotion recognition using acoustic features and textual content. In *Proceedings of IEEE international conference on multimedia and expo, 2004*, Vol. 1, pp. 53–56.
 82. Hoch, S., Althoff, F., McGlaun, G., & Rigoll, G. (2005). Bimodal fusion of emotional data in an automotive environment. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing, 2005*, Vol. 2, pp. 1085–1088.
 83. Yu, F., Chang, E., Xu, Y. Q., & Shum, H. Y. (2001). Emotion detection from speech to enrich multimedia content. In *2nd IEEE Pacific-Rim conference on multimedia*, Beijing, China.
 84. Wang, Y., Du, S., & Zhan, Y. (2008). Adaptive and optimal classification of speech emotion recognition. In *4th international conference on natural computation, 2008*, pp. 407–411.
 85. Zhang, S. (2008). Emotion recognition in Chinese natural speech by combining prosody and voice quality features. In F. Sun et al. (Eds.), *Advances in neural networks. Lecture Notes in Computer Science* (pp. 457–464). Berlin: Springer.
 86. Luengo, I., Navas, E., Hernez, I., & Snchez, I. (2005). Automatic emotion recognition using prosodic parameters. In *INTERSPEECH, Lisbon, Portugal, 2005*, pp. 493–496.
 87. Luger, M., & Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition. In *International conference in acoustic, speech, signal processing*, Honolulu, Hawaii, USA, 2007, Vol. 4, pp. 17–20.
 88. Iliou, T., Anagnostopoulos, C. N. (2009). Statistical evaluation of speech features for emotion recognition. In *4th international conference on digital telecommunications*, Colmar, France, 2009, pp. 121–126.
 89. Lee, C. M., Narayanan, S. S., & Pieraccini, R. (2001). Recognition of negative emotions from the speech signal. In *Proceedings of Automatic Speech Recognition and Understanding workshop*, 2001, pp. 240–243.
 90. Litman, D. J., & Forbes-Riley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48, 559–590.
 91. Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., & Boufaden, N. (2009). Cepstral and long-term features for emotion recognition. In *Proceedings Interspeech*, Brighton, 2009, pp. 344–347.
 92. Vlasenko, B., & Wendemuth, A. (2009). Processing affected speech within human machine interaction. In *Proceedings Annual Conference of the International Speech Communication Association*, 2009.
 93. Shami, M. T., & Kamel, M. S. (2005). Segment-based approach to the recognition of emotions in speech. In *Proceedings of IEEE international conference on multimedia and expo, 2005*, pp. 4–7.

94. <http://www.youtube.com/watch?v=uvhNyAXFTMQ>, “Egypt today show”, Alfaraiin channel, May 25, 2012.
95. http://www.youtube.com/watch?v=S1T_EKDP1R8, “New cairo show”, Al-hayat channel, May 28, 2012.
96. <http://www.youtube.com/watch?v=2v6X2VEjb4k>, “Laka sumt, AIUraify show”, September 12, 2012.
97. <http://www.youtube.com/watch?v=MQv3tKTwm7k>, Zain telecommunication, January 22, 2009.
98. <http://www.youtube.com/watch?v=16qNcn03G3s>, Prince Sultan bin Fahed call, Alriyadiya sport channel, October 6, 2011.
99. <http://www.youtube.com/watch?v=E4TqhBo1SCK>, Althaqafiya channel, January 7, 2012.
100. <http://www.youtube.com/watch?v=Wpf3OxEdJak>, “Dairat al daw’e”, Noon channel, Haifa webhe call, November 19, 2009.
101. <http://www.youtube.com/watch?v=eBzmv9QNU7M>, “Musal-salati show”, Mona Zaki call, June 12, 2011.
102. <https://www.kaggle.com/suso172/arabic-natural-audio-dataset>, Kaggle website for public datasets, 2017.
103. Schiel, F., & Heinrich, C. (2009). Laying the foundation for in-car alcohol detection by speech. In *Proceedings of Interspeech*, Brighton, UK, 2009, pp. 983–986.
104. Schuller, B., & Burkhardt, F. (2010). Learning with synthesized speech for automatic emotion recognition. In *Proceedings of international conference on digital signal processing*, Dallas, TX, 2010, pp. 5150–5155.
105. Burkhardt, F., Schuller, B., Weiss, B., & Weninger, F. (2011). ‘Would you buy a car from me?’—On the likability of telephone voices. In *Proceedings of Interspeech*, Florence, 2011, pp. 1557–1560.
106. Fisher, W., Doddington, G., & Goudie-Marshall, K. (1986). The DARPA speech recognition research database: Specifications and status. In *Proceedings of the DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.
107. Schuller, B., Steidl, S., & Batliner, A. (2009). The interspeech 2009 emotion challenge. In *Proceedings of Interspeech*, Brighton, UK, 2009, pp. 312–315.
108. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., & Muller, C. (2010). The interspeech 2010 paralinguistic challenge. In *Proceedings of Interspeech*, Makuhari, Japan, 2010, pp. 2794–2797.
109. Schuller, B., Steidl, S., Batliner, A., & Krajewski, J. (2011). The interspeech 2011 speaker state challenge. In *Proceedings of Interspeech*, Florence, 2011, pp. 3201–3204.
110. Schuller, B., et al. (2012). The interspeech 2012 speaker trait challenge. In *Proceedings of Interspeech*, Portland, OR, 2012, pp. 254–257.
111. Schuller, B., Valstar, M., Cowie, R., & Pantic, M. (2011). *AVEC 2011—The first international audio/visual emotion challenge* (Vol. 2, pp. 415–424). Berlin: Springer.
112. Schuller, B., Valstar, M., Cowie, R., & Pantic, M. (2012). AVEC 2012—The continuous audio/visual emotion challenge. In *Proceedings of the 2nd international audio/visual emotion challenge and workshop, AVEC, grand challenge and satellite of ACM ICMI*, CA, 2012.
113. Eyben, F., Wöllmer, M., & Schuller, B. (2010). *openSMILE—The munich versatile and fast open-source audio feature extractor*. ACM.
114. <https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php>, “Kruskal-Wallis H Test using SPSS Statistics”, Leard Statistics website.
115. Witten, I. A., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann.
116. Platt, J. C. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines*. Technical Report MSR-TR-98-14, April 1998.
117. Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.
118. Meddeb, M., Karray, H., & Alimi, A. (2016). Automated extraction of features from Arabic emotional speech corpus. *International Journal of Computer Information Systems and Industrial Management Applications*, 8, 184–194.
119. Meddeb, M., Hichem, K., & Alimi, A. (2015). Speech emotion recognition based on Arabic features. In *15th International conference on Intelligent Systems design and Applications (ISDA15)*, Marrakesh, Morocco, IEEE conference, 2015, pp. 14–16.
120. Elliott, C. (1992). *The affective reasoner: A process model of emotions in a multi-agent system*. Ph.D. thesis, Inst. Learning Sciences, Northwestern University, Tech. Rep. 32, 1992.
121. Landweh, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning*, 59(1–2), 161–205.
122. Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832–844.
123. <http://weka.sourceforge.net/doc.dev/allclasses-noframe.html>, Machine Learning Group at the University of Waikato, 2016.
124. Frank, E., Wang, Y., Inglis, S., Holmes, G., & Witten, I. H. (1998). Using model trees for classification. *Machine Learning*, 32(1), 63–76.
125. Aha, D., & Kibler, D. (1991). Instance-based learning algorithms. *Machine Learning*, 6, 37–66.
126. Meddeb, M., Karray, H., & Alimi, A. (2016). Content-based Arabic speech similarity search and emotion detection. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics*, 2016, pp. 530–539.
127. Wu, S., Falk, T. H., & Chan, W. Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53, 768–785.
128. Dellaert, F., Polzin, T., & Waibel, A. (1996). Recognizing emotion in speech. In *Proceedings of ICSLP*, Philadelphia, 1996, pp. 1970–1973.
129. Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E. (2000). Desperately seeking emotions: Actors, wizards, and human beings. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Newcastle, 2000, pp. 195–200.
130. Schuller, B., Muller, R., Lang, M., & Rigoll, G. (2005). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Proceedings of Interspeech*, Lisbon, 2005, pp. 805–809.
131. McGilloway, S., Cowie, R., Doulas-Cowie, E., Gielen, S., Westerdijk, M., & Stroeve, S. (2000). Approaching automatic recognition of emotion from voice: A rough benchmark. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Newcastle, 2000, pp. 207–212.



Samira Klaylat is a Ph.D. student at Beirut Arab University since 2010. She received a B.S. in computer science with distinction and M.S. in information system from Beirut Arab University in 2004 and 2006 respectively. She is a Microsoft Certificated Software Developer since 2006. She worked as a full time web developer and head of I.T. department at GL-Network Development Company in Beirut 2007 till end of 2008. She is currently a part time university

instructor at Lebanese International University since 2009. Her research interest is signal processing and data mining.



Ziad Osman is Professor and Chair of the Department of Electrical and Computer Engineering at Beirut Arab University. He received a B.S. from Beirut Arab University in 1987, and an M.S. and his Ph.D. in Electrical Engineering from the University of Florida at Gainesville in 1991 and 1994, respectively. Afterwards he worked as a postdoc at the University of Florida for a year. Since 1995 he has been a faculty member in the Faculty of Engineering

at Beirut Arab University (BAU). His research interests centered on semiconductor photonics, image/text processing and encryption.



Lama Hamandi received her B.E. degree in Electrical Engineering from the American University of Beirut and her M.S. and Ph.D. degrees from the Ohio State University with a major in Computer Engineering and minor in Computer Science. She taught various courses in several universities in Lebanon and abroad. After that she joined the Electrical and Computer Engineering department in the American University of Beirut. Her research interests

include Parallel Computing, Digital Image Processing, Natural Language Processing and Data mining.



Rached Zantout received his B.E. from The American University of Beirut, Lebanon in 1988, his MSc from the University of Florida in 1990 and Ph.D. from the Ohio State University in 1994, all degrees being in Electrical Engineering. He was a Research Associate and Teaching Associate for most of his graduate studies. Directly after finishing his Ph.D. he joined Scriptel Corporation and worked on several R&D projects to develop a new generation of graphic input devices. Between 10/1995 and 8/2000 Dr. Zantout was an Assistant Professor at King Saud University in Riyadh (Saudi Arabia). Then Dr. Zantout moved to Lebanon and taught as an Assistant Professor at the University of Balamand for the period between 9/2000 and 9/2002. He also worked as a part-time faculty members at reputed Lebanese universities like the American University of Beirut, Lebanese American University and Beirut Arab University. Between 9/2003 and 8/2009 Dr. Zantout was at the Hariri Canadian University where he became an Associate professor at the Electrical and Computer Engineering Department. Between 9/2009 and 9/2012 Dr. Zantout was an Associate professor at the College of Computer and Information Sciences at Prince Sultan University, Riyadh, Saudi Arabia. Between 9/2012 and 9/2014, Dr. Zantout was an Associate Professor at the Mathematics and Computer Science Department of the Faculty of Science at Beirut Arab University, Lebanon. Currently Dr. Zantout is Associate Professor at the Electrical and Computer Engineering Department of the College of Engineering at Rafik Hariri University. Dr. Zantout's research interests cover Robotics, Artificial Intelligence, and Natural Language Processing. He currently works on developing components for Machine Translation and Natural Language Processing with a special focus on tools related to the Arabic Language. He also has active research in the area of autonomous robot navigation, Computer Vision, Digital Image Processing and Embedded Systems Design.

er of graphic input devices. Between 10/1995 and 8/2000 Dr. Zantout was an Assistant Professor at King Saud University in Riyadh (Saudi Arabia). Then Dr. Zantout moved to Lebanon and taught as an Assistant Professor at the University of Balamand for the period between 9/2000 and 9/2002. He also worked as a part-time faculty members at reputed Lebanese universities like the American University of Beirut, Lebanese American University and Beirut Arab University. Between 9/2003 and 8/2009 Dr. Zantout was at the Hariri Canadian University where he became an Associate professor at the Electrical and Computer Engineering Department. Between 9/2009 and 9/2012 Dr. Zantout was an Associate professor at the College of Computer and Information Sciences at Prince Sultan University, Riyadh, Saudi Arabia. Between 9/2012 and 9/2014, Dr. Zantout was an Associate Professor at the Mathematics and Computer Science Department of the Faculty of Science at Beirut Arab University, Lebanon. Currently Dr. Zantout is Associate Professor at the Electrical and Computer Engineering Department of the College of Engineering at Rafik Hariri University. Dr. Zantout's research interests cover Robotics, Artificial Intelligence, and Natural Language Processing. He currently works on developing components for Machine Translation and Natural Language Processing with a special focus on tools related to the Arabic Language. He also has active research in the area of autonomous robot navigation, Computer Vision, Digital Image Processing and Embedded Systems Design.