Speech Communication 122 (2020) 19–30

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features

Lamiaa Abdel-Hamid

Department of Electronics & Communication, Faculty of Engineering, Misr International University, Cairo, Egypt

ARTICLE INFO

Keywords:
Speech emotion recognition
Arabic speech emotion database
Prosodic features
Mel-frequency cepstral coefficients (MFCC)
Long-term average spectrum (LTAS)
Wavelet transform

ABSTRACT

Speech emotion recognition (SER) has recently been receiving increased interest due to the rapid advancements in affective computing and human computer interaction. English, German, Mandarin and Indian are among the most commonly considered languages for SER along with other European and Asian languages. However, few researches have implemented Arabic SER systems due to the scarcity of available Arabic speech emotion databases. Although Egyptian Arabic is considered one of the most widely spoken and understood Arabic dialects in the Middle East, no Egyptian Arabic speech emotion database has yet been devised. In this work, a semi-natural Egyptian Arabic speech emotion (EYASE) database is introduced that has been created from an award winning Egyptian TV series. The EYASE database includes utterances from 3 male and 3 female professional actors considering four emotions: angry, happy, neutral and sad. Prosodic, spectral and wavelet features are computed from the EYASE database for emotion recognition. In addition to the classical pitch, intensity, formants and Mel-frequency cepstral coefficients (MFCC) widely implemented for SER, long-term average spectrum (LTAS) and wavelet parameters are also considered in this work. Speaker independent and speaker dependent experiments were performed for three different cases: (1) emotion vs. neutral classifications, (2) arousal and valence classifications and (3) multi-emotion classifications. Several analysis were made to explore different aspects related to Arabic SER including the effect of gender and culture on SER. Furthermore, feature ranking was performed to evaluate the relevance of the LTAS and wavelet features for SER, in comparison to the more widely used prosodic and spectral features. Moreover, anger detection performance is compared for different combinations of the implemented prosodic, spectral and wavelet features. Feature ranking and anger detection performance analysis showed that both LTAS and wavelet features were relevant for Arabic SER and that they significantly improved emotion recognition rates.

1. Introduction

Humans exhibit various emotions throughout their daily life, such as happiness, anger, disgust, sadness, fear, etc., in response to the different situations they encounter. Emotions tend to have a direct effect on our relationships and interactions with others, as well as on our mental health and decision making. Consequently, researchers in several multidisciplinary domains including psychology, neurology and cognitive science have taken great interest in understanding, studying and detecting human emotions (Swain et al., 2018). Moreover, rapid advancements in artificial intelligence technologies have led to increased interest in affective computing in which systems can recognize and accordingly respond to the different human emotions.

The cognitive appraisal theory states that the way people interpret a specific situation and their judgment about the extent that the situation positively or negatively affects them and meets their goals, determines their emotional reaction to that situation (Thagard, 2019). Nevertheless, emotional states commonly occur in parallel with various physiological changes in bodily functions such as heart rate, breathing rate, brain signals, perspiration, skin temperature, hormone levels, facial expressions, voice, etc. Emotions can thus be defined as being complex elicited mental states associated with physiological (bodily) responses. Physiological signals such as electrocardiography (ECG), electromyography (EMG), galvanic skin response (GSR), respiration rate (RR), electroencephalography (EEG), as well as facial expressions have been successfully used to detect different emotional states (Mohammadi et al., 2017).

Recently, speech signals have also been shown to convey information relevant to the emotion of the speaker (Tawari and Trivedi, 2010) with the advantage of being more easily recorded than other physiological signals that require special equipment and settings. Speech emotion recognition (SER) has thus been gaining increased attention as well as being adopted in several applications including criminal investigations, robot interactions, computer games, smart TVs and call centers (Khalil et al., 2018; Meddeb et al., 2017; Sinith et al., 2016). Moreover,

E-mail address: lamiaa.a.hamid@miuegypt.edu.eg

https://doi.org/10.1016/j.specom.2020.04.005
Received 11 June 2019; Received in revised form 13 March 2020; Accepted 28 April 2020
Available online 22 May 2020

detecting emotions from speech can be useful for psychological medical diagnosis (Kamińska and Pelikant, 2012; Likitha et al., 2018).

Emotion recognition from speech can, however, be a challenging task due to the somewhat ambiguous nature of emotions in addition to their variability across different cultures, languages and genders. There are two popular approaches for emotion representation: categorical and dimensional (Alarcao and Fonseca, 2017). The categorical approach indicates a set of basic emotions that are universal among all humans regardless of their culture. In the 20th century, Paul Ekman suggested there are six basic emotions: anger, disgust, fear, happiness, sadness and surprise (Ekman, 1999). Another work by Robert Plutchik however identified eight basic emotions namely anger, anticipation, disgust, fear, happiness, trust, sadness and surprise (Plutchik, 1991). On the other hand, the dimensional approach categorizes emotions based on cognition into an n-dimensional space. The two-dimensional circumplex emotional model is the most commonly used in which emotions are describe based on valence (positive or negative) and arousal (intensity) (Posner, Russell and Peterson, 2005). In SER research, basic emotions introduced by the categorical approach are the most widely considered alongside neutral speech (Mustafa et al., 2018).

In term of languages, most speech emotion databases implemented consider European languages such as English, German and Spanish (El Ayadi et al., 2011; Mustafa et al., 2018). Recently, an increasing number of emotional speech databases in Asian languages such as Mandarin, Telegu, Japanese, Hindi, Malay, Persian and Korean have also been emerging. However, African speech emotion databases are scarcely available in literature (Mustafa et al., 2018). Specifically for the Arabic language, extremely limited speech emotion databases exists despite it being one of the six official languages of the United Nations and being spoken by over 400 million persons in the Arab world ("Arabic Population").

In this work, an Egyptian Arabic speech emotion database is presented that includes four different emotions: angry, happy, neutral and sad. The introduced database includes a total of 579 speech utterances for 3 male and 3 female subjects. Prosodic, spectral and wavelet features are computed from the different speech utterances. Specifically, features implemented in this work include a combination of the widely used prosodic and spectral features, in addition to long-term average spectrum (LTAS) and wavelet features which are being used for the first time for Arabic SER. Several binary and multi-emotion classification experiments are performed considering both the categorical and dimensional emotion models. Furthermore, classification results are reported for the cases of speaker independent and speaker dependent emotion recognition. Moreover, feature ranking is employed in order to study the relevance of the LTAS and wavelet features for SER and compare it to those of the more commonly implemented prosodic and spectral features.

The rest of the paper is divided as follows: Section 2 summarizes relevant SER literature. Section 3 mentions the details of the introduced Egyptian Arabic speech emotion database. Section 4 gives a detailed description of the implemented prosodic, spectral and wavelet features. Section 5 illustrates the results from the different performed SER experiments. Section 6 discusses the presented results, evaluates the relevance of the implemented features as well as compare performance of anger detection for different combinations of the implemented features. Finally, Section 7 wraps up with conclusions.

2. Literature review

Speech emotion recognition systems vary among them in the database characteristics (language of speech, how emotions were induced, number of considered emotions) as well as in the computed features from the speech signal and the implemented classifier for emotion recognition. Prosodic features, such as pitch and intensity, along with spectral features, specifically Mel-frequency cepstral coefficients (MFCC) and linear prediction coefficients (LPC) based features, are the most commonly implemented features for SER (Mustafa et al., 2018; Palo and Mohanty, 2018; Swain, Routray and Kabisatpathy, 2018 ; Vogt and André, 2006). In term of classifiers, support vector machine (SVM) and neural networks (NN) are the most widely used in speech emotion recognizers (Mustafa et al., 2018). Other implemented classifiers include hidden Markov models (HMM), k-nearest neighbor (kNN), decisions trees, Naïve Bayes (NB) and Gaussian mixture model (GMM).

Table 1 gives a brief summary of SER research including database (name, language and type), considered emotions, classifier, features and recognition rates. More SER comprehensive surveys can be found in Koolagudi et al. (2018), Mustafa et al. (2018) and Akçay and Oğuz (2020).

3. Egyptian Arabic emotion speech (EYASE) database

Speech emotion databases can be categorized into natural, semi-natural, acted and elicited based on how emotions are induced (Koolagudi et al., 2018; Mustafa et al., 2018). Natural emotion databases consider spontaneous speech that are typically attained from radio programs, television programs or call centers. Semi-natural emotion databases are collected from emotional scenes in movies or series (Koolagudi et al., 2018). Acted emotion databases rely on professional actors deliberately simulating specific predefined emotions for the sake of data collection. Elicited emotion databases, on the other hand, are collected by inducing the required emotions by people with no professional acting experience. Natural and semi-natural speech databases typically have a wide variability in the recorded speech as they consist of as many phrases as the number of utterances within the database. On the other hand, acted and elicited emotion databases are recorded for a specific set of phrases that may or may not be repeated per emotion, and that rarely exceed 15 different phrases.

Most existing SER research consider acted or elicited emotion databases (Klaylat et al., 2018). Acted and elicited emotion databases rely on self-reported emotions as opposed to the emotion labeling required in natural and semi-natural emotion databases. Furthermore, creating natural and semi-natural speech emotion databases can be an extremely hard task, specifically when considering several emotions, as it requires analyzing, recording and labeling a large amount of conversation. Consequently, acted and elicited emotion databases are relatively easier to collect. Moreover, acted and elicited emotion databases are recorded in noise-free environments which makes feature extraction simpler and more reliable than in the case of natural and semi-natural emotion databases. Nevertheless, whether or not the acted or elicited emotions in these databases are realistic is a controversial issue as it is argued how far emotions can be simulated in controlled lab settings (Mustafa et al., 2018). Natural and semi-natural emotion databases have thus been gaining increased interest in the last few years.

Although a large number of speech emotion databases are available for various European and Asian languages, very limited Arabic speech emotions databases exist in literature (Mustafa et al., 2018). Among the Arabic databases, Meftah et al. (2015) built the KSUEMotions elicited database which was recorded in Modern Standard Arabic spoken by Syrian, Saudi and Yemen native speakers. The KSUEMotions database includes five different emotions: angry, happy, neutral, sad and surprised. Furthermore, two Arabic natural speech emotion databases have been introduced in literature that were recorded for different Arabic dialects. The first was recorded from TV shows to include angry, happy and surprised emotions (Klaylat et al., 2018), whereas the second was recorded from customer service phone calls along with from TV shows to consider angry and neutral emotions (Khalil et al., 2018). Among the challenges of Arabic speech emotion recognizers, thus, is the limited available databases in addition to the wide variability between the spoken Arabic dialects in the different Arabic countries.

Egyptian Arabic is among the most popular Arabic dialects, spoken by nearly 100 million Egyptian citizens ("Egyptian Arabic"). Moreover,

Table 1
Speech emotion recognition literature review.

Reference	Database			Emotions	Classifier	Features	Recognition Rate
	Name	Language	Type				
(Esmaileyan and Marvi, 2014)	BerlinEmo	German	Acted	anger, fear, happiness, neutral and sadness (5)	Linear Discriminant Analysis (LDA)	prosodic (181), spectral (2280)	*Males*: 73.40%, *Females*: 78.64%
	PDREC	Persian	Semi-natural (radio programs)				*Males*: 47.28%, *Females*: 55.74%
(Lalitha et al., 2015b)	BerlinEmo	German	Acted	anger, boredom, disgust, fear, happiness, neutral, sadness (7)	SVM	pitch, intensity, entropy, jitter, shimmer, autocorrelation, zero crossings, harmonic to noise ratio	81.13%
(Ram and Ponnusamy, 2014)	BerlinEmo	German	Acted	anger, fear, happiness, neutral, sadness (5)	SVM	pitch, intensity, MFCC	63.8%
	–	Tamil (Indian)	Natural				71.3%
(Lalitha et al., 2015a)	BerlinEmo	German	Acted	anger, boredom, disgust, fear, happiness, neutral, sadness (7)	Artificial neural networks (ANN)	MFCC, frequency scaled MFCC, cepstrum	85.7%
(Meftah, Selouani and Alotaibi, 2015)	KSU Emotions	Modern Standard Arabic	Elicited	anger, happiness, neutral, sadness, surprise (5)	Multilayer Perception	pitch, intensity, formants, speech rate	*Males*: 83.33%, *Females*: 55.67%, *Both*: 78.33%
(Sinith et al., 2016)	BerlinEmo	German	Acted	anger, happiness, neutral, sadness (4)	SVM	pitch, intensity, MFCC	*Males*: 67.5 %, *Females*: 70%, *Both*: 75%, 61.25%
	SAVEE	English	Acted				*Females*: 95.83%
	–	Malayalam	Elicited				
(Swain, Routray, Kabisatpathy and Kundu, 2017)	Odia Language Database	Cuttacki Sambalpuri Berhampuri	Elicited	anger, disgust, fear, sadness, happiness, surprise (6)	SVM	Prosodic (72)	75.27% 78.34% 70.17%
(Chatterjee et al., 2018)	TESS	English	Elicited	anger, happiness, neutral (3)	SVM	intensity, zero crossing rate, formants, MFCC, spectral centroid, short term energy	91.3%
(Khalil, Al-Khatib, El-Alfy and Cheded, 2018)	–	Arabic	Natural (call-center, TV shows)	anger, neutral (2)	SVM	pitch, intensity, formants, MFCC	87.24%
(Klaylat, Osman, Hamandi and Zantout, 2018)	–	Arabic (Egyptian, Gulf, Jordan Lebanese)	Natural (talk shows)	anger, happiness, surprise (3)	Sequential minimal optimization (SMO)	pitch, intensity, zero crossing rate, probability of voicing, MFCC, linear spectral frequency (LSP)	95.52%
(Koolagudi, Murthy and Bhaskar, 2018)	–	Telugu	Semi-natural (movies)	anger, fear, happiness, neutral, sadness (5)	GMM	Pitch, intensity, jitter, shimmer, formants, MFCC	84.78%
(Hifny and Ali, 2019)	KSU Emotions	Modern Standard Arabic	Elicited	anger, happiness, neutral, sadness, surprise (5)	Deep NN	MFCC	85.0%

Table 2
Description of the basic emotions considered in the EYASE database.

Emotion	Description	Arousal-Valence
Anger (A)	triggered when a person feels physically or psychologically harmed or threatened. Anger can thus be caused by insults, attacks or frustration.	negative valence high arousal
Happiness (H)	feeling enjoyed by a person. The most common reported causes of happiness are social relationships followed by success.	positive valence high arousal
Sadness (S)	painful emotion associated with failure to achieve a goal or loss of a person/object to which one is attached to.	negative valence low arousal

Table 3
EYASE database details.

	Angry	Happy	Neutral	Sad	Total
Male 1	30	30	30	27	117
Male 2	30	20	30	30	110
Male 3	30	22	30	30	112
Female 1	20	20	20	20	80
Female 2	20	20	20	20	80
Female 3	20	20	20	20	80
Total	150	132	150	147	579

Egyptian dialect is commonly recognized by Arabic speakers due to the huge popularity of Egyptian movies, TV series and songs within the Arab world. To the best of the author's knowledge, no Egyptian Arabic emotion database has been previously developed for speech emotion recognition.

In this work, an Egyptian Arabic semi-natural emotion speech database is created from the award winning Egyptian drama series Hatha Al-Masaa (هذا المساء) ("Hatha Almasaa"). Four basic emotions were considered in the introduced Egyptian Arabic speech emotion (EYASE) database: angry (A), happy (H), neutral (N) and sad (S). The EYASE database was recorded for three male and three female lead professional actors. At the time of filming, the actors were within the age range from 22 to 45 years old and had between 12 and 22 years of professional experience, with the exception of the youngest female actor who had about six years of acting experience. Initially, sound clips were recorded and labelled based on visual, audio and story narrative as well as on the depicted actor emotion. The sound clips labeling into angry, happy or sad was based on the Ekman description of the basic emotions (Ekman and Cordaro, 2011; Universit and Strack, 1991) summarized in Table 2, whereas the neutral case considered normal speech which is void of basic emotions. Next, context unaware annotation based solely on the recorded sound clips was performed by two different labelers. Finally, only utterances on which all labelers agreed on were included in the EYASE database. In total, the EYASE database includes 579 utterances. All speech samples were recorded using the open source Audacity software (https://www.audacityteam.org/) at a sampling rate of 44.1 kHz where a single utterance duration ranged from 1 to 6 s. The EYASE database can be provided for research purposes upon request from the following link: https://www.researchgate.net/publication/341001383_EYASE_Database. Table 3 summarizes the details of the EYASE database.

4. Methods

Fig. 1 shows the flow diagram of the implemented SER system. Speech emotion recognition preliminary relies on the extraction of a set of features from the speech signal. Next, the computed feature set is input into a classifier whose task would be to identify the different emotions. In this work, a combination of prosodic, spectral and wavelet features were extracted from the introduced EYASE database for emotion recognition. In addition to the commonly implemented prosodic (pitch-intensity) and spectral (formants – MFCC) features, LTAS and wavelet parameters were also included in the feature vector. The details of the computed features are given in this section.

4.1. Prosodic features

Prosodic features are acoustic parameters computed from the speech utterance. They include the speech pitch and intensity. Pitch is the fundamental frequency (f_0) of the speech signal created by the vibration of the speaker's vocal cords. Pitch ranges tend to vary among individuals as well as for different emotions (Kostoulas and Fakotakis, 2006; Ververidis and Kotropoulos, 2006). For example, male adults generally have lower pitch ranges than female adults. Also, the anger emotion typically has higher pitch ranges relative to other basic emotions (Ververidis and Kotropoulos, 2006). In the present study, pitch was computed based on the autocorrelation method (Boersma, 1993) considering a pitch range of 75–300 Hz for males and 100–500 Hz for females. Intensity (energy) refers to the loudness of the speech signal, and is thus related to the degree of arousal of the speaker. Both pitch and intensity were found to be highly correlated with the speaker's emotion and hence are widely used in SER (Meftah et al., 2015).

Pitch and intensity contour statistical features, including the mean, standard deviation, maximum, minimum and range, were computed for emotion detection. Jitter and shimmer were also considered, which measure the variation in pitch and intensity, respectively.

4.2. Spectral features

Spectral features computed in this work include the formants, MFCC and LTAS. Formants correspond to the resonance frequencies of the human vocal tract system at which high energy peaks occur. Formants tend to change with emotion variation, and are hence commonly used for SER (Koolagudi et al., 2018). MFCC features are computed from the nonlinear Mel-scale which emphasizes lower frequency components over higher ones. They are widely implemented in speech and speaker recognition systems as well as in SER, as they mimic the perception of the human auditory system by being more sensitive to sound variations at lower tones. LTAS represents the logarithmic signal power density of the voiced parts of the signal while correcting away the pitch influence (Boersma and Kovacic, 2006). LTAS features also have the advantage of being computationally inexpensive relative to the MFCC features (Kinnunen et al., 2006). Despite being commonly implemented for speech analysis (Bahmanbiglu et al., 2017; Fletcher et al., 2017;

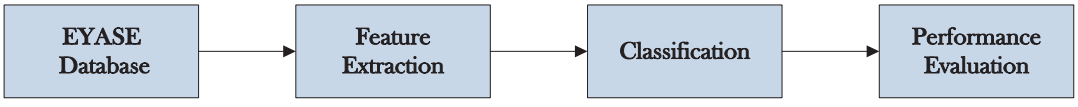

Fig. 1. Introduced Arabic SER flow diagram.

Muckenhirn et al., 2017; Yüksel and Gündüz, 2018), LTAS features were scarcely used for SER.

Spectral features computed include the first three formant, the mean of the first twelve Mel-frequency cepstral coefficients, along with the LTAS mean, standard deviation, slope, maximum, minimum and range.

4.3. Wavelet features

Wavelet transform (Mallat, 1989) is a multiresolution technique commonly used for analysis and processing of acoustic signals (Haridas et al., 2018; Tirumala, Shahamiri, Garhwal and Wang, 2017). Wavelet transform decomposes the speech signal by passing it through low and high pass filters resulting in approximation and detail coefficients, respectively. Wavelet decomposition has the advantage of being localized in both time and frequency. Moreover, wavelet decomposition of the subsequent approximation coefficients can be performed resulting in different scales of the analyzed signal.

In this work, four level wavelet decomposition was performed using the Daubechies4 (db4) wavelet, then wavelet energy and entropy (Coifman and Wickerhauser, 1992) of the detail and approximation subbands from the four wavelet decompositions were computed.

In literature, pitch, intensity, formants and MFCC features have been widely and successfully implemented for SER in different languages including Arabic. However, very limited SER research considered LTAS (Eyben et al., 2015) or wavelet features (Han and Wang, 2013; Joshi and Zalte, 2013; Krishna Kishore and Krishna Satish, 2013; Saste and Jagdale, 2017). Moreover, to the best of the author's knowledge neither LTAS nor wavelet features were previously used for Arabic SER. Accordingly, among the contributions of this work is the implementation of the LTAS and wavelet features for Arabic SER along with the more classical prosodic and spectral features. In Section 6 feature ranking is performed to explore the usefulness of the LTAS and wavelet features for Arabic SER and to compare their relevance for Arabic SER with respect to the more widely implemented prosodic and spectral features.

The final feature vector used in this work includes a total of 49 prosodic, spectral and wavelet features. Prosodic and spectral features were computed using the Praat software (Boersma and Weenink, 2018), whereas MATLAB (Mathworks, Inc., Natick, MA, USA) was used to calculate the wavelet features. Arabic speech emotion classification results using the proposed feature vector are presented in the next section for the introduced EYASE database.

5. Results

Most SER literature consider speaker independent systems which generally are more challenging than speaker dependent systems. Nevertheless, speaker dependent emotion recognition can be useful for a wide range of applications including psychiatric diagnosis, as well as human interactive machines (e.g. games, robots, etc.…) which can be easily trained to recognize emotions for a specific user.

In this section, classification results are presented for both speaker independent and speaker dependent Arabic SER using the EYASE database. For each case, the following analyses were performed:

(a) **Emotion Classifications**: Experiments were performed to identify each of the angry, happy and sad basic emotions versus neutral speech.

Table 4
Speaker independent emotion versus neutral classification accuracies (%).

	SVM			kNN		
	Males	Females	Both	Males	Females	Both
Angry	95.0	85.8	90.3	92.2	83.3	89.0
Happy	70.4	69.2	66.3	66.7	66.7	64.9
Sad	86.4	80.0	81.5	85.3	74.2	79.1

(b) **Arousal & Valence Classifications**: Angry/sad and angry/happy binary classifications were made in order to consider arousal and valence, respectively.

(c) **Multi-emotion Classifications**: Classifier was built to separate angry, happy, neutral and sad (AHNS) emotions. Also, angry, neutral and sad (ANS) multi-emotion classification was considered.

Support vector machine is among the most commonly used classifiers for SER owing its good performance and rapid training speed (Koolagudi et al., 2018; Mustafa et al., 2018; Özseven, 2019). The kNN classifier is considered a simple and efficient classifier that is easy to tune for good performance. For the speaker independent experiments, an SVM classifier with radial basis function (rbf) kernel as well as the kNN classifier were considered. For the speaker dependent experiments, SVM with linear kernel was used as it easy to tune while giving reliable performance (Sinith et al., 2016). All experiments were performed using 10 fold cross validated classifications in the Weka Platform (Hall et al., 2009). For the SVM classifier, the cost and gamma parameters were tuned for best performance. As for the kNN classifier, the k parameter was varied over the range from 1 to 15 and best classification results were reported.

5.1. Speaker independent experiments

Speaker independent emotion recognition is performed assuming no previous knowledge of the identity of the subjects. In order to study whether gender differences would affect the SER performance, all experiments were carried out for each gender separately as well as for both male and female subjects combined. Tables 4-6 summarize the speaker independent SER results for emotion, arousal & valence and multi-emotion classifications, respectively for both the SVM (rbf) and kNN classifiers.

Table 4 illustrates the speaker independent classification accuracies for each of the angry, happy and sad emotions versus neutral speech for males, females and both. Experiments show better recognition rates by the SVM classifier as compared to the kNN classifier. Results also showed that the angry emotion was detected with the highest accuracies (SVM: 85.8–95%) followed by the sad emotion (SVM: 80–86.4%), whereas the happy emotion was the hardest to detect (SVM: 66.3–70.4%). Generally, male SER gave better or just as good accuracies than female SER. For the SVM classifier, emotion recognition was better for males than for females by 9.2%, 6.4% and 1.2% for the angry, sad and happy emotions respectively.

Table 5 summarizes the results for the arousal (angry/sad) and valence (angry/happy) classifications. Generally, arousal detection gave better accuracies than valence detection where arousal recognition re-

Table 5
Speaker independent arousal and valence classification accuracies (%).

	SVM			kNN		
	Males	Females	Both	Males	Females	Both
Arousal (Angry/Sad)	97.2	94.2	94.3	94.9	87.5	93.3
Valence (Angry/Happy)	89.5	80.8	86.2	88.3	77.5	83.7

Table 6
Speaker independent multi-emotion classification accuracies (%).

	SVM			kNN		
	Males	Females	Both	Males	Females	Both
AHNS	69.9	66.3	66.8	66.4	55.4	61.7
ANS	84.3	77.2	81.0	80.5	71.7	78.3

Table 7
Speaker dependent emotion versus neutral classification accuracies (%).

	Male 1	Male 2	Male 3	Female 1	Female 2	Female 3
Angry	96.7	98.3	95.0	85.0	97.5	80.0
Happy	58.3	74.0	75.0	67.5	72.5	60.0
Sad	86.0	88.3	93.3	85.0	80.0	72.5

sults were between 7.7–13.4% higher than valence recognition results. For the SVM classifier, the arousal recognition accuracies were all found to be higher than 94%, whereas valence recognition accuracies were within the range from 80–90%. Classification performance for the male subjects was also found to be more superior to performance for female subjects by approximately 3% and 9% for arousal and valence, respectively when considering the SVM classifier.

Table 6 shows the results for the multi-emotion classification experiments. Considering the angry, happy, neutral and sad emotions (AHNS), better performance was achieved by the SVM classifier where recognition rates of 69.9%, 66.3% and 66.8% were achieved for males, females and both, respectively. Results in Table 4 have shown that the happy emotion was the most challenging to separate from neutral speech. By omitting the happy emotion and considering only the angry, neutral and sad (ANS) emotions, recognition rates were found to significantly improve to become 84.3%, 77.2% and 81.0% for males, females and both, respectively using the SVM classifier. As in the previous experiments, male SER was observed to result in higher accuracies than female SER.

5.2. Speaker dependent experiments

Speaker dependent emotion recognition is performed assuming previous knowledge of the identity of each of the subjects. The EYASE database includes utterances from 3 male and 3 female subjects. In this section, the speaker dependent experiments were performed for each of the six different subjects. Tables 7–9 summarize the speaker dependent SER results for the emotion, arousal & valence and multi-emotion classifications, respectively for the SVM (linear) classifier.

Table 7 demonstrates the speaker dependent classification accuracies for each of the angry, happy and sad emotions versus neutral speech for the six subjects within the EYASE database. Similar to the speaker independent experiments, best accuracies were achieved for the angry emotion (males: 95–98.3%, females: 80–97.5%), followed by the sad emotion (males: 86–93.3%, females: 72.5–85.0%) while least accuracies were attained for the happy emotion (males: 58.3–75%, females: 60–72.5%).

Table 8 illustrates the arousal (angry/sad) and valence (angry/happy) classification results. Arousal classification accuracies (males: 96.5–100%, females: 92.5–95%) were found superior to those the valence accuracies (males: 82.7–96.7%, females: 72.5–90%). However for male1, arousal and valence classification accuracies were found to be very close. Moreover, it is observed that arousal was more readily detected for male subjects (96.5–100%) than for female subjects (92.5–95%).

Table 9
Speaker dependent multi-emotion classification accuracies (%).

	Male 1	Male 2	Male 3	Female 1	Female 2	Female 3
AHNS	64.9	76.4	80.4	66.3	62.5	62.5
ANS	85.1	91.1	91.1	81.7	81.7	73.3

Table 9 summarizes the speaker dependent multi-emotion classification results. For the AHNS classifications, accuracies for the different male subjects (64.9–80.4%) were significantly superior to accuracies for the female subjects (62.5–66.3%). Classification results were generally improved when the happy emotion was omitted. ANS classifications resulted in accuracies within the range from 85.1–91.1% and 73.3–81.7% for the different male and female subjects, respectively.

Experiments performed in this section show that the SVM classifier achieved consistently better results than the kNN classifier which is in agreement with previous literature (Dahake et al., 2017; Özseven, 2019). Furthermore for both speaker independent and speaker dependent SER, the anger emotion was found to be the most readily recognized whereas happiness was the most challenging. Also for both cases, arousal detection resulted in better performance than valence detection and three emotion classifications (ANS) yielded higher recognition rates than four emotion classifications (AHNS). However, speaker dependent SER tends to generally outperform speaker independent SER which is in agreement with findings in previous literature (Rybka and Janicki, 2013). Moreover for all experiments, male emotion recognition resulted in significantly better performance than female emotion recognition. Further analysis of these results is given in the next section.

6. Discussion

Arabic SER is a relatively new field of research due to the scarcity of available databases. Generally, acted and elicited emotion databases are the most and widely used for speech emotion detection as they are easier to record and label. However, the authenticity of emotions within these databases is controversial as it is argued how far emotions can be simulated in controlled lab settings (Mustafa et al., 2018). Recent studies are hence starting to rely more on natural and semi-natural databases which include more realistic emotions (Mustafa et al., 2018). However, emotions in natural and semi-natural speech can be more difficult to detect compared to acted and elicited emotions (Sinith et al., 2016).

In this work, a semi-natural Egyptian Arabic speech emotion (EYASE) databases was introduced that includes four different emotions: angry, happy, neutral and sad. Several classification experiments were performed to detect emotions: (1) emotion vs. neutral classifications, (2) arousal & valence classifications and (3) multi-emotion classifications for both speaker independent and dependent experiments. In the next subsections, results attained within the performed experiments are discussed.

6.1. Arabic SER performance analysis

Multi-emotion classifications results summarized in Tables 6&9 showed that for the EYASE database, the emotion recognition rate was affected by the considered number of emotions. In order to evaluate

Table 8
Speaker dependent arousal and valence classification accuracies (%).

	Male 1	Male 2	Male 3	Female 1	Female 2	Female 3
Arousal (Angry/Sad)	96.5	100	98.3	92.5	95.0	95.0
Valence(Angry/Happy)	96.7	96.0	82.7	72.5	90.0	87.5

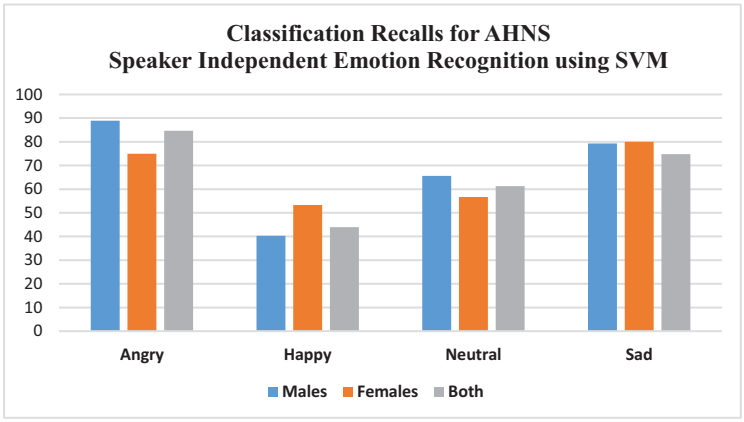

Fig. 2. Classification recalls for AHNS speaker independent emotion recognition using SVM.

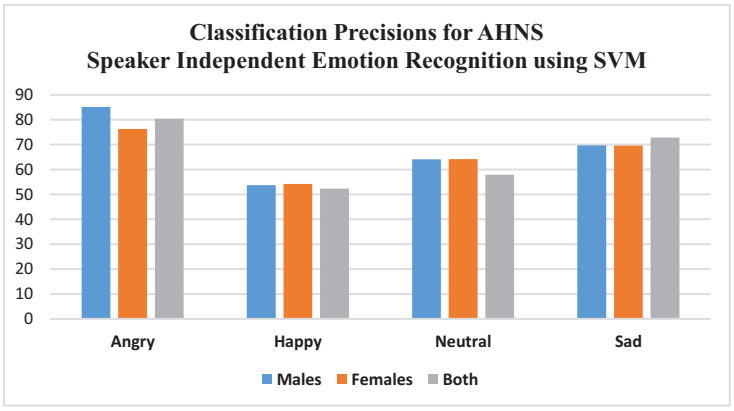

Fig. 3. Classification precisions for AHNS speaker independent emotion recognition using SVM.

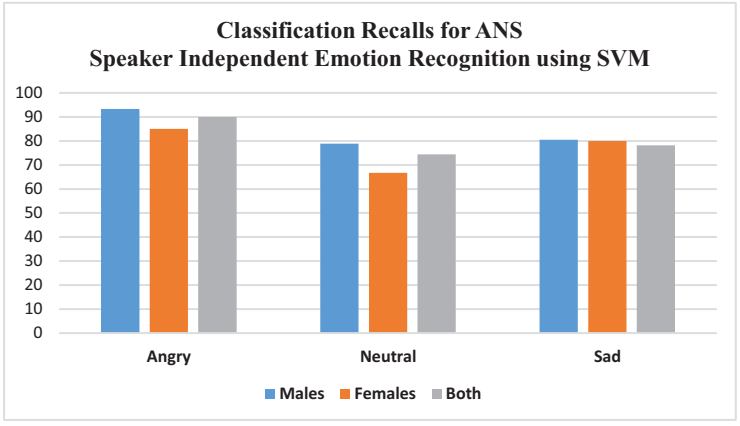

Fig. 4. Classification recalls for ANS speaker independent emotion recognition using SVM.

which emotion was more challenging to detect in a multi-emotion classifications scenario, the recall and precision were computed for each of the different emotions considering three cases: males, females and both genders. Figs. 2 & 3 demonstrate the AHNS speaker independent recall (rec) and precision (prc) SVM classification results, respectively for each of the four different emotions. Results show that the anger emotion (Rec: 75–88.9%, Prc: 76.3–85.1%) was the easiest to detect whereas the happiness emotion (Rec: 40.3–53.3%, Prc: 52.3–54.2%) was the most challenging. Omitting the happiness emotion for the multi-emotion classification thus led to significant improvement in classification results. For ANS emotion detection, Figs. 4 & 5 show that the anger emotion (Rec.: 85–93.3%, Prc: 79.7–91.3%) remained easier to detect than the sadness (Rec: 78.2–80.5%, Prc: 76.2–79.5%) and neutral (Rec: 66.7–78.9%, Prc: 75.5–81.6%) emotions.

Overall, few Arabic SER systems were implemented in literature owing to the limited availability of Arabic speech emotion databases. Khalil et al. (2018) focused their research on anger detection from a natural Arabic database including different dialects, that was recorded from call centers and TV shows. Several experiments were performed and best accuracy of 87.24% was achieved considering both prosodic and MFCC features. In this work, the anger emotion was detected at an accuracy of 90.3% and an average accuracy of 96.7% for the speaker independent and speaker dependent experiments, respectively. Other works have considered various emotions for Arabic SER. Meftah et al. (2015) introduced KSUEmotions which is a Modern Standard Arabic elicited database that includes utterances for five different emotions: anger, happiness, neutral, sadness and surprise. Best reported results for emotion recognition were 83.3%, 55.7% and 78.3% for male, female

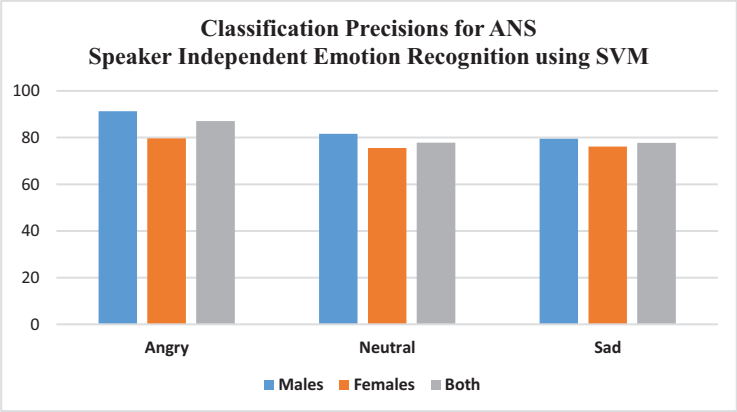

Fig. 5. Classification precisions for ANS speaker independent recognition using with SVM.

and both, respectively. Later work by Hifny and Ali (2019) reported a total accuracy of 85% for the same database. However, elicited speech emotion databases are recorded in controlled lab environments, thus may result in overly optimistic performance. For the EYASE database presented in this study, speaker independent experiments resulted in accuracies of 84.3%, 77.2% and 81% considering the anger, neutral and sadness emotions for male, female and both, respectively.

The TV series ("Hatha Almasaa") from which the EYASE database was recorded is a serious drama discussing social issues within the Egyptian community. Consequently, the anger and sadness emotions were more intensely expressed by the actors in the different scenes. On the other hand, the happiness emotion was usually mild, as it was commonly associated with family gathering, appreciation or happy moments occurring in the midst of the main dramatic storyline. The observations that anger and happiness are the easiest and hardest emotions to detect, respectively is similar to that given in (Meftah et al., 2015) for a Modern Standard Arabic SER corpus (KSUEmotions). Meftah et al. have found that considering angry, happy, neutral, sad and surprised emotions, anger was the most readily recognized emotion whereas happiness was the most challenging. Moreover, other Arabic SER works in literature have also shown that the anger emotion was commonly associated with the highest recognition rates in comparison to other basic emotions (Klaylat et al., 2018; Meddeb et al., 2017).

In general, emotion expression depends on a variety of factors including culture, gender, age and social roles (Brody, 2009; Sagha, Deng and Schuller, 2018). Middle Eastern cultures are considered honor cultures (Pudalov, 2016) where threats or insults provoke extremely intense anger emotions expressed with severe aggressiveness and excessive shouting. Although anger suppression is recommended and honored by religion (Pudalov, 2016), cultural norms seem to override religious ones within the Middle Eastern countries where in most cases, expressing extreme anger is considered as a sign of dominance leading to the submission of people to which this anger is targeted. Overblown anger emotions in honor cultures can thus rationalize the superiority of the anger emotion recognition rate in this work as well as in (Meftah et al., 2015). Several papers considering different cultures have reported results indicating that other emotions were more reliably detected than the anger emotion. In (Joshi and Zalte, 2013) in which emotion recognition was performed using a Marathi speech database (one of the Indian languages), among the five considered emotions (angry, bored, happy, neutral and sad), anger detection achieved the third highest accuracy after sadness and boredom. In (Pan, Shen and Shen, 2012), anger detection was second best after sadness when considering seven emotions (angry, bored, disgust, fear, happy, neutral and sad) for German speech. Nevertheless, in some cases the anger emotion was also found to achieve the highest accuracy rates (Palo and Mohanty, 2018; Swain, Routray, Kabisatpathy and Kundu, 2017). However, in others works results varied depending on the implemented features and/or classifier for emotion recognition (Cao, Verma and Nenkova, 2015 ; Dahake et al., 2017; Krishna Kishore and Krishna Satish, 2013; Sinith et al., 2016).

With respect to gender, SER for male subjects was found in this work to significantly outperform SER for female subjects. Such finding is in agreement with the results in (Meftah et al., 2015) who reported that for Arabic speakers, average accuracies for males and females were 83.33% and 56.67%, respectively. However, other SER works considering Western or Asian subjects have reported opposite findings where emotion recognition rates were higher for female subjects than for male subjects (Esmaileyan and Marvi, 2014; Sinith et al., 2016). Several researches and meta-analysis for American and European cultures have concluded that overall, women show greater emotion expression than men (Chaplin, 2015; Sagha, Deng and Schuller, 2018). Conversely in Eastern Arabic cultures, it is considered more socially acceptable for men to express extreme emotions than it is for women. Women on the other hand are expected to conform to a more subtle profile, specifically within middle class communities. Male expression of anger and happiness for example are highly associated with severe yelling and excited loud speech, respectively which would be considered rude and non-lady like for females. However, few research on the effect of ethnicity on emotion expression in different genders is present in literature since majority of research is conducted considering white Western middle-class subjects (Chaplin, 2015).

6.2. SER feature analysis

Generally, the performance of speech emotion recognizers depends on the relevance of the considered features in addition to the extent by which the speakers express their emotions. Prosodic features are the most commonly implemented features for SER since they were shown to be highly related to expressed emotions (Koolagudi et al., 2018). MFCC features were also shown to be highly relevant for SER, thus are widely used alongside the classical prosodic features (Chatterjee et al., 2018).

In this work, prosodic (pitch – intensity), spectral (formants- MFCC – LTAS) and wavelet (from detail and approximation subbands) features were considered for emotion detection in the EYASE database. Although LTAS features have been commonly used for speech analysis (Bahmanbiglu et al., 2017; Fletcher et al., 2017; Muckenhirn et al., 2017; Yüksel and Gündüz, 2018), they have only been implemented in very limited SER research (Eyben et al., 2015). Wavelet based features have been previously used for SER in early researches (Han and Wang, 2013; Joshi and Zalte, 2013; Krishna Kishore and Krishna Satish, 2013), however they have become less common in recent work (Saste and Jagdale, 2017). Nevertheless to the best of the author's knowledge, both LTAS and wavelet features were not previously considered for Arabic SER.

Feature ranking is used, in this subsection, in order to identify the most relevant features within for each of the performed experiments

Table 10
Most relevant features for emotion vs neutral speech classifications.

	Angry		Happy		Sad	
	Male	Female	Male	Female	Male	Female
Prosodic	Pitch (2) Intensity (1)	Pitch (1) Intensity (1)	**Pitch (3)**	**Pitch** (5)	**Intensity** (3)	–
Spectral	**LTAS** (4)	**LTAS** (4) MFCC (1)	LTAS (2) MFCC (1)	LTAS (1) Formants (1)	LTAS (1) MFCC (2) Formants (1)	LTAS (1) MFCC (1) Formants (1)
Wavelet	–	–	Approx. (1)	–	–	Approx. (1) **Details (3)**

Table 11
Most relevant features for arousal and valence classifications.

	Arousal (angry/sad)		Valence (Angry/happy)	
	Male	Female	Male	Female
Prosodic	Pitch (2) Intensity(2)	Pitch (2) Intensity (1)	Intensity (2)	Intensity (2)
Spectral	**LTAS** (3)	**LTAS** (4)	**LTAS** (5)	**LTAS** (5)
Wavelet	–	–	–	–

Table 12
Most relevant features for multi-emotion classifications.

	AHNS		ANS	
	Male	Female	Male	Female
>Prosodic	Pitch (2)	Pitch (1) Intensity (1)	Pitch (2) Intensity (1)	Pitch (2) Intensity (1)
Spectral	**LTAS** (3)	**LTAS** (3)	**LTAS** (3)	**LTAS** (3) MFCC (1)
Wavelet	Details (2)	Details (2)	Details (1)	–

(emotion vs. neutral – arousal & valence classifications – multi-emotion classifications). Feature ranking was implemented in Weka using correlation feature selection (Hall, 1999) which is based on the intuition that relevant feature subsets contain features that are highly correlated with the predictive class (i.e. emotion), yet uncorrelated to each other. Pearson's correlation measure is thus computed between each feature and the predictive emotion class in order to estimate the relevance of a specific feature with respect to the considered emotion. Pearson correlation has been previously implemented in several SER research (Li and Akagi, 2019; Mencattini et al., 2014; Vogt, André and Bee, 2008), hence considered in the present study. Tables 10–12 show the most relevant features for each of the performed speaker independent analysis (emotion vs. neutral – arousal & valence classifications – multi-emotion classifications) considering both male and female subjects. In each of the tables, the type of feature is indicated followed by the number of statistical features that were among the seven highest ranked of the forty-nine implemented features. Moreover, the feature category including the most relevant feature is shown in bold.

Feature ranking results show that prosodic features (pitch and intensity) were highly significant for all the performed emotion detection experiments, regardless of the number and type of emotions considered in the classification. The MFCC and formant spectral features were specifically relevant for the detection of a specific emotion as opposed to neutral speech. Interestingly, the LTAS features, like prosodic features, were found to be highly relevant for all performed experiments using the EYASE database. Moreover, LTAS features were ranked highest among all other features for the angry, arousal, valence and multi-emotion classifications, by that outperforming the classical prosodic and MFCC features. As for the wavelet features, they were found to be considerably useful specifically within the emotion vs. neutral and multi-emotion classifications. LTAS is typically used in speech analysis to represent the distribution of the average speech energy with respect to frequency (Kacha et al., 2020). LTAS thus has the advantage of providing information on the spectral distribution of the speech signal while reducing the effect of phonetic segment variability (Fletcher et al., 2017; Kacha et al., 2020). Nevertheless, the advantage of the wavelet features is that they separate the low and high frequency speech components in the approximation and detail subbands, respectively thus capturing different spectra information related to the speech utterance. It is hence highly recommended to integrate both LTAS and wavelet features for SER in order to explore their relevance for other speech emotion databases of different languages.

The averages of the prosodic (pitch –intensity) and LTAS (slope – minimum) parameters are illustrated in Figs. 6–9, respectively considering one male and one female speaker from the EYASE database. Fig. 6 shows that for both genders, highest pitches were associated with the anger emotion followed by the happiness, sadness and neutral emotions, respectively. In Fig. 7, the intensity parameter was highest for the anger emotion for both males and females. However, intensity values were inconsistent for the other emotions for the different genders. For males, intensities were higher for angry, happy, neutral, then sad; whereas for females they were higher for angry, sad, happy, then neutral. From Figs. 8 and 9, it can be depicted that the LTAS slope and minimum values were significantly different for the angry emotion with respect to the happiness, neutral and sadness emotions. Nevertheless for all emotions, the LTAS slope and minimum average values differed for the male and female subjects. Parameter statistics thus indicate that the considered emotion features tend to be gender dependent, which is in agreement with the findings in (Koolagudi et al., 2009). Such variances are among the yet unresolved challenges of SER.

In order to further study the relevance of the LTAS and wavelet features in comparison to the classical prosodic, formants and MFCC features, anger detection performance is compared for the different combinations of the prosodic, spectral and wavelet features implemented in this work. Anger detection was chosen since anger was found to be the most readily detected emotion in the EYASE database. Anger detection has several applications from which the most relevant is measuring customer satisfaction in call centers. Manually going through all recorded phone calls would be an extremely tedious and time-consuming task. Anger recognition thus has the potential to help organizations easily monitor and improve their call center's quality of service by automatically detecting calls with angry or frustrated customers for evaluation and follow up (Khalil et al., 2018).

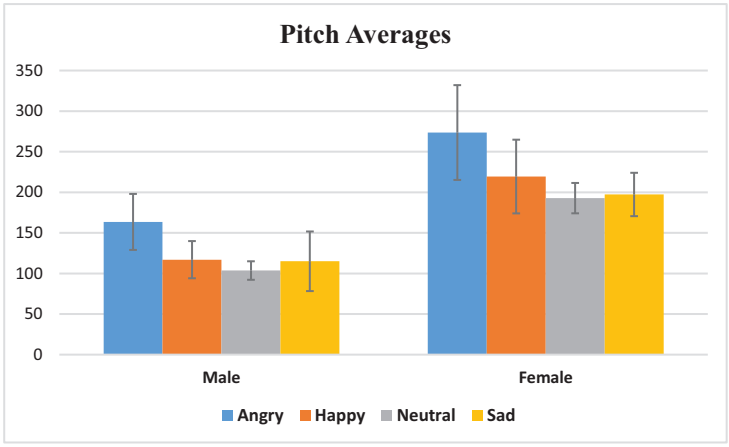

Fig. 6. Pitch parameter averages for the different emotions.

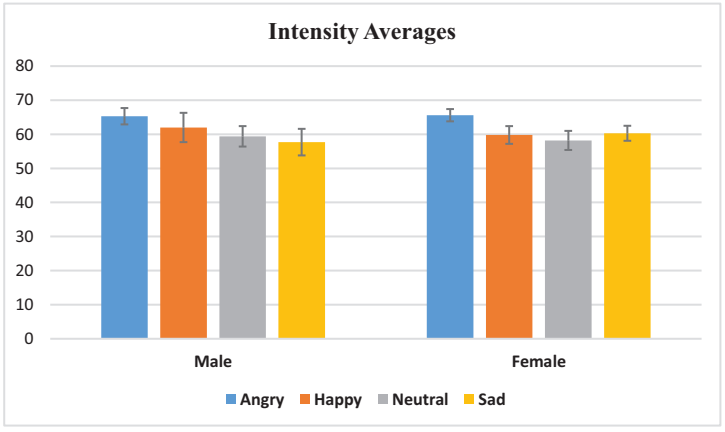

Fig. 7. Intensity parameter averages for the different emotions.

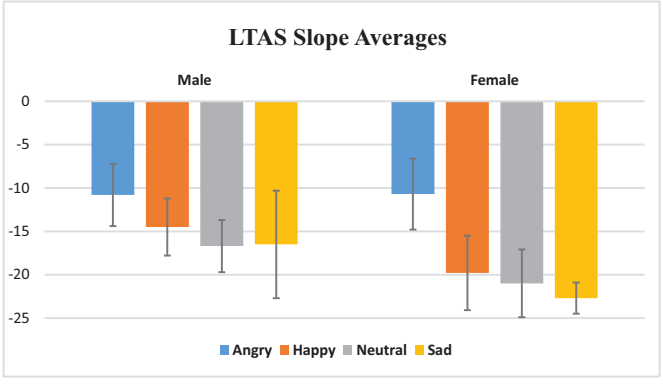

Fig. 8. LTAS slope parameter averages for the different emotions.

Table 13 summarizes the recognition rates for the different features using linear SVM classifier. LTAS features was found to give the best accuracies for males (90%), females (85%) and both (88.3%), followed by wavelet, prosodic and MFCC features. Table 14 shows recognition rates for various combinations of the different features. For males, females and both best accuracies were achieved for combination of features including both the LTAS and wavelet features among others. Another interesting observation is that although different feature combinations achieved the highest recognition rates for males, females and both genders, utilizing the complete feature vector resulted in considerably close performance to the highest rates.

Emotion perception from speech can be a challenging task even for humans. Meftah et al. (2014) showed that human raters correctly iden-

Table 13
Anger vs. neutral recognition rates for each feature category using linear SVM.

Features	Males	Females	Both
Prosodic	88.9	80.8	82.0
Formants	68.9	60.8	55.7
MFCC	85.0	79.2	82.0
LTAS	90.0	85.0	88.3
Wavelet	89.4	80.0	86.3

Table 14
Anger vs. neutral recognition rates for different feature combinations using linear SVM.

Features	Males	Females	Both
Prosodic - Formants	92.2	80.0	84.3
Prosodic - MFCC	91.1	80.0	86.7
Prosodic - LTAS	92.0	85.8	89.7
Prosodic - Wavelet	92.8	81.7	88.7
MFCC - Formants	89.4	82.5	81.3
MFCC - Wavelet	87.8	78.3	85.7
MFCC - LTAS	93.3	82.5	88.0
LTAS – Formants	93.9	82.5	89.3
LTAS – Wavelet	95.0	82.5	90.3
Wavelet - Formants	87.8	81.7	86.0
Prosodic - MFCC - LTAS	93.9	85.0	89.0
Prosodic - LTAS - Wavelet	94.4	86.7	91.0
MFCC - LTAS - Wavelet	93.9	80.0	88.7
MFCC – Formants - LTAS -Wavelet	96.1	80.8	88.7
Prosodic - MFCC – Formants -Wavelet	92.2	79.2	87.0
All	95.0	84.2	90.7

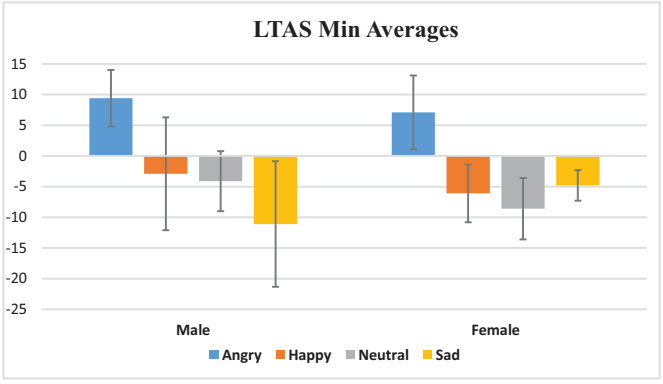

Fig. 9. LTAS min parameter averages for the different emotions.

tified only 88.5% of elicited emotions, happiness being the hardest to detect. Furthermore, Koolagudi et al. (2009) reported that only 61% and 66% of emotions acted by professional male and female actors, respectively were correctly recognized by human observers, which was even worse than the performance of their introduced SER system (males: 69%, females:75%). SER is thus a challenging field that has been gaining increased interest due its application in several emerging domains.

In this work, the Egyptian Arabic speech emotion (EYASE) database was introduced. Speaker independent and speaker dependent experiments were performed for three different cases: (1) emotion classifications, (2) arousal and valence classifications and (3) multi-emotion classifications. Furthermore, LTAS and wavelet features were implemented for SER, both which were shown to have the potential to improve SER performance when they are combined with the classical prosodic and spectral emotion features that are widely used in SER literature.

7. Conclusions

Arabic speech emotion recognition is a relatively new research field owing to the limited available speech emotion databases. In this work, a semi-natural Egyptian Arabic speech emotion (EYASE) database was introduced that includes 579 utterances from 3 male and 3 female professional actors for the angry, happy, neutral and sad emotions. Prosodic (pitch-intensity), spectral (formants, MFCC, LTAS) and wavelet features were computed for emotion detection. Both LTAS and wavelet features were seldom considered for SER in addition to not being previously implemented for Arabic SER. For the EYASE database, emotion, arousal & valence, as well as multi-emotion classifications were performed for both speaker independent and speaker dependent cases.

Generally, emotion expression depends on several factors including culture and gender. For the EYASE database, anger emotion was found to be the most readily detected whereas happiness was the most challenging. Arousal (angry/sad) recognition rates were shown to be superior to valence (angry/happy) recognition rates. Furthermore, higher accuracies were attained for male subjects than for female subjects in all performed experiments, which can be attributed to cultural aspects. All these observations were consistent for both speaker independent and dependent experiments. However, speaker dependent SER in most cases gave better performance than speaker independent SER.

Feature ranking showed that in addition to the classical prosodic features, LTAS and wavelet features were also highly relevant for Arabic SER in all the performed experiments. Specifically, LTAS features were among the highest relevant features for emotion vs. neutral, arousal & valence and multi-emotion classifications, whereas the wavelet features were relevant for the emotion vs. neutral and multi-emotion classifications. Furthermore for anger detection, adding the LTAS and wavelet parameters to the classical prosodic and spectral features was found to enhance the overall performance. LTAS and wavelet features are thus recommended for implementation within SER systems.

Declaration of Competing Interest

I wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

CRediT authorship contribution statement

Lamiaa Abdel-Hamid: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing - original draft, Visualization.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.specom.2020.04.005.

References

Akçay, M.B., Oğuz, K., 2020. Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. 116.
Alarcao, S.M., Fonseca, M.J., 2017. Emotions Recognition Using EEG Signals: a Survey. IEEE Trans. Affect. Comput. 3045, 1–20. doi:10.1109/TAFFC.2017.2714671.
Arabic Population [WWW Document], n.d. URL http://worldpopulationreview.com/countries/arab-countries/ (accessed 4.29.19).
Bahmanbiglu, S.A., Mojiri, F., Abnavi, F., 2017. The Impact of Language on Voice: an LTAS Study. J. Voice 31 (249). doi:10.1016/j.jvoice.2016.07.020, e9-249.e12.
Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proc. Inst. Phon. Sci. 97–110.
Boersma, P., Kovacic, G., 2006. Spectral characteristics of three styles of Croatian folk singing. J. Acoust. Soc. Am. 119, 1805–1816.
Boersma, P., Weenink, D., 2018. Praat: doing phonetics by computer [Computer program].
Brody, L., 2009. Gender, Emotion, and the Family. Harvard University Press.
Cao, H., Verma, R., Nenkova, A., 2015. Speaker-sensitive emotion recognition via ranking : studies on acted and spontaneous speech. Comput. Speech Lang. 29, 186–202.
Chaplin, T.M., 2015. Gender and emotion expression: a developmental contextual perspective. Emot. Rev. 7, 14–21. doi:10.1177/1754073914544408.
Chatterjee, J., Mukesh, V., Hsu, H.H., Vyas, G., Liu, Z., 2018. Speech emotion recognition using cross-correlation and acoustic features. In: Proceedings of the IEEE 16th International Conference on Dependable, Autonomic and Secure Computing IEEE 16th International Conference on Pervasive Intelligence and Computing IEEE International Conference on Big Data Intelligence, 3. Comput. IEEE, pp. 250–255. doi:10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00050.
Coifman, R.R., Wickerhauser, M.V., 1992. Entropy-based algorithms for best basis selection. IEEE Trans. Inf. Theory 38, 713–718. doi:10.1109/18.119732.
Dahake, P.P., Shaw, K., Malathi, P., 2017. Speaker dependent speech emotion recognition using MFCC and Support Vector Machine. In: Proceedings of the International Conference on Automatic Control and Dynamic Optimization Techniques. ICACDOT 2016. IEEE, pp. 1080–1084. doi:10.1109/ICACDOT.2016.7877753.
Egyptian Arabic [WWW Document], n.d. URL https://www.statista.com/statistics/377302/total-population-of-egypt/ (accessed 4.29.19).
Ekman, P., 1999. Basic emotions. Handb. Cogn. Emot. 445–460.
Ekman, P., Cordaro, D., 2011. What is meant by calling emotions basic. Emot. Rev. 3, 364–370. doi:10.1177/1754073911410740.
El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit 44, 572–587 10.1016/j.patcog.2010.09.020.
Esmaileyan, Z., Marvi, H., 2014. A database for automatic persian Speech Emotion Recognition: collection, processing and evaluation. Int. J. Eng. Trans. A Basics 27, 79–90. doi:10.5829/idosi.ije.2014.27.01a.11.
Eyben, F., Salomão, G.L., Sundberg, J., Scherer, K.R., Schuller, B.W., 2015. Emotion in the singing voice—A deeperlook at acoustic features in the light of automatic classification. Eurasip J. Audio Speech Music Process. 19. doi:10.1186/s13636-015-0057-6, 2015.
Fletcher, A.R., Wisler, A.A., McAuliffe, M.J., Lansford, K.L., Liss, J.M., 2017. Predicting Intelligibility Gains in Dysarthria Through Automated Speech Feature Analysis. J. Speech. Lang. Hear. Res. 60, 3058–3068. doi:10.1044/2017_JSLHR-S-16-0453.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software. ACM SIGKDD Explor. Newsl 11, 10. doi:10.1145/1656274.1656278.
Hall, M.A., 1999. Correlation-based Feature Selection for Machine Learning.
Han, Z., Wang, J., 2013. Speech emotion recognition based on wavelet transform and improved HMM. In: Proceedings of the 2013 25th Chinese Control and Decision Conference. CCDC 2013, pp. 3156–3159. doi:10.1109/CCDC.2013.6561489.

Haridas, A.V., Marimuthu, R., Sivakumar, V.G., 2018. A critical review and analysis on techniques of speech recognition: the road ahead. Int. J. Knowledge-Based Intell. Eng. Syst. 22, 39–57. doi:10.3233/KES-180374.

H. Almasaa [WWW Document], n.d. URL https://www.imdb.com/title/tt7046200/ (accessed 4.29.19).

Hifny, Y., Ali, A., 2019. Efficient Arabic emotion recognition using deep neural networks. In: Proceedings of the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6710–6714. doi:10.1109/ICASSP.2019.8683632.

Joshi, D.D., Zalte, M.B., 2013. Recognition of Emotion from Marathi Speech Using MFCC and DWT Algorithms 59–63.

Kacha, A., Grenez, F., Orozco-Arroyave, J.R., Schoentgen, J., 2020. Principal component analysis of the spectrogram of the speech signal: interpretation and application to dysarthric speech. Comput. Speech Lang. 59, 114–122 10.1016/j.csl.2019.07.001.

Kamińska, D., Pelikant, A., 2012. Recognition of human emotion from a speech signal based on plutchik's model. Int. J. Electron. Telecommun. 58, 165–170. doi:10.2478/v10177-012-0024-4.

Khalil, A., Al-Khatib, W., El-Alfy, E.S., Cheded, L., 2018. Anger detection in Arabic speech dialogs. In: Proceedings of the International Conference on Computing Sciences and Engineering, ICCSE 2018 - Proceedings. IEEE, pp. 1–6. doi:10.1109/ICCSE1.2018.8374203.

Kinnunen, T., Hautamäki, V., Fränti, P., 2006. On the use of long-term average spectrum in automatic speaker recognition. In: Proceedings of the 5th International *Symposium on Chinese Spoken Language Processing (ISCSLP'06)*, Singapore, pp. 559–567.

Klaylat, S., Osman, Z., Hamandi, L., Zantout, R., 2018. Emotion recognition in Arabic speech. Analog Integr. Circuits Signal Process. 96, 337–351. doi:10.1007/s10470-018-1142-4.

Koolagudi, S.G., Maity, S., Kumar, V.A., Chakrabarti, S., Rao, K.S., Koolagudi Shashidhar, G., Maity, S., K.V.A, C.S., R.K.S., 2009. IITKGP-SESC: speech database for emotion analysis. Commun. Comput. Inf. Sci. 40, 485–492. doi:10.1007/978-3-642-03547-0_46.

Koolagudi, S.G., Murthy, Y.V.S., Bhaskar, S.P., 2018. Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. Int. J. Speech Technol. 21, 167–183. doi:10.1007/s10772-018-9495-8.

Kostoulas, T., Fakotakis, N., 2006. A speaker dependent emotion recognition framework. In: Proceedings of the 5th International Symposium on Communication Systems, Networks and Digital Signal Processing, pp. 305–309.

Krishna Kishore, K.V., Krishna Satish, P., 2013. Emotion recognition in speech using MFCC and wavelet features. In: Proceedings of the 3rd International Advance Computing Conference IACC 2013 842–847 doi:10.1109/IAdCC.2013.6514336.

Lalitha, S., Geyasruti, D., Narayanan, R., Shravani, M., 2015a. Emotion Detection Using MFCC and Cepstrum Features. Proc. Comput. Sci. 70, 29–35. doi:10.1016/j.procs.2015.10.020.

Lalitha, S., Madhavan, A., Bhushan, B., Saketh, S., 2015b. Speech emotion recognition. In: Proceedings of the International Conference on Advances in Electronics, Computers and Communications, ICAECC 2014. IEEE, pp. 1–4. doi:10.1109/ICAECC.2014.7002390.

Li, X., Akagi, M., 2019. Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model. Speech Commun 110, 1–12.

Likitha, M.S., Gupta, S.R.R., Hasitha, K., Raju, A.U., 2018. Speech based human emotion recognition using MFCC. In: Proceedings of the International Conference on Wireless Communications, Signal Processing and Networking 2017, pp. 2257–2260. doi:10.1109/WiSPNET.2017.8300161 2018-Janua.

Mallat, S.G., 1989. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. 674–693.

Meddeb, M., Karray, H., Alimi, A.M., 2017. Content-based Arabic speech similarity search and emotion detection. In: Hassanien, A.E., Shaalan, K., Gaber, T., Azar, A.T., Tolba, M.F. (Eds.), Proceedings of the International Conference On Advanced Intelligent Systems and Informatics, 2016. Springer International Publishing, Cham, pp. 530–539.

Meftah, A., Alotaibi, Y., Selouani, S.-A., 2014. Designing, building, and analyzing an Arabic speech emotional corpus. Work. Free. Arab. Corpora Corpora Process. Tools Work. Program. 22.

Meftah, A., Selouani, S.A., Alotaibi, Y.A., 2015. Preliminary Arabic speech emotion classification. In: Proceedings of the IEEE International Symposium on Signal Processing and Information Technology ISSPIT 2014, pp. 179–182. doi:10.1109/ISSPIT.2014.7300584.

Mencattini, A., Martinelli, E., Costantini, G., Todisco, M., Basile, B., Bozzali, M., Di Natale, C., 2014. Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. Knowl. Based Syst. 63, 68–81.

Mohammadi, Z., Frounchi, J., Amiri, M., 2017. Wavelet-based emotion recognition system using EEG signal. Neural Comput. Appl. 28, 1985–1990. doi:10.1007/s00521-015-2149-8.

Muckenhirn, H., Korshunov, P., Magimai-Doss, M., Marcel, S., Muckenhirn, H., Korshunov, P., Magimai-Doss, M., Marcel, S., 2017. Long-term spectral statistics for voice presentation attack detection. IEEE/ACM Trans. Audio, Speech Lang. Process. 25, 2098–2111.

Mustafa, M.B., Yusoof, M.A.M., Don, Z.M., Malekzadeh, M., 2018. Speech emotion recognition research: an analysis of research focus. Int. J. Speech Technol. 21, 137–156. doi:10.1007/s10772-018-9493-x.

Özseven, T., 2019. A novel feature selection method for speech emotion recognition. Appl. Acoust. 146, 320–326. doi:10.1016/j.apacoust.2018.11.028.

Palo, H.K., Mohanty, M.N., 2018. Wavelet based feature combination for recognition of emotions. Ain Shams Eng. J. 9, 1799–1806. doi:10.1016/j.asej.2016.11.001.

Pan, Y., Shen, P., Shen, L., 2012. Speech emotion recognition using support vector machine. Int. J. Smart Home 6, 101–108. doi:10.1109/KST.2013.6512793.

Plutchik, R., 1991. The Emotions. University Press of America.

Posner, J., Russell, J.A., Peterson, B.S., 2005. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev. Psychopathol. 17, 715–734. doi:10.1017/S0954579405050340.

Pudalov, A.R., 2016. IiCA and MAPP - positive anger expression in middle eastern refugee children IiCA and MAPP - positive anger expression in middle eastern refugee. Master Appl. Posit. Psychol. Serv. Learn. Proj..

Ram, C.S., Ponnusamy, R., 2014. An effective automatic speech emotion recognition for Tamil language using support vector machine. In: Proceedings of the 2014 International Conference On Issues and Challenges in Intelligent Computing Techniques. ICICT 2014. IEEE, pp. 19–23. doi:10.1109/ICICICT.2014.6781245.

Rybka, J., Janicki, A., 2013. Comparison of speaker dependent and speaker independent emotion recognition. Int. J. Appl. Math. Comput. Sci. 23, 797–808. doi:10.2478/amcs-2013-0060.

Sagha, H., Deng, J., Schuller, B., 2018. The effect of personality trait, age, and gender on the performance of automatic speech valence recognition. In: Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction, ACII 2017. IEEE, pp. 86–91. doi:10.1109/ACII.2017.8273583.

Saste, S.T., Jagdale, S.M., 2017. Emotion recognition from speech using MFCC and DWT for security system. In: Proceedings of the International Conference of Electronics, Communication and Aerospace Technology (ICECA), pp. 701–704. doi:10.1109/ICECA.2017.8203631.

Sinith, M.S., Aswathi, E., Deepa, T.M., Shameema, C.P., Rajan, S., 2016. Emotion recognition from audio signals using Support Vector Machine. In: Proceedings of the IEEE Recent Advances in Intelligent Computational Systems, RAICS 2015. IEEE, pp. 139–144. doi:10.1109/RAICS.2015.7488403.

Swain, M., Routray, A., Kabisatpathy, P., 2018. Databases, features and classifiers for speech emotion recognition: a review. Int. J. Speech Technol. 21, 93–120. doi:10.1007/s10772-018-9491-z.

Swain, M., Routray, A., Kabisatpathy, P., Kundu, J.N., 2017. Study of prosodic feature extraction for multidialectal Odia speech emotion recognition. In: Proceedings of the IEEE Region 10 Annual International Conference, Proceedings/TENCON. IEEE, pp. 1644–1649. doi:10.1109/TENCON.2016.7848296.

Tawari, A., Trivedi, M.M., 2010. Speech emotion analysis: exploring the role of context. IEEE Trans. Multimed. 12, 502–509. doi:10.1109/TMM.2010.2058095.

Thagard, P., 2019. Mind Society: From Brains to Social Sciences and Professions. Oxford University Press (March 1, 2019).

Tirumala, S.S., Shahamiri, S.R., Garhwal, A.S., Wang, R., 2017. Speaker identification features extraction methods: a systematic review. Expert Syst. Appl. doi:10.1016/j.eswa.2017.08.015.

Universit, E., Strack, F., 1991. Subjective well-being: An Interdisciplinary Perspective.

Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: resources, features, and methods. Speech Commun 48, 1162–1181. doi:10.1016/j.specom.2006.04.003.

Vogt, T., André, E., 2006. Improving automatic emotion recognition from speech via gender differentiation. Proc. Lang. Resour. Eval. Conf. 1123–1126.

Vogt, T., André, E., Bee, N., 2008. EmoVoice—A framework for online recognition of emotions from voice. In: International Tutorial and Research Workshop On Perception and Interactive Technologies For Speech-Based Systems. Springer, pp. 188–199.

Yüksel, M., Gündüz, B., 2018. Long term average speech spectra of Turkish. Logop. Phoniatr. Vocology 43, 101–105. doi:10.1080/14015439.2017.1377286.