

COVER PAGE

Document downloaded by @DAEL

Tue Nov 10 15:03:38 2026
For personal use

When automatic English translation is provided, only the original document is authentic.
The EAA cannot be held responsible of any translation error

Bibliographical reference

Contribution to the Creation of an Arabic Expressive Speech Corpus, L. Demri, L. Falek and H. Teffahi, Acta Acustica **vol. 101** (Number 5), 2015, pp. 1052-1063

DOI
https://doi.org/10.3813/AAA.918899

Acta Acustica united with Acustica
Vol. 101 (2015) 1052 – 1063

Technical and Applied Papers
DOI 10.3813/AAA.918899

Contribution to the Creation of an Arabic Expressive Speech Corpus

L. Demri, L. Falek, H. Teffahi
Department of Electronics and Computer Science, University of Science and Technology
Houari Boumediene, BP 32, El Alia, 16111, Bab-Ezzouar, Algiers.
ldemri1987@hotmail.fr, falek.leila@gmail.com, hteffahi@gmail.com

Summary
In this article we present the methodology employed for the design and evaluation of a Basic Arabic Expressive Speech corpus (BAES-DB). The corpus, which has a total length of approximately 150 minutes, is constituted of 13 speakers uttering a set of 10 sentences while simulating 3 emotions (joy, anger and sadness) in addition to a neutral utterance. The 10 sentences have been selected to meet phonetic equilibrium and absence of emotional content criteria, from a corpus of sentences proposed in Boudraa [1]. The corpus was evaluated through various tests of guided categorization, performed through a website. The overall good recognition rate is currently at 83.03%, with sadness being the most well recognized type of expressive speech (90.93%) and joy being the least well recognized (73.87%).

PACS no. 43.70.+i, 43.72.+q

1. Introduction

The field of affective computing has been gaining popularity for the last two decades [2]. The growing interest for human-machine interactions attracts numerous researchers from various regions and backgrounds. Expressive speech synthesis (ESS) is one of the many branches of affective computing. Its aim is to synthesize speech that contains, in addition to the semantic information of the uttered text, some emotional information that must provide a new layer of information to the listener.

Expressivity can be defined, as described by Beller in [3], as "a layer of information in communication, which consists of external manifestations, both controlled and uncontrolled, of internal states of a person that are inaccessible to other persons' perceptions." This is the definition that will be used throughout this paper.

One crucial point in the design of an ESS system is the availability of an expressive speech database in order to extract as much information as possible on the various human emotions. Such information allows us to characterize types of expressive speech in order to be able to imitate them during synthesis. Such a corpus can also be exploited in the field of emotion recognition. In the field of speaker recognition, it allows us to broaden the intra-speaker variability. In speech recognition. it also allows us to study the influence of a given emotion on the various characteristics of speech in order to deduce the uttered text regardless of the speaker's emotional state.

Received 24 June 2014,
accepted 06 August 2015.

Many expressive speech corpora already exist. Douglas-Cowie [4] and Ververidi [5] present overviews of the various existing corpora up until 2006. More recently, the HUMAINE corpus [6] contains recordings of natural and simulated emotions in English, French, and Hebrew. Grimm's VAM corpus [7] consists of 12 hours of audio-visual data in German, extracted from a talk show in which the speakers talked to each other without knowledge that their utterances would later be analyzed. The RECOLA corpus [8] contains audio, video, ECG and EDA data of 23 pairs of interlocutors (46 speakers) interacting in order to solve a hypothetical life or death situation (a "survival test", described in the paper). The discussion between the two interlocutors is supposed to contain emotional data because of the urgency (although hypothetical) of the situation.

Up until recently, research on Arabic emotion and expressivity has been very limited. The earliest works on expressive speech synthesis in Arabic seem to be those of Al-Dakkak, in 2005 [9, 10]. In them, Al-Dakkak records sentences in various emotional states, and analyzes them via a statistical study of the pitch, intensity, and duration (speech rate), in order to extract rules to be applied to the synthesized speech, e.g. for anger: "*F0 mean increases by 40–75%, F0 slope increases, intensity mean decreases*", and so on. This is a classic approach used in expressive speech synthesis, but it does not take the voice quality into consideration, and therefore could produce a more natural speech (if voice quality was modifiable). It is not clear how many speakers were recorded, nor whether the speech corpus was made available online.

Since 2013, however, there seems to be a growing interest in Arabic emotion and expressivity, especially in the

© S. Hirzel Verlag · EAA

Document downloaded by @DAEL, 2026-11-10 15:03:40 - personal use only

field of emotion classification and recognition in Arabic text. [11], [12], [13] and [14] deal with emotion detection from Twitter feeds and children story books.

In the field of spoken language, a few contributions can be mentioned in emotion recognition [15, 16] and Arabic expressive speech synthesis [17, 18].

Despite the abundance of expressive speech corpora in various languages, the phonological and phonemic differences make it impossible to exploit certain pieces of emotions to imitate them in a different language. In other words, even though some characteristics of certain expressive speech types are common for every language (e.g. the increase of intensity of the voice for anger), the differences in accents and phonemes between languages constitute an obstacle to the analysis of characteristics specific to the studied language (whenever they exist). Furthermore, cultural differences might arise, as pointed out by the works of Rilliard, Shochi *et al.* [19, 20, 21, 22]

Many speech corpora exist for the Arabic language [23, 24, 25], but most of these corpora have been designed for specific applications, and since they do not contain expressive speech, they cannot be exploited for the analysis of expressivity in the Arabic language. The low amount of Arabic expressive speech compared to other languages motivated us to create an Arabic expressive speech corpus, very similar to that of Burkhardt for the German language [26], which will hopefully provide new data in the field of affective computing in Arabic. The aim is to enrich the global online resources of Arabic expressive speech with a freely downloadable, annotated Arabic expressive speech corpus. Expressive speech databases are usually used for unit selection synthesis, and therefore have to be quite large; however, the primary aim of this database is to provide information on how expressivity modifies the voice quality and how it manifests in the prosodic and micro-prosodic level in the Arabic language. We present in this paper the methodology employed to design and evaluate such a corpus.

2. Recording protocol

A total of 13 speakers have so far been recorded in two different locations in the Department of Electronics and Computer Science at the USHTB (Algiers). This section provides details on the recording conditions, the chosen sentences, and the speakers.

2.1. Equipment, software, and recording conditions

The recordings were performed with a dynamic hypercardioïd Shure® Beta 58A microphone. This microphone reduces the sounds coming from afar or from behind the microphone, allowing a good signal-to-noise ratio.

During the recordings, the microphone is placed at approximately 20 cm in front of the mouth of the speaker and plugged to an M-Audio® MobilePre® audio interface. The files were recorded on 16 bits at 44100 Hz. The acquisitions were made using GoldWave®.



Figure 1. Recording conditions for the first 4 speakers.

Table I. Values of the theoretical RT60 for each proposed frequency in [2]1.

Frequency (Hz)	RT_{60} (s)
125	1.48
250	1.74
500	0.61
1000	0.58
2000	0.72
4000	1.02

The first 4 speakers were recorded at the Spoken Communication and Signal Processing Laboratory (Lab 66), which is a 12 m × 9.2 m × 3.4 m room with 3 painted concrete walls and one large window as the 4th wall. The ceiling is painted concrete, and the floor is made of ceramic tiles. Figure 1 shows the equipment used and the recording configuration in Lab 66. The speakers were recorded late in the afternoon when the laboratory was empty and there were less people crossing the faculty hall. The theoretical values of the reverberation time (RT60) for these room characteristics were calculated by the online tool on the website of the SAE institute and are provided in Table I.

As Lab 66 is not an empty room and contains many tables and furniture, a "hand clap test" was performed to obtain an empirical value of RT_{60}. Two hand claps were recorded in Lab 66 with two different microphones (the Shure® Beta 58A mentioned above, and a simple laptop-integrated microphone), and the RT_{60} values were read after normalization at 0 dB using Audacity®. The obtained values were approximately 0.609 s, and 0.640 s. The experimental RT_{60} is therefore below 0.65 s for a hand clap. The signal-to-noise ratio was calculated "post-hoc", by calculating the mean energies for the vocal parts and for the silent parts, and dividing the former by the latter.

[2] http://www.sae.edu/reference_material/pages/Reverberation%20Time%20Calculator.htm

Document downloaded by @DAEL, 2026-11-10 15:03:40 - personal use only



Figure 2. Anechoic chamber described in [27].

The mean signal-to-noise ratio, for the first 4 speakers, is 56.41 dB.

The 9 other speakers were recorded in an anechoic chamber built in the Acoustics Laboratory of the Department of Electronics and Computer Science (Lab 73). The room is 3.5 m × 2.5 m × 2.3 m wide, with an RT_{60} below 0.6s for all frequencies, and about 0.1s at 500 Hz. More details about the construction and the characteristics of the room can be found in [27]. The mean signal-to-noise ratio for the 9 speakers recorded in this room is 59.19 dB.

Though the first set of recordings was performed as a preliminary experiment, we feel they would be of value for machine classification tasks that need to be robust to the variations of acoustic recording conditions. The database can therefore be viewed as two sets of data, with the first set consisting of the first 4 speakers and the second consisting of speakers 5 to 13.

2.2. Selected sentences

In this work, our primary focus was mainly to gather data for the study of the variations of the prosody and microprosody and of the voice quality when expressivity is involved in speech. We also wanted the phonemic frequencies of the sentences to reflect those of the Arabic language. Therefore, the sentences of the corpus were selected to satisfy two criteria: on one hand, we wanted phonetically balanced sentences, in order to have the greatest number of phonemes while maintaining the frequency of occurrence of each phoneme; on the other hand, we did not want the sentences to contain emotional clues, so that the meaning of a sentence does not influence the perception of the emotion during the evaluation of the corpus. These precautions lead us to assume that the perceived emotions are caused by the acoustic characteristics of the recorded voices rather than the semantic meaning of the sentences.

In [1], Boudraa proposes 20 lists of 10 phonetically balanced sentences in standard Arabic. For our corpus, one of the lists has been selected in order to meet the absence of emotional clues criteria. Since it is not possible to totally remove semantic emotional clues, we decided that only those sets that contained sentences that clearly stated the emotional state of the speaker should be rejected, e.g. sentences such as "I am so happy today". The selected sentences, their phonetic transcription and their meaning are given in Table II.

Table II. Selected list of 10 phonetically balanced sentences in Arabic along with their meanings and phonetic transcriptions.

Number Sentence	Meaning Phonetic transcription
Sentence 1: أَحفَظُ مِنَ الأَرضِ	I preserve the earth ʔaħfazʀu mina lʔardʀi
Sentence 2: أَينَ المُسَافِرُونَ ؟	Where are the voyagers? ʔayna lmusa:firu:na
Sentence 3: لَا لَم يَستَمتِع بِثَمَرِهَا	No, he did not enjoy her fruits la: lam yastamtiʔʀ biθamariha:
Sentence 4: سَيُؤَذِيهِم زَمَانُنَا	Our epoch will hurt them sayuʔði:him zama:nuna:
Sentence 5: كُنتُ قُدوَةً لَهُم	I was a model to them kuntu qudwatan lahum
Sentence 6: اَزَرَ صَائِمًا	He rewarded a fasting man ʔa:zara sʀa:ʔiman
Sentence 7: كَالَ وَ غَبَطَ الكَبشَ	He weighed and tied the sheep ka:la wa ɣabatʀa lkabʃa
Sentence 8: هَل لَذَعَتهُ بِقَولٍ ؟	Did she hurt him with words? hal laðaʔʀathu biqawlin
Sentence 9: عَرَفَ وَالِيًا وَ قَائِدًا	He knew a Wali and a Caïd ʔʀarafa wa:liyan wa qa:ʔidan
Sentence 10: خَلَا بَالُنَا مِنكُمَا	Our mind was distracted from you xala: ba:luna: minkuma:

Phonetic context was not taken into consideration for this first set of sentences. However, it can be mentioned that these sentences cover 123 diphones of the 1521 possible pairs of Arabic phones [28]. This number is calculated by assuming that the typical number of phones in a language is the square of the number of phones in that language, which is 39 for Arabic [29]; however, a lot of possible phone-phone combinations do not naturally occur, meaning there are certainly less than 1521 diphones in Arabic. In order to form an idea of the average number of diphones that are used in Arabic texts, as well as their frequency of occurrence, we counted the number of their occurrences from two short texts; the first is a news article which can be found online[3], and the second is the introductory text for a religious booklet. A total of 274 di-

[3] http://www.alhayat.com/Articles/9061553/

Document downloaded by @DAEL, 2026-11-10 15:03:40 - personal use only

phones were counted from both of the texts combined, 92 of which were found both in the sentences and the texts.

2.3. Information about the speakers

Speaker information is shown in Table III. All speakers were born in Algeria. They all fluently speak dialectal Arabic (also called Algerian Arabic) which is a mixture of Standard Arabic, Berber (an ancient pre-Islamic language spoken in North Africa) French, and some Turkish words. All have learned Modern Standard Arabic (MSA) at school. However, the parents of the speakers all have different geographical origins. Despite the geographic differences among the speakers, the pronunciations are not too variable, because Standard Arabic has a tendency to standardize pronunciations of speakers of different regions. The "character trait" column of Table III is based on our personal general impression of each speaker during the recording session. It should be noted here that speaker 12 had 2 years of theater school, and participated in a theater competition in which their group was rewarded as the second best in Algeria. Apart from that, no speaker had any training in acting. We deemed the selected speakers adequate mostly because the selected expressive types are not very hard to simulate. Natural conversations very frequently contain simulations of emotions for the purpose of comedic effects, cynicism, emphasis, lying, etc. Therefore, we assume that most people are capable of simulating certain emotions and using various types of expressive speech, by the simple fact that they live in a human community where this communication skill is necessary.

2.4. Recording instructions

The selected expressions are neutral, joy, sadness and anger, as these basic expressions are distinguishable in every culture and even in animals [30, 31]. Each speaker recorded the 10 sentences for a given expressive situation before moving on to the next one. The first four speakers were recorded while sitting, whereas the 9 others were standing.

For the neutral state, the speakers were asked to utter the sentences as a news presenter. It is therefore a reading style, as the speakers do not utter the sentences as they naturally would in a real conversation, but in a more "formal" and "official" way. This choice was made in an attempt to standardize the neutral utterances, so that they are unaffected by the speaker's personal way of speaking. For the recordings of the other expressions, the speakers were asked to simulate the appropriate expressive situations, joy, sadness and anger. Although the expressive speech is not spontaneous, the linguistic content of the sentences remains the same, which is beneficial for acoustic analysis.

2.5. Difficulties in simulating expressive speech and strategies employed by the speakers

On one hand, we wanted the speaker to pronounce the 10 utterances in an identical manner in order to compute accurate statistics and limit the effects of noise. On the other hand, however, we wanted the expression to be as realistic as possible. For the recording of "joy", we used various techniques to amuse the speakers before or during the recordings. This caused bursts of laughter, as well as inconsistency in producing the expressive speech. Another problem was that it was not easy to produce joy while repeating a single sentence 10 times. The first 4 or 5 utterances might be sincerely "joyous", but the next ones have a tendency to be more "mechanical". In fact, when a speaker is trying to simulate the expression of joy over several repetitions of the same sentence, this expression can be interpreted as sadness [32].

Various strategies were employed for the simulation of "joy". Speaker 1 performed the recordings alone, and therefore had to resort to reading jokes before recording the sentences. His expression of joy tended to be heard as "neutral".

Speakers 2 and 3 exaggerated their performances, opting for very simulated and unnatural performances, but this resulted in better recognition rates for "joy".

Speakers 4, 5, 7, 8, 11, and 13 produced genuine or almost genuine joy because of their "light-hearted" and "joyful" nature, or because they were amused by the recording process (sometimes both).

Speaker 6 was particularly (even surprisingly) talented in the ability to genuinely laugh on purpose. Speaker 12 was amused by the experiment, in addition to her aforementioned theatrical training.

Speaker 9 was very shy, therefore some volunteers were asked to quietly attempt to make her laugh while she uttered the sentences.

The negative expressions seemed easier to simulate than the positive ones. For "sadness", the speakers uttered the sentences with a plaintive tone of voice. For anger, the speakers were asked to use hot anger, but some (Speakers 8 and 9) were rather uncomfortable producing this expression because of their rather calm and reserved nature.

2.6. Formatting of the speech corpus

The recording phases resulted in 520 uncompressed sound files (13 speakers, 4 emotions, 10 sentences). Each file contains the 10 utterances of a sentence in one of the four expressive states for each speaker.

An algorithm based on the computation of an energy threshold and on the duration of the silent parts was designed in order to parse each file into separate saved sound files.

2.7. Corpus annotations

In order to facilitate the automatic analysis of the corpus, the utterances will be annotated manually using Praat. For each speech file, 2 interval tiers will be created, namely "Words" and "Phones". The "Words" tier contains the word boundaries along with an orthographic transcription of the words in the sentence (in Arabic). The "Phones" tier contains phone boundaries labeled with a symbol (one or two characters) representing the phone; these symbols are loosely based on the standard phonetic symbols but are

Document downloaded by @DAEL, 2026-11-10 15:03:40 - personal use only

Table III. Information on the selected speakers.

Speaker		Parents' origins	Character trait	Gender	Age
Speaker 1	L. D	Tizi-Ouzou, Bejaia	Indifferent, Calm	Male	26
Speaker 2	A. B	Tlemcen, Mahares	Light-hearted	Male	25
Speaker 3	Sal. B	Algiers, Tlemcen	Calm, Serious	Male	25
Speaker 4	H. S	Algiers	Extraverted, Joyful	Female	25
Speaker 5	N. O	Tizi-Ouzou	Joyful, Calm	Male	25
Speaker 6	K. K	Algiers	Active, Impatient, Joyful	Male	24
Speaker 7	A. M	Tizi-Ouzou	Extraverted, Joyful	Female	30
Speaker 8	H. T	Tizi-Ouzou	Light-hearted	Female	27
Speaker 9	S. S	Algiers	Very shy	Female	25
Speaker 10	H. Z	Souk-Ahras	Shy, Joyful	Female	27
Speaker 11	A. H	Algiers	Light hearted, Joyful	Male	24
Speaker 12	Sar. B	Algiers	Light hearted, Joyful	Female	24
Speaker 13	R. R	Jijel	Light hearted, Joyous	Female	21

also inspired by Arabic "SMS language" (e.g. 7 for letter "ح" and 3 for letter "ع"). So far only speaker 1 and some of speaker 2's utterances have been annotated; however the final goal is to provide annotations for all 13 speakers.

3. Validation of the recorded speech corpus

In order to validate the expressive speech corpus and thus to verify that the recordings do indeed contain the desired expressive information, a website[4] was created to facilitate the evaluation of the database. Three tests were performed on the recorded corpus, as described in the following subsections.

3.1. "Per-speaker" evaluation

3.1.1. Description

In this test, the evaluator selects a speaker and rates only that speaker's performances. The evaluator must make a judgment regarding the expression contained in the utterance, and its naturalness. Figure 3 is an explanatory diagram about the design of the test.

Before beginning the test, the evaluator is told that it is about recognizing expressive speech[5] (direct test, [3]), what the expressions are, and that the meaning of the sentences does not matter. He/she is also asked to reply instinctively, without too much thinking. The evaluator then chooses one of the speakers and the test begins.

The evaluator hears one of the 10 utterances (randomly chosen) of one of the sentences (randomly chosen) in one of the expressive states (randomly chosen) of the speaker he/she selected. The evaluator has the possibility to listen to the utterance as many times as desired, although a spontaneous response is expected. He/she is asked to pick, out of 4 possible answers (neutral, anger, joy and sadness),

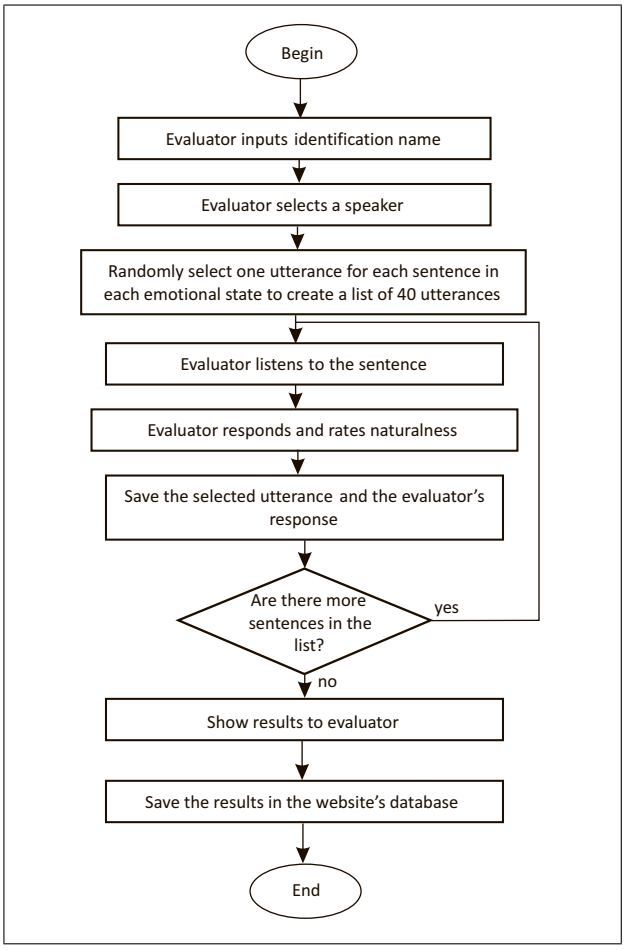


Figure 3. Evaluation algorithm.

the expressive state in which the speaker seemed to be. It is therefore a direct test of guided categorization [3]. The evaluator must also rate, on a scale from 1 to 5, the naturalness of the utterance, i.e, whether the simulated expression seems "forced", with 1 being the least natural and 5 being the most natural utterance.

The random selection of utterances is made before the test begins, by creating a randomized list of 40 utterances

[4] http://www.voice-research.hostingsiteforfree.com www.voice-research.org/synth/

[5] In actuality, for simplicity sake, it was assumed that the evaluators did not know the difference between emotive and expressive speech, thus the word "emotion" was used rather than "expressivity".

Document downloaded by @DAEL, 2026-11-10 15:03:40 - personal use only

chosen so that the evaluator tests each sentence in each expressive state once. Only one of the 10 utterances is randomly chosen for each sentence. The drawback to this is that a large number of tests must be performed before each of the utterances has been tested.

In total, the corpus is made up of 5200 speech files (13 speakers, 4 expressions, 10 sentences, 10 utterances), each containing one utterance. Since each evaluator tests 40 utterances in each test, a minimum of 130 evaluations must be performed before each utterance is evaluated (actually much more, because the utterances are picked randomly). However, since the utterances are meant to be identical, we can have a general idea of the quality of the corpus before each utterance has been tested. The precision of the evaluation becomes more and more accurate as the number of evaluations increases.

When the utterances are not identical, we can distinguish the good from the bad ones by saving, for each utterance, the speaker ID, the expression, and the number of the utterance, along with the answer given by the evaluator.

3.1.2. Results and discussion

The results for this test are based on those of 75 passes by 23 different evaluators. 2178 of the 5200 utterances have been evaluated, that is 41.88% of all the available utterances. The global rate of correct answers is 83.03%. These results are slightly lower than those in [26] for the Neutral, Anger, and Joy (Happiness) expressions, but higher for the expression of Sadness. Compared to the overall good recognition rate, the results are within a 3% range. The mean naturalness score for all tests is 3.93 on a scale from 1 to 5, but it is 4.02 for correct answers only. Table IV is the confusion matrix for the obtained results.

Table IV shows the answers given by the evaluators, for each simulated expression. For example, when neutral utterances were presented, the evaluators answered "neutral" 605 times, "anger" 16 times, "joy" 11 times and "sadness" 120 times.

Much information can be extracted from this table. First, it is clear that negative expressions, i.e. sadness and anger, are better recognized than the others. This can be due to the fact that these expressions are easy to simulate (a plaintive tone of voice for sadness, faster and louder speech for anger), or that they are easy to recognize, or most probably both of these reasons. The "neutral" and "joy" expressions are not recognized as well, but their results are still higher than chance. It may be that positive emotions are harder to simulate for persons who are not professional actors. Considering the emotional dimensions [33], it seems that the activation dimension is easier to simulate than the valence dimension. It is also possible that humans are more sensitive to negative expressions.

Table IV also shows the most frequent confusions that occur between the expressions. We can see that when "joy" is presented, the most common mistake is "neutral". The contrary is not true: "neutral" is not mostly mistaken with joy. Interestingly enough, the "neutral" expression is mostly mistaken with "sadness". There is therefore a tendency to confuse each expression with a more negative one.

Table IV. Confusion matrix of the obtained results for 75 evaluations. Ordinate: expressions simulated by the speakers; Abscissa: expressions perceived by the evaluators.

	Neutral	Anger	Joy	Sadness
Neutral	605 (80.67%)	16 (2.13%)	11 (1.47%)	120 (16%)
Anger	72 (9.6%)	650 (86.53%)	23 (3.07%)	5 (0.67%)
Joy	126 (16.8%)	26 (3.33%)	554 (74%)	43 (5.73%)
Sadness	55 (7.33%)	6 (0.8%)	7 (0.93%)	682 (90.93%)

Table V. Correct recognition rates and naturalness ratings for each speaker.

Speaker	Correct recognition rate	Naturalness
Speaker 1	74.64%	N/A
Speaker 2	89.00%	N/A
Speaker 3	85.75%	N/A
Speaker 4	91.39%	4.43
Speaker 5	72.00%	3.73
Speaker 6	80.63%	3.91
Speaker 7	88.13%	3.65
Speaker 8	72.63%	3.93
Speaker 9	71.67%	3.96
Speaker 10	88.33%	3.86
Speaker 11	92.50%	4.52
Speaker 12	89.17%	4.01
Speaker 13	83.33%	3.92

Since the speakers used different strategies to simulate expressions, the results vary considerably. The recognition rates as well as the naturalness ratings for each speaker are given in Table V, and Table VI shows the confusion matrices for each of the speakers.

It is somewhat surprising that speaker 12 has not offered the overall best performances, despite having gone through theatrical training. Also, speaker 9's "joy" is poorly recognized, despite the fact that she was not simulating the emotion but actually laughing. On the other hand, the best results for "joy" are those of speaker 6, who, as previously mentioned, was very talented in the ability to laugh on purpose, without actually being joyful or amused. This raises the question of whether it is better to record simulated emotions rather than natural ones as the latter might not contain the acoustic features we wish to analyze. Also, humans rely significantly on facial cues to infer the emotional state of the speaker, therefore, for an audio-only database, having exaggerated performances might be a good choice.

Table VI allows us to decide which speaker should be chosen in order to study a given expression; for example,

Document downloaded by @DAEL, 2026-11-10 15:03:40 - personal use only

Table VI. Confusion matrices of the global results for each speaker. Ordinate: Expressions simulated by the speakers; abscissa: expressions perceived by the evaluators.

		Neutral	Anger	Joy	Sadness
Speaker 1 (14 tests)	Neutral	106 (75.71%)	4 (2.86%)	5 (3.57%)	25 (17.86%)
	Anger	14 (10.00%)	111(79.29%)	15 (10.71%)	0 (0.00%)
	Joy	61 (43.57%)	4 (2.86%)	65 (46.43%)	10 (7.14%)
	Sadness	1 (0.71%)	1 (0.71%)	2 (1.43%)	136(97.14%)
Speaker 2 (10 tests)	Neutral	87 (87.00%)	2 (2.00%)	4 (4.00%)	9 (9.00%)
	Anger	5 (5.00%)	92 (92.00%)	0 (0.00%)	2 (2.00%)
	Joy	12 (12.00%)	3 (3.00%)	82 (82.00%)	2 (2.00%)
	Sadness	6 (6.00%)	0 (0.00%)	0 (0.00%)	94 (94.00%)
Speaker 3 (10 tests)	Neutral	82 (82.00%)	1 (1.00%)	0 (0.00%)	17 (17.00%)
	Anger	4 (4.00%)	95 (95.00%)	0 (0.00%)	0 (0.00%)
	Joy	21 (21.00%)	4 (4.00%)	73 (73.00%)	2 (2.00%)
	Sadness	7 (7.00%)	0 (0.00%)	1 (1.00%)	93 (93.00%)
Speaker 4 (9 tests)	Neutral	88 (97.78%)	0 (0.00%)	0 (0.00%)	3 (3.33%)
	Anger	8 (8.89%)	75 (83.33%)	4 (4.44%)	3 (3.33%)
	Joy	9 (10.00%)	1 (1.11%)	79 (87.78%)	1 (1.11%)
	Sadness	0 (0.00%)	0 (0.00%)	2 (2.22%)	87 (96.67%)
Speaker 5 (5 tests)	Neutral	31 (62.00%)	2 (4.00%)	0 (0.00%)	18 (36.00%)
	Anger	5 (10.00%)	44 (88.00%)	0 (0.00%)	0 (0.00%)
	Joy	9 (18.00%)	7 (14.00%)	31 (62.00%)	3 (6.00%)
	Sadness	11 (22.00%)	1 (2.00%)	0 (0.00%)	38 (76.00%)
Speaker 6 (4 tests)	Neutral	24 (60.00%)	5 (12.50%)	0 (0.00%)	10 (25.00%)
	Anger	0 (0.00%)	40(100.00%)	0 (0.00%)	0 (0.00%)
	Joy	1 (2.50%)	0 (0.00%)	39 (97.50%)	1 (2.50%)
	Sadness	13 (32.50%)	1 (2.50%)	0 (0.00%)	26 (65.00%)
Speaker 7 (4 tests)	Neutral	33 (82.50%)	0 (0.00%)	1 (2.50%)	6 (15.00%)
	Anger	4 (10.00%)	35 (87.50%)	1 (2.50%)	0 (0.00%)
	Joy	1 (2.50%)	1 (2.50%)	33 (82.50%)	5 (12.50%)
	Sadness	0 (0.00%)	0 (0.00%)	0 (0.00%)	40(100.00%)
Speaker 8 (4 tests)	Neutral	30 (75.00%)	2 (5.00%)	1 (2.50%)	7 (17.50%)
	Anger	9 (22.50%)	31 (77.50%)	0 (0.00%)	0 (0.00%)
	Joy	3 (7.50%)	5 (12.50%)	29 (72.50%)	3 (7.50%)
	Sadness	6 (15.00%)	1 (2.50%)	1 (2.50%)	32 (80.00%)

it would be wiser to choose speakers 7 or 11 if we were to analyze the characteristics of sadness. And finally, as previously mentioned, we have saved the answers given by the evaluators for each utterance, which allows us to choose the best performance for each speaker. If a speaker has offered a very good performance on a given utterance, but mediocre performances on the other utterances of that sentence, it will be noticeable, even if his/her recognition rate is low for that particular expression.

3.1.3. Effect of the semantic content

As previously stated, it is impossible to totally remove any emotion/expression from meaningful sentences. In our selected sentences, words like "hurt", "reward", and negative sentences might be perceived as "emotionally negative" by the evaluators. In this subsection we will look at the data in a different light to check whether this is the case. The obtained results for each sentence will be analyzed to check whether the negative ones give more negative results.

Table VII gives the confusion matrices for each sentence. To determine whether the semantic content of the sentences does have an effect on the evaluator's responses, we can look at the recognition rates for Joy (because the three other expressive types are "less positive"), and we can also look at the number of times where Neutral is mistaken for Anger or Sadness. Looking at sentence 3's results for "joy" (which has the lowest recognition rate amongst the sentences), it does seem like the negative modality of the sentence did affect the evaluator's perception of the expressivity of the sentence.

However, sentence 8, ("Did she hurt him with words?"), despite containing a negative word ("hurt"), has the best recognition rate for "joy", even better than sentence 6, which contains the possibly positive word "reward".

Document downloaded by @DAEL, 2026-11-10 15:03:40 - personal use only

Table VI. Continuation.

		Neutral	Anger	Joy	Sadness
Speaker 9 (3 tests)	Neutral	19 (63.33%)	0 (0.00%)	0 (0.00%)	11 (36.67%)
	Anger	2 (6.67%)	27 (90.00%)	1 (3.33%)	0 (0.00%)
	Joy	3 (10.00%)	0 (0.00%)	20 (66.67%)	7 (23.33%)
	Sadness	9 (30.00%)	1 (3.33%)	0 (0.00%)	20 (66.67%)
Speaker 10 (3 tests)	Neutral	28 (93.33%)	0 (0.00%)	0 (0.00%)	1 (3.33%)
	Anger	11 (36.67%)	20 (66.67%)	0 (0.00%)	0 (0.00%)
	Joy	0 (0.00%)	0 (0.00%)	29 (96.67%)	1 (3.33%)
	Sadness	0 (0.00%)	1 (3.33%)	0 (0.00%)	29 (96.67%)
Speaker 11 (3 tests)	Neutral	27 (90.00%)	0 (0.00%)	0 (0.00%)	3 (10.00%)
	Anger	1 (3.33%)	30(100.00%)	0 (0.00%)	0 (0.00%)
	Joy	3 (10.00%)	0 (0.00%)	24 (80.00%)	2 (6.67%)
	Sadness	0 (0.00%)	0 (0.00%)	0 (0.00%)	30(100.00%)
Speaker 12 (3 tests)	Neutral	26 (86.67%)	0 (0.00%)	0 (0.00%)	4 (13.33%)
	Anger	5 (16.67%)	24 (80.00%)	1 (3.33%)	0 (0.00%)
	Joy	1 (3.33%)	0 (0.00%)	29 (96.67%)	0 (0.00%)
	Sadness	1 (3.33%)	0 (0.00%)	1 (3.33%)	28 (93.33%)
Speaker 13 (3 tests)	Neutral	24 (80.00%)	0 (0.00%)	0 (0.00%)	6 (20.00%)
	Anger	4 (13.33%)	25 (83.33%)	1 (3.33%)	0 (0.00%)
	Joy	2 (6.67%)	0 (0.00%)	22 (73.33%)	6 (20.00%)
	Sadness	1 (3.33%)	0 (0.00%)	0 (0.00%)	29 (96.67%)

As for the amount of times when the "normal" expression is mistaken with "sadness" or "anger", it also seems like semantic context might play a role in the perception of expressivity, when looking at sentence 4 ("Our epoch will hurt them"), which is confused 24% of the time with a more negative expressivity. However, sentence 6 is more often viewed as negative than sentence 8, despite the former containing the word "reward" and the latter containing the word "hurt".

From all this, it can be said that the semantic content of the sentences might play a role in the evaluators' perceptions, but that role is not clearly evident and remains fuzzy. The effect of semantic content on perception of affective expressions needs to be researched as an additional, possibly separate study. In this study, we did not control for how well the evaluators understood the sentences spoken in Modern Standard Arabic. One reason is that Algerians typically do not fluently speak and understand Modern Standard Arabic, and therefore, it is probably untrue that every evaluator understood every word in every sentence. For another thing, one evaluator was French and another was Chinese, with recognition rates of 75% and 67.5% respectively.

3.2. Reversed-signal evaluation

Due to the possible effect of the semantic context, another evaluation experiment was designed and performed which consisted in playing the recorded sentences to the evaluators "backwards", thereby ensuring that no evaluator can understand the meaning of the sentences. To the best of our knowledge, this evaluation method has never been tried in the literature. All utterances from speaker 11 were reversed in the time domain and the test was performed by 10 evaluators. The overall correct recognition rate was sensibly lower (67%), but still better than chance. A closer look at the results (Table VIII) shows the same tendencies as for the normal test, i.e. neutral is often confused with sadness, and joy is mostly confused with neutral. However, sadness becomes the least well-recognized emotion and anger becomes the most well-recognized.

The lower correct recognition rate can probably be attributed to the fact that people do not usually hear backwards speech. For instance, in addition to the evaluators not understanding the sentences, the prosodic lines which are usually decreasing in an affirmative sentence (in most languages) are now increasing.

The purpose of this test was to get rid of any semantic clue in the sentences so that the evaluators can only judge the expressivity based on acoustic cues. The rate of well-recognized emotions shows that the sentences do contain enough acoustic cues for the evaluators to categorize them correctly. Though this experiment only considered speaker 11, we assume this conclusion to be true for other speakers as well, even though the results would vary depending on the speaker and the expressive type.

3.3. "All speakers" evaluation

This test only differs from the "per-speaker" evaluation in that the evaluator does not select a speaker in the beginning of the test - instead, the speakers are selected randomly. Also, only 10 utterances are tested by each evaluator, so that the evaluation becomes a less fastidious task. This can

Document downloaded by @DAEL, 2026-11-10 15:03:40 - personal use only

Table VII. Confusion matrices for each sentence (in %).

Sentence 1	Neutral	Anger	Joy	Sadness
Neutral	77.33	1.33	0.00	22.67
Anger	18.67	74.67	6.67	1.33
Joy	28.00	1.33	62.67	8.00
Sadness	20.00	1.33	1.33	76.00
Sentence 2	Neutral	Anger	Joy	Sadness
Neutral	85.33	2.67	0.00	13.33
Anger	16.00	80.00	1.33	1.33
Joy	25.33	4.00	66.67	2.67
Sadness	9.33	0.00	0.00	90.67
Sentence 3	Neutral	Anger	Joy	Sadness
Neutral	77.33	5.33	0.00	16.00
Anger	10.67	88.00	1.33	1.33
Joy	29.33	6.67	58.67	5.33
Sadness	6.67	0.00	1.33	92.00
Sentence 4	Neutral	Anger	Joy	Sadness
Neutral	73.33	2.67	2.67	22.67
Anger	12.00	84.00	4.00	0.00
Joy	12.00	8.00	70.67	8.00
Sadness	2.67	0.00	1.33	97.33
Sentence 5	Neutral	Anger	Joy	Sadness
Neutral	74.67	0.00	1.33	24.00
Anger	6.67	89.33	2.67	1.33
Joy	8.00	4.00	77.33	10.67
Sadness	2.67	2.67	2.67	92.00
Sentence 6	Neutral	Anger	Joy	Sadness
Neutral	80.00	2.67	0.00	17.33
Anger	8.00	88.00	2.67	0.00
Joy	18.67	0.00	73.33	9.33
Sadness	2.67	0.00	0.00	97.33
Sentence 7	Neutral	Anger	Joy	Sadness
Neutral	93.33	1.33	0.00	5.33
Anger	6.67	89.33	2.67	0.00
Joy	14.67	0.00	84.00	2.67
Sadness	8.00	1.33	0.00	90.67
Sentence 8	Neutral	Anger	Joy	Sadness
Neutral	84.00	4.00	2.67	9.33
Anger	2.67	92.00	5.33	0.00
Joy	10.67	1.33	86.67	1.33
Sadness	8.00	1.33	0.00	90.67
Sentence 9	Neutral	Anger	Joy	Sadness
Neutral	90.67	1.33	2.67	4.00
Anger	8.00	88.00	4.00	0.00
Joy	10.67	5.33	76.00	6.67
Sadness	6.67	0.00	0.00	93.33
Sentence 10	Neutral	Anger	Joy	Sadness
Neutral	70.67	0.00	5.33	25.33
Anger	6.67	92.00	0.00	1.33
Joy	10.67	2.67	84.00	2.67
Sadness	6.67	1.33	2.67	89.33

Table VIII. Confusion matrix of the obtained results for the reversed-signal test (in %).

	Neutral	Anger	Joy	Sadness
Neutral	65	3.75	0	30
Anger	5	87.5	3.75	3.75
Joy	21.25	3.75	67.5	7.5
Sadness	30	20	5	64.25

Table IX. Confusion matrix of the obtained results for the reversed-signal test (in %).

	Neutral	Anger	Joy	Sadness
Neutral	87.5	3.125	0	9.38
Anger	29.17	70.83	0	0
Joy	17.39	8.70	65.22	70
Sadness	14.29	0	9.52	76.19

be considered a more difficult version of the "per-speaker" test, because the evaluator is now unable to simply memorize the 4 types of expressive speech of one evaluator. This test is therefore more reliable and accurate. The overall confusion matrix for 10 evaluations) is given by Table VIII.

The results are again lower than for the "per-speaker" test (with an average of 74.94%), while still showing that the recorded corpus is reliable.

More analyses can be performed using the evaluation results, because many details are saved as the evaluator's answers are provided. All of the SQL databases (containing the tests' answers) will be freely accessible along with the audio files.

4. Basic speech synthesis application

Since the corpus was primarily built in order to help us gather data for an expressive speech synthesis system, we will briefly describe, in this section, the production of synthetic voices for a text-to-speech (TTS) application. The approach is similar to that of Iida *et al.* in [34], and even more so to that of Bulut *et al.* in [35], i.e. diphones are recorded in various voice qualities corresponding to various expressivities. During synthesis of an expressive sentence, e.g. anger, only diphones from the inventory of anger are used. Since controlling the voice quality is difficult in unit concatenation text-to-speech systems, this choice has provided good results.

In this work, we recorded Arabic diphones for "neutral" and "sad" expressivities, that is, they were uttered with a normal voice quality for the "neutral" expressivity, and with a more brittle and slightly high-pitched voice for the "sad" expressivity, by speaker 1 of the aforementioned speakers (cf. section 2.3). Unlike in Bulut's work, the diphones are recorded in a monotone voice, i.e. the fundamental frequency (f0) is the same for all diphones, and the diphone duration is constant. A common approach for the recording of diphones is to include the desired diphone as

Document downloaded by @DAEL, 2026-11-10 15:03:40 - personal use only

Table X. Mean expressivity ratings for the various combinations of speech parameters.

Expressivity	Neutral prosody	Sad prosody
Neutral V.Q	1.49	2.15
Sad V.Q	2.88	4

Table XI. Mean naturalness ratings for the various combinations of speech parameters.

Naturalness	Neutral prosody	Sad prosody
Neutral V.Q	3.83	3195
Sad V.Q	3.25	3.06

the middle syllable of a nonsensical word. Our approach was to form a nonsensical word for each consonant (or silence), with each of the Arabic vowel sounds included in that word. For instance, for letter "ب" (Arabic sound "b"), the nonsense word would simply be "babubibɛb", for letter "ت", "tatutitɛt", and so on.

The diphones were then concatenated via a TTS system that outputs monotone speech, to produce the 10 previously cited sentences (Section 2.2), both with "neutral" and "sad" diphones. The prosody was then manually and carefully applied for each sentence to these monotone synthesized utterances by copying the natural prosody of the recorded utterances of the database using Praat. Four combinations were tested: neutral prosody + neutral voice quality, neutral prosody + "sad" voice quality, sad prosody + neutral voice quality, and sad prosody + sad voice quality.

These resulting expressive utterances[6] were evaluated by 10 Algerian listeners (some of whom had already passed the first test) who were asked to rate, on a scale from 1 to 5, the "sadness" and the "naturalness" of the produced utterances. For the rating of "sadness", 1 is equivalent to a completely neutral utterance, and 5 is equivalent to a completely sad utterance. For the rating of "naturalness", 1 is a very robotic, unnatural voice, whereas 5 is a very "human" and natural voice. The results are given in Tables VIII and IX.

Table VIII shows that the expressivity is better recognized when the voice quality matches the prosody. The ratings are closer to 1 when the neutral prosody is used in conjunction with a neutral voice quality, and they are closer to 5 when the sad prosody is used in conjunction with a sad voice quality. The ratings for the combinations "sad prosody + neutral V.Q." and "neutral prosody + sad V.Q." are in between these two mean ratings, which was predictable. However, it seems that voice quality plays a slightly more important role in conveying "sadness" than prosody, as the mean for "sad V.Q. + neutral prosody" is slightly closer to 5 than the "neutral V.Q. + sad prosody" combination. This is in agreement with previous studies such as [36].

Regarding naturalness, the mean ratings are close to each other, with "Neutral prosody + neutral V.Q" being the most natural utterance and "Sad prosody + sad V.Q." being the least natural. This might be because the production of sadness required more processing during the transformation phase using Praat.

5. Future works

This expressive speech corpus can be improved in many aspects.

The corpus would certainly benefit from the inclusion of more professional speakers, especially when noticing that acted expressive speech can be more convincing than real emotions (see section 3.2). The number of speakers could also be increased in order to diversify the speech corpus. Speakers from other Arabic countries would perform the expressions with other accents, therefore providing a richer database. The corpus would also obviously benefit from the recording of different sentences that would bring new phonemic and phonetic contexts. Also, maybe some other sentences should be designed in order to avoid emotionally biased sentences such as "did she hurt him with words?". Unfortunately, designing a set of sentences that is both emotionally neutral and phonetically balanced seems like a challenging task.

Another obvious improvement would be to increase the number of simulated expressions. We have focused, for this first work, on basic primary expressive speech types, because they are simpler to simulate for non-professional persons, but the availability of more expressive speech types would evidently improve the speech corpus.

In [37], Campbell insists on the importance of spontaneity in expressive corpora and underlines the difference of pronunciation between sentences read by a speaker that distances him/herself from the content of the utterances, and those uttered naturally in everyday life. This suggests another, more natural recording of the "Neutral" expression in order to have both versions available.

Regarding the evaluation phase, other evaluation methods could also be performed. A guided categorization test does not necessarily reflect the expressive speech actually perceived by the evaluators. In this kind of test, the evaluator might proceed by elimination in order to find the correct answer. He/she already knows that the presented expressive speech is one of the presented answers. A possible improvement on this point should therefore be to perform a non-guided categorization test, or to add other expressive speech types to the possible answers (e.g, tiredness, fear), in order to give the evaluators more choices.

Up until now, the test has been administered only to Algerian evaluators with the exception of one French evaluator, and a Chinese one, both of whom did not understand the meaning of the sentences. An idea of improvement would therefore be to diversify the linguistic and cultural

[6] http://voice-research.hostingsiteforfree.com/synth/ www.voice-research.org

Document downloaded by @DAEL, 2026-11-10 15:03:40 - personal use only

backgrounds of the evaluators. This would be helpful in order to compare the results between people who understand Arabic and those who do not, and to determine to what extent does the semantic content influence the perception of expression. Finally, the number of evaluations should be increased, even though the global recognition rate seems to stabilize itself between 80 and 85%. The implemented website allows these improvements to be easily realized, even though finding volunteers to evaluate the corpus is way harder than expected.

6. Conclusion

We hope that the recording of this expressive speech corpus will help in the advancement of the synthesis and recognition of expressive speech in the Arabic language, and for the field of affective computing in general. This modest contribution, even though it can be improved in many ways, offers a corpus characterized by a good recognition rate of the expressive speech types it contains. The corpus will be made public and freely available (along with the Praat annotations and evaluation results), thus helping the Arabic expressive speech research community. It is, for us, a first step towards the synthesis of expressive speech in the Arabic language.

Acknowledgement

The authors wish to acknowledge the speakers for their patience and efforts during the recording process and the evaluators for the time they spent evaluating the sentences. We also thank Y. Remram who kindly allowed us to use the aforementioned anechoic chamber.

References

[1] M. Boudraa, B. Boudraa, B. Guerin: Twenty lists of ten Arabic sentences for assessment. Acta Acustica united with Acustica **86** (2000) 870–882.

[2] R. W. Picard: Affective computing. MIT Media Laboratory Perceptual Computing Section Technical Report No. 321, 1995.

[3] G. Beller: Analyse et modèle génératif de l'expressivité. Thèse Ph.D, Univ. Paris VI, France, p.5, p.125, 2009.

[4] E. Douglas-Cowie, N. Campbell, R. Cowie, P. Roach: Emotional speech: Towards a new generation of databases. Speech Communication **40** (2003) 33–60.

[5] D. Ververidis, C. Kotropoulos: Emotional speech recognition: Resources, features, and methods. Speech Communicatione **48** (2006) 1162–1181.

[6] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRoie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, K. Karpouzis: The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. – In: Proc. ACII, LNCS 4738,. Springer, Berlin Heidelberg, Germany, 2007, 488–500.

[7] M. Grimm, K. Kroschel, S. Narayanan: The Vera Am Mittag german audio-visual emotional speech database. Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hannover, Germany, 2008.

[8] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. 10th IEEE Int. Conf. Work. Autom. Face Gesture Recognit.,, Apr. 2013, 1–8.

[9] O. Al-Dakkak: Emotion inclusion in an arabic text-to-speech. 13th European Signal Processing Conference, Antalya, Turkey, September 4-8, 2005.

[10] O. Al-Dakkak: Prosodic feature introduction and emotion incorporation in an Arabic TTS. Information and Communication Technologies, ICTTA '06, 2006.

[11] A. F. El Gohary, T. I. Sultan, M. A. Hana, M. M. El Dosoky: A computational approach for analyzing and detecting emotions in Arabic text. International Journal of Engineering Research and Applications (IJERA) **3** 100–107.

[12] O. Alsharif, D. Alshamaa, N. Ghneim: Emotion classification in Arabic poetry using machine learning. International Journal of Computer Applications **65** 10–15, March 2013.

[13] O. Rabie, C. Sturm: Feel the heat: Emotion detection in Arabic social media content. Proceedings of the International Conference on Data Mining, Internet Computing and Big Data, Kuala Lumpur, Malaysia, 2014.

[14] E. Refaee, V. Rieser: Can we read emotions from a smiley face? emoticon-based distant supervision for subjectivity and sentiment analysis of Arabic twitter feeds. 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data, LREC'2014.

[15] Y. Alotaibi, A. Meftah, S. A. Selouani: Investigation of emotion classification using speech rhythm metrics. Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE), 2013, 204–209.

[16] M. Meddeb, H. Karray, A. M. Alimi: Intelligent remote control for TV program based on emotion in Arabic speech. International Journal of Scientific Research & Engineering Technology (IJSET) (2014) Vol. 1.

[17] W. M. Azmy, S. Abdou, M. Shoman: Arabic unit selection emotional speech synthesis using blending data approach. International Journal of Computer Applications **81** (2013) 22–28.

[18] W. M. Azmy, S. Abdou, M. Shoman: The creation of emotional effects for an Arabic speech synthesis system. The Egyptian Society of Language Engineering, International Workshop, ESOLE'13, Cairo, 12/2013.

[19] T. Shochi, D. Erickson, A. Rilliard, V. Aubergé, J. C. Martin: Recognition of Japanese attitudes in audio-visual speech. Speech prosody (2008) 689–692.

[20] T. Shochi, G. Gagnié, A. Rilliard, D. Erickson, V. Aubergé: Learning effect of prosodic social affects for Japanese learners of French language. Proceedings of Speech Prosody, 2010, 155.

[21] A. Rilliard, J. A. Moraes, D. Erickson, T. Shochi: Prosodic analysis of brazilian portuguese attitudes. Proc. of the Speech Prosody Conf., 2012.

[22] A. Rilliard, D. Erickson, T. Shochi, J. A. de Moraes: US English attitudinal prosody performances in L1 and L2 speakers. Proc. of the Speech Prosody Conf., 2014.

[23] M. a. M. Abushariah, R. N. Ainon, R. Zainuddin, O. O. Khalifa, M. Elshafei: Phonetically rich and balanced arabic speech corpus: An overview. Int. Conf. Comput. Commun. Eng., 2010, 1–6.

[24] K. Almeman, M. Lee, A. A. Almiman: Multi dialect Arabic speech parallel corpora. 1st Int. Conf. Commun. Signal Process. their Appl., Feb. 2013, 1–6.

Document downloaded by @DAEL, 2026-11-10 15:03:40 - personal use only

[25] M. O. Eltayeb, M. E. Mustafa: Acoustic-support vector machines approach to detect spoken Arabic language. Int. Conf. Comput. Electr. Electron. Eng., Aug. 2013, 525–529.

[26] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss: A database of German emotional speech. Proceedings of Interspeech, Lissabon, 2005.

[27] Y. Remram, H. Khelladi: Construction of an inexpensive anechoic chamber for teaching aid. Proceedings of INTER-NOISE 2009, Ottawa, Canada, 23-26 August 2009.

[28] M. Assaf: A prototype of an Arabic diphone speech synthesizer in festival. Master's Thesis, Univ. Uppsala, Sweden, p. 26, 2005.

[29] A. W. Black, K. A. Lenzo: Building synthetic voices. Language Technologies Institute, Carnegie Mellon University and Cepstral, LLC, p.95, 2003.

[30] J. Panksepp, J. Burgdorf: 'laughing' rats and the evolutionary antecedents of human joy? Physiology & Behavior **79** (2003) 533–547.

[31] M. Bekoff: The emotional lives of animals: "a leading scientist explores animal joy, sorrow and empathy-and why they matter". New world library, 2008.

[32] P. Ekman, W. V. Friesen, M. O'Sullivan: Smiles when lying. Journal of personality and social psychology **54** (1988) 414.

[33] K. R. Scherer: What are emotions? And how can they be measured? Social Science Information **44** (2005) 695–729.

[34] A. Iida, N. Campbell, F. Higuchi, M. Yasumura: A corpus-based speech synthesis system with emotion. Speech Communication **40** (2003) 161–187.

[35] M. Bulut, S. S. Narayanan, A. K. Syrdal: Expressive speech synthesis using a concatenative synthesizer. ICSLP, 2002.

[36] I. Grichkovtsova, A. Lacheret, M. Morel: The role of intonation and voice quality in the affective speech perception. Interspeech, 2007, 2245–2248.

[37] N. Campbell: Databases of expressive speech. J. Chinese Lang. Comput. **14** (2004) 295–304.

Document downloaded by @DAEL, 2026-11-10 15:03:40 - personal use only