*INTERSPEECH 2017*
August 20–24, 2017, Stockholm, Sweden

# Building audio-visual phonetically annotated Arabic corpus for expressive text to speech

*Omnia Abdo*[1], *Sherif Abdou*[2], *Mervat Fashal*[1]

[1]Alexandria University, Phonetics and linguistics department, Egypt
[2]Cairo University, Faculty of Computers & Information, Egypt

omnia.abdo@alexu.edu.eg , s.abdou@fci-cu.edu.eg , mervat.fashal@alexu.edu.eg

## Abstract

The present research aims to build an MSA audio-visual corpus. The corpus is annotated both phonetically and visually and dedicated to emotional speech processing studies. The building of the corpus consists of 5 main stages: speaker selection, sentences selection, recording, annotation and evaluation. 500 sentences were critically selected based on their phonemic distribution. The speaker was instructed to read the same 500 sentences with 6 emotions (Happiness- Sadness- Fear- Anger- Inquiry - Neutral). A sample of 50 sentences was selected for annotation. The corpus evaluation modules were: audio, visual and audio –visual subjective evaluation.

The corpus evaluation process showed that happy, anger and inquiry emotions were better recognized visually (94%, 96% and 96%) than audibly (63.6%, 74% and 74%) and the audio visual evaluation scores (96%, 89.6% and 80.8%). Sadness and fear emotion on the other hand were better recognized audibly (76.8% and 97.6%) than visually (58% and 78.8 %) and the audio visual evaluation scores were (65.6% and 90%).

**Index Terms**: expressive audio-visual corpus, human-computer interaction, modern standard Arabic resource, under-resourced language resource

## 1. Introduction

Recent advancement in human-computer interaction (HCI) technology goes back to the successful transfer of data between humans and the machine, improving thus user friendly and natural interactions. Human speech perception is bimodal in nature; humans combine audio and visual information in deciding what has been spoken. Integration of audio and video signals for speech processing systems has become an important field of study. In text to speech, the generated speech needs to be intelligible to listeners and have a natural expressive melody.

Informal survey was held with professional researchers in the field of speech processing, computer vision and animation industry. The survey was gathered through meeting, emails and telephone calls. The aim of the survey was to collect and determine the specific problems which face expressive talking heads. The collected information helped the researchers in determining the present corpus specifications.

Unfortunately, Arabic expressive TTS systems face a number of challenges; the corpora are sometimes recorded by unprofessional actors, which affect the quality of the generated speech. Another problem is the lack of enough recorded multimodal data of emotions disabling to the generalization of facial and acoustic parameter patterns corresponding to various emotions.

One of the crucial improvements that are needed to achieve major progress in the speech processing field is the collection of new corpus that tries to overcome the limitations of the existing corpora. A survey on the existing emotion corpora modules was held. It reviews 35 text emotional corpora, audio emotional corpora, visual emotional corpora, and finally audio-visual emotional corpora.

Some examples of emotional corpora:

- Chinese emotion blog corpus (text corpus): contained eight emotional categories and useful for research in emotion text expressions in Chinese [1].
- KSUEmotion corpus (audio corpus): includes natural, sadness, happiness and question emotions. 20 speakers record 10 newspaper sentences in an uncontrolled environment [2].
- The Indian Spontaneous Expression Database (visual corpus): consists of 428 annotated video clips and can be used in spontaneous expressions recognition [3].
- Japanese Audio-Visual Emotion Database (audio – visual corpus): is the first audio-visual emotion corpus in Japanese and contains 100 minutes of annotated recording [4].

Previous stated works have been limited to collecting from languages other than Arabic, such as English, German …etc. The present study builds the first Modern Standard Arabic audio visual expressive corpus which is annotated both visually and phonetically. The recorded corpus is dedicated to research in expressive speech processing field and especially to Text to speech application.

## 2. Corpus building

Accurate corpus design is one of the primary aspects in building high quality expressive text to speech systems. And the procedures of collecting and annotating the corpus are the vital part of the process.

### 2.1. Speaker selection stage

Selection of the corpus' speaker is one of the problematic issues in designing a TTS system [5]. Accordingly critical speaker selection process had to be held with subjective evaluation of expressing the right emotion. Therefore, a mini experiment for selecting a speaker was carried out in order to elucidate the speaker who is suitable for the present corpus.

7 speakers were participated in the selection experiment. Their ages ranged from 20:35 years old. They have different backgrounds (professional actors and students



http://dx.doi.org/10.21437/Interspeech.2017-1357

in theater department). They are well trained actors. All the participants have no hearing or articulation deficit.

Three sentences, extracted from the final 500 corpus, were chosen for the speaker selection experiment and were: one short in length (3 words), one medium (6 words) and the third was a long one (9 words) to reflect different sentences' length of the final corpus.

Each participant preformed the three sentences with the four emotional states (Happy, Sad, Anger and Fear).

Ten human raters were chosen for the speaker selection experiment. The raters' ages range from 20:60 years old. They were recruited to evaluate the expressive emotion of the seven participants and were asked to subjectively rate 84 utterances. They watched the speakers' video and then record their responses in a sheet of paper. The subjects could play each stimulus as many times as they wished.

Based on the results, the speaker number 2 has the highest perceptual scores in the subjective evaluation in all the emotional states. The selected speaker is professional actor and proficient MSA speaker. In addition to all of that he has more than 12 experience years in acting and direction.

### 2.2. Sentences selection stage

The present corpus has to represent the acoustic properties of all sounds that occur in MSA to be used as a baseline for training expressive text to speech systems. This section presents the methodology of sentences selection procedure for phonetically balanced speech corpus. The following chart represents the steps.

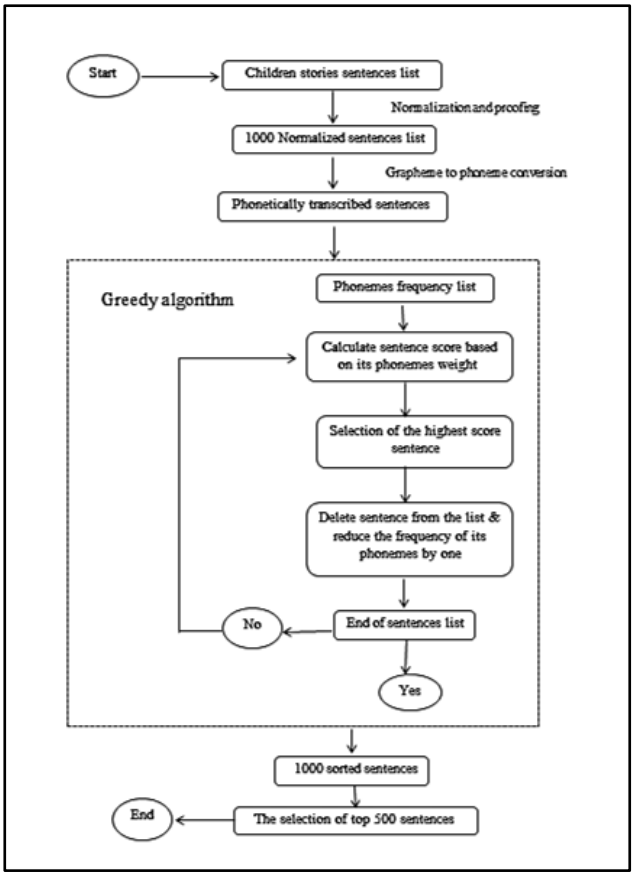


Figure 1: *Sentence selection procedure using greedy algorithm.*

The first important issue to consider when collecting sentences from stories is to obtain copyright from the authors. For the current study, the recorded sentences were selected from 10 children stories. Copyright of those stories are in the public domain.

The second step in the selection procedure is normalize and proofing the raw sentences. A professional Arabic linguist manually added full diacritical marks to the written sentences. The reason for that is to avoid any ambiguity in pronunciation and ensure correct articulation. For example see the following sentence:

- Original sentence: لا يستدعي الأمر مني لحظة من التفكير(I don't think twice about it)
- After adding diacritics: لَا يَسْتَدْعِي الأَمْرُ مِنِّي لَحْظَةً مِنِ التَّفْكِيرْ

Sentences which have difficult pronounced word or awkward grammar have been eliminated. To ensure reading fluency, the length of sentences is between minimum 4 to 10 words maximum.

Grapheme to phoneme system converted the diacratized sentences into a string of phonemes. Then, the greedy selection algorithm [6] has been used in the present corpus. This is an optimization technique for constructing a subset of sentences.

The output of applying the greedy selection algorithm is 1000 balanced sorted sentences list. Then the top 500 sentences were chosen to be recorded with each emotion resulting 3000 recorded file.

### 2.3. Recording stage

This stage is divided into 3 sub-stages: Pre-recording, Recording and Post-recording phases.

#### *2.3.1. Pre-recording phase*

This phase is not a part of the recorded corpus itself; it is the only way to eliminate all the bugs and problems in the proposed procedure.
The following section will briefly describe the recording procedure which was followed in order to capture the corpus from the speaker. It describes the facial motion capture technique, the setting of the laboratory and the recording equipment.
**First: Facial motion capture:** The marked-based facial motion capture technique was used in this study based on the Psychological and physiological evidences [7], therefore the researcher choose 29 marks and placed them on the movable facial parts for each emotion. The chosen color was blue to distinguish the marks from the green background.
**Second: Laboratory setup:** The dynamic of every facial expression during speech as well as the respective acoustic signal was captured on a profession audio studio (dimensions 3.5 X 2.9 m). The studio was designed to eliminate any outside environmental noise.

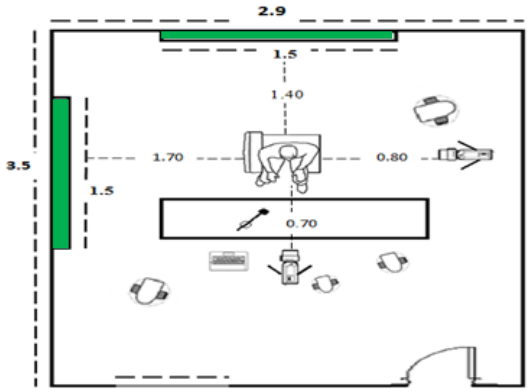

Figure 2: *laboratory setting*.

The studio room was readjusted in certain specification in order to be suitable for the recording process as follows:

The microphone was located on approximately 40 cm distance from the speaker's lips in order to minimize the distortion of the acoustic sound signal. The two recorded cameras were placed approximately 70 cm distance from the speaker. One camera was placed in the front of the speaker and the other in his left side on 90 degree. For lighting the studio, the primary light was fixed on height level of the speaker's face and the secondary light on 45 degrees to the right of the speaker. The background has been illuminated by a separated light source at 1.70m height. Green backdrops covered the studio's walls (speaker's background). The speaker was placed in the center of the studio with 1 m distance from cameras and 1.5 m from background.

**Third: Recording equipment:** The following section illustrates the used recording equipment: microphone, cameras and light equipment.

The selection of particular microphone depends heavily on the intended usage. Hence, high quality stereo microphone "Zoom H4n Handy Recorder" was used in the present corpus. Zoom H4n Handy Recorder has a couple of advantages, first its stereo microphones are direct to sound source (speaker's voice) which helps eliminating any background noise.

Two full HD cameras were used for recording: Canon LEGRIA HF R 506 and Nikon D3200. The resolution of the two cameras is 1920 X1080 pixels with 30 fps/ 24 Mbps. The two full HD camcorders were placed at the level of the speaker's eye height (~ 120 cm from the floor).

While recording, some light sources were needed to illuminate the speaker's face. So at the pre- recording stage, the researcher tried various types of light sources, some were too bright and other produced background noise.

For the studio illumination, ARRI 650 Plus is ideal for use especially in small spaces. Its wide angle lenses helps producing well light distribution.

The goal is to eliminate shadows and create soft, flat lighting on the speaker's face. Because the top light source causes a dark shadow under the speaker's eyes and nose, two light sources were needed to light the speaker's face; one key (primary) light was fixed on front of the speaker behind the main camera and fill (secondary / assistant) light on 45 degrees to the right of the speaker. The background has been illuminated by a separated light source behind the speaker used to distinguish the speaker from the background and avoiding any shadows in the background at 1.70m height. The following figures illustrate the light sources placement.


Figure 3: light sources placement.

### *2.3.2. Recording phase*

Recording held in the studio described above. Each pure hour of speech (~500 sentences) will take one recording session (1 day). The speaker was placed in the center of the studio with 1 m distance from cameras and 1.5 m from background (to prevent shadows). The corpus recording took 6 sessions each last for 4 or 5 hours.

The recorded emotions were happiness, sadness, fear, anger, inquiry and neutral state.

### *2.3.3. Post recording phase*

In the previous phase, 500 sentences were recorded per emotion. There were number of video files from the two cameras and audio files from the H4n microphone. The current phase has two main parts. The first part was synchronization of the acoustic signal and video. It was done by finding a clapper sound at the beginning of the video file.

Then match video audio track with the microphone recording. Finally after matching, the video sound was muted leaving only the audio of the microphone.

The second step is that the video files were manually splitted into individual sentences. The recorded audio was saved in (.wav) file format with bit depth: 16 bits and sampling: 48 kHz, while the recorded video was saved as AVCHD (.mov) file.

## 3. Annotation stage

Speech processing systems like TTS or ASR strongly rely on the availability of well-designed speech corpora. For their development, the speech corpus should be combined with information (full or partial) about its contents and its phonetic units (syllable, word, phrase, etc.). The recorded corpus was visual and acoustical annotated.

There are nine visemes in Arabic language. Those visemes were extracted from the recorded sentences in the present study. The visemes duration was also calculated and compared for the recorded emotions. The fear emotion presents the highest value because it has a slow speaking rate.

The entire recorded corpus was transcribed orthographically because of many reasons; first, it provides further researchers with a simple symbolic representation of the recorded data. With this representation it is easy to navigate through the corpus. Secondly, the orthographic transcription formed the basis for all other transcriptions and annotations.

Using stratified sampling method, a sample of 300 recorded data (50 sentences X 6 emotion states) was selected for further analysis and annotation using Praat software. The sampled sentences were phonemically transcribed and timed-alignment using IPA symbols. The researcher used waveform, spectrogram and auditory discrimination cues in determining phonemes boundaries.

Considering the intended application of this corpus (developing expressive TTS), the researcher used ToBI system for intonation transcription. Five basic pitch accents symbols have been found in the recorded corpus:

- 2 monotones: H* and L*
- 3 Bi-tones: L+H*, L*+H and H!

Six boundary tones:

- 2 phrase boundary tones: L- and H-
- 2 final intonation boundary tones: L% and H%
- 2 initial intonation boundary tones: %L and %H

Break indices distribution is as follows: Sad emotion has higher rate of occurrence of the normal "1" break index than other recorded emotions resulting from slow articulation. Silence gaps within sentences cause the appearance of "2" and "3" break indices in sad and fear emotions. Happy, anger and inquiry emotions have higher occurrence of "0" break index resulting from their fast speaking rates.

Praat speech analysis software, version 5.3.78 [8], was used in acoustic annotation in this research. In Praat, the TextGrid in this corpus consists of six tiers as follows:

1. Phonemes tier.
2. Visemes tier.
3. Syllable tier.
4. Tones tier.
5. Orthography tier.
6. Break indices tier.

ELAN software, version 4.9.3 [2], was used in video annotation. It enables importing transcription from Praat TextGrid files; it links Praat annotations to video timeline.


Figure 4: *ELAN example of the combination of two videos with Praat TextGrid.*

## 4. Corpus evaluation

In order to assess the quality of the corpus, we resorted to human observers for evaluating expressed emotions in three modules: audio only, visual only and audio visual.

The goal was to collect five subjects for each sentence, therefore 90 human raters were chosen for the corpus evaluation process, and each one rated 50 sentences of the sample in one module. The raters' ages range from 18:40 years old with a mean age of 20. They were recruited to evaluate the expressive emotion of the speaker.

The rating process was held in a quiet room with day light. The raters were exposed to the stimuli (watch the video, list to the audio file or watch the muted video) and then they were instructed to choose the expressed emotion and record their responses in a sheet of paper.

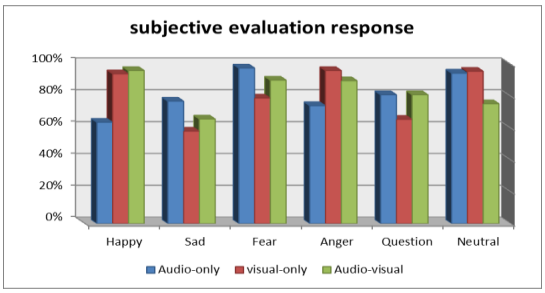


Figure 5: *subjective evaluation response.*

The visual smile has a great contribution in the recognition of happiness as an emotion in visual alone and audio-visual evaluation. The audio- evaluation lacks this visual cue and has low score of correct rating responses. Sad emotion is combined with pause or lengthening near the end of the utterance. Visually, sad emotion lacks a distinctive visual feature and sometimes is misrecognized as neutral emotion. That explains higher audio responses than other modules. Fear emotion is also characterized by the repetition of words or phonemes. Most of the inquiry sentences have a clear pause towards the beginning of the utterance, and also have a high rising tone at the end.

Table 1: *Audio - visual subjective evaluation results*

| | Happy | Sad | Fear | Anger | Inquiry | Neutral |
|---|---|---|---|---|---|---|
| Happy | 96 | 0 | 0 | 0.4 | 0.8 | 2.8 |
| Sad | 1.2 | 65.6 | 2.4 | 2.8 | 3.2 | 24.8 |
| Fear | 0 | 9.6 | 90 | 0.4 | 0 | 0 |
| Anger | 0 | 4.4 | 0 | 89.6 | 0 | 3.6 |
| Inquiry | 1.6 | 1.2 | 0.8 | 2.8 | 80.8 | 12.8 |
| Neutral | 0.4 | 17.2 | 3.6 | 2.4 | 0.8 | 75.2 |

The confusion between intended recorded emotion and subjective ratting was examined. When raters misrecognized the intended emotion, they perceived it as neutral in the audio evaluation. This was not the case in the other modules; fear and anger were misrecognized for sadness (19.2% and 2.8% respectively) in visual evaluation whereas happiness is misrecognized for inquiry (3.2 %). In audio visual evaluation, the recorded emotions were mistaken for sad or neutral emotions.

## 5. Conclusions

The present study aims at building high quality expressive audio-visual modern standard Arabic corpus for text to speech; to achieve this goal it requires establishing accurate procedures which are represented in 6 main stages. The recorded audio-visual corpus is contributing to the field of speech processing specifically TTS applications, and can generally be used in various applications: expressive animation, human - robotics interactions, and game development. The collected corpus and the annotation will be publically available for general research purposes.

# 6. References


[1] C. Quan and F. Ren. “A blog emotion corpus for emotional expression analysis in Chinese”, Computer Speech & Language, Vol. 24, Issue 4, pp. 726-749,2010

[2] A. Meftah, Y. Alotaibi, S. Selouani, "Designing, Building, and Analyzing an Arabic Speech Emotional Corpus", ninth international conference on language resources and evaluation, at Reykjavik, Iceland, 2014

[3] S L Happy, P. Patnaik, A. Routray, R. Guha , "The Indian Spontaneous Expression Database for Emotion Recognition", IEEE Transactions on Affective Computing (Volume: PP, Issue: 99 ), 2015

[4] N. Lubis, R. Gomez, S. Sakti, K. Nakamura, K. Yoshino, S. Nakamura , and K. Nakadai, "Construction of Japanese Audio-Visual Emotion Database and Its Application in Emotion Recognition", LREC conference proceedings 2016

[5] Busso, Shrikanth, S Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances", 7th International Seminar on Speech Production, 2006.

[6] L. Buchsbaum and J. P. H. van Santen. "Selecting training inputs via greedy rank covering". In Proceedings of the 28th ACM Symposium on Theory Of Computing (STOC), pages 288–295, Philadelphia, PA, USA, 1996

[7] Ekman, P., Sorenson, E.R., Friesen, W.V. "Pan-cultural elements in the Facial Display of Emotions. Science", 164, 86-88, 1969

[8] P. Boersma, D. Weenink, "Praat: doing phonetics by computer [Computer program]". Version 5.3.78, from http://www.praat.org/, 2015

[9] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, H. Sloetjes,
"ELAN: a Professional Framework for Multimodality Research".
In: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation, 2006.