King Saud University Emotions (KSUEmotions) Corpus

Authors:
Ali Hamid Meftah[1], Yousef Ajami Alotaibi[1], Sid-Ahmed Selouani[2]

Affiliations:
[1]College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
[2]Université de Moncton, 218 bvd. J.-D.-Gauthier, Shippagan, E8S 1P6, Canada
{ameftah, yaalotaibi}@ksu.edu.sa, selouani@umcs.ca

==========

KSUEmotions corpus was recorded for Modern Standard Arabica (MSA) using 23 speakers (10 males and 13 females) from three Arabic countries: Yemen, Saudi Arabia and Syria. Recording took place in two phases.

In Phase 1, 10 male speakers were selected from Saudi Arabia, Yemen, and Syria and also 10 female speakers from two countries Saudi Arabia and Syria. All speakers reading the 16 MSA sentences selected from the originally corpus King Abdulaziz City for Science and Technology Text-To-Speech Database (KTD) [1]. In this phase, Neutral, sadness, happiness, surprise, and questioning emotions were selected. Questioning was considered as an emotion because it was incorporated in the corpus originally used (KTD) [1].

To evaluate Phase 1 recordings, a blind human perceptual test was performed. Nine listeners (6 males and 3 females) were involved to listen the recorded files to test whether were able to recognize the recorded emotion. According to the results of the human perceptual test, and by avoiding defective speakers and/or files, and to ensure uniformity among different variables such as speakers' gender, Phase 2 was produced. 7 male speakers from Phase 1 and 7 female speakers (4 of them from Phase 1 and the other new three female speakers were added from Yemen nationality) and 10 sentences were chosen for Phase 2. In this phase, he questioning emotion was excluded while the anger emotion was added to this corpus to be more consistent with other similar corpora in the field. In Phase 2 each sentence was spoken over two trials. The total duration of the all recorded files was 2 hours and 55 minutes for Phase 1 and 2 hours and 15 minutes for Phase 2. Again, a blind human perceptual test was performed for Phase 2 with the same nine listeners who reviewed Phase 1. PRAAT software [2] was used for the KSUEmotions corpus recording process.

A. Design of prompt texts

We selected 16 sentences from the KTD corpus, as shown in Table 1. The KTD corpus contains many different types of sentences, but we selected only those sentences that simulated the four targeted emotions without any increase or decrease in the word count.

All 16 sentences were selected from a newspaper that can be accessed either visually or aurally through a variety of different media. In the case of the "questioning" emotion, we added the word **" هل/hal/**." In addition, for this emotion the shortest sentences contain four words, and the longest sentences contain 16 words.

All Arabic phonemes are included, and their frequencies are relatively representative of Arabic, with /a/ and /l/ being the most frequent phonemes while phoneme /ðˤ/ and / ʁ / had the lowest frequency.

Based on the analyzed results of the human perceptual test related to the sentences, in Phase 1, we select the sentences that obtained the highest recognition rate and do not include those that obtained the lowest recognition rate, we have also added two single words, namely, yes and no. Table 1 shows all sentences, words, and phase that used in, and the words statistics are shown in Table 2.

Table 1. Selected sentences

Sen. ID	Sentence in Arabic and IPA symbols	Phase
S1:	إِصَابَةٌ جَدِيدَةٌ بِشَلَلِ الْأَطْفَالْ، وَأَرْبَعُمِئَةَ وَخَمْسَةَ عَشَرَ بِالْجُذَامْ فِي الْيَمَنْ. ?isʔaabatun ʒadiidatun biʃalalili l?atʔfaal wa?arbaʕumi?ata waxamsata ʕaʃara bilʒuðaam fil jaman	1
S2:	الْعَدَدُ الْكُلِّيِّ لِمَرْضَى اَلْجُذَامْ، بَلَغَ مَعَ مَطْلَعِ الْعَامِ الْجَارِي، سَبْعَةَ آلَافٍ وَتِسْعِمِئَةٍ وَثَمَانِيَةً وَعِشْرِينَ حَالَةْ. ?alʕadadul kullijji limardʔal ʒuðaam balaʁa maʕa matʔlaʕil ʕaamil ʒaarii sabʕata ?aalaafin watisʕimi?atin waθamaaniijatan waʕiʃriina ħaalah	1
S3:	وَفَاةُ الشَّيْخِ الْغَزَالِيِّ، فِي الْمَدِينَةِ الْمُنَوَّرَةْ، فِي شَهْرِ مَارِسْ، عَامَ أَلْفٍ وَتِسْعِمِئَةٍ، وَسِتَّةٍ وَتِسْعِينْ. wafaatuʃ ʃajxil ʁazaalijj fil madiinatil munawwarah fii ʃahri maaris ʕaama ?alfin watisʕimi?atin wasittatin watisʕiin	1
S4:	وَفَاةُ اَلشَّيْخ جَادِ الْحَقْ، وَدَفْنُهُ فِي قَرْيَتِهِ بِطُرَّةْ، بِمَرْكَزِ طَلْخَا، بِالدَّقَهْلِيَّةْ. wafaatu ʃʃajx ʒaadi lħaq wadafnuhu fii qarjatihi bitʔurrah bimarkazi tʔalxaa biddaqahlijjah	1
S5:	الشَّيْخُ الشَّعْرَاوِيُّ، يُوَارَى الثَّرَى فِي دَقَادُوسْ. ?a ʃʃajxuʃ ʃaʕraawijj juwaraθ θaraa fii daqaaduus	1 & 2
S6:	زَغْلُولِ النَّجَّارْ، يَتَكَلَّمُ عَنْ زِلْزَالِ تْسُونَامِي. zaʁluulin naʒʒaar jatakallamu ʕan zilzaalit suunaamii	1 & 2
S7:	أَحْمَدْ فُؤَادْ بَاشَا، عُضْوًا بِمَجْمَعِ الْخَالِدَيْن فِي اَلْقَاهِرَة. ?aħmad fu?aad baaʃaa ʕudʔwan bimaʒmaʕi lxaalidajn fi lqaahirah	1 & 2
S8:	بُورِسْ يِلْسِن، يَتَنَحَّى عَنِ السُّلْطَةِ لِأَسْبَابٍ صِحِّيَّةْ. buuris jilsin jatanaħħaa ʕanis sultʔati li?asbaabin sʔiħħijjah	1 & 2
S9:	جُورْج بُوش، يُقَدِّمُ وَسَاطَتَهُ لِحَلِّ الْأَزْمَةْ، بَيْنَ رُوسْيَا وَجُورْجِيَا. ʒuurʒi buuʃ juqaddimu wasaatʔatahu liħallil ?azmah bajna ruusjaa waʒuurʒijaa	1 & 2
S10:	مُحَمَّدْ رَجَبِ الْبَيُّومِي، رَئِيسُ تَحْرِيرِ مَجَلَّةِ الْأَزْهَرْ. muħammad raʒabil bajjuumii ra?iisu taħriiri maʒallatil ?azhar	1 & 2
S11:	السَّادَاتْ، بَطَلُ اَلْحَرْبِ وَالسَّلَامْ. ?assaadaat batʔalul ħarbi wassalaam	1 & 2
S12:	كَامْبِ دِيفِيدْ، اتِّفَاقِيَّةٌ مُلْزِمَةٌ بَيْنَ فِلَسَطِين وَإِسْرَائِيلْ. kaambi diifiid ?ittifaaqijjatun mulzimatun bajna filasatʔiin wa?israa?iil	1 & 2
S13:	الْعَاهِلُ الْمَغْرِبِيُّ، الْمَلِكُ اَلْحَسَنْ، فِي زِيَارَةٍ لِلْعَاهِلِ الْأَرْدُنِيِّ، الْمَلِكِ حُسَيْن بِن طَلَالْ. ?alʕaahilul maʁribijj ?almalikul ħasan fii zijaaratin lilʕaahilil ?ardunijj ?almaliki ħusajn bin tʔalaal	1
S14:	الْمَلِكْ فَهْدْ، يَتَوَعَّدُ تَنْظِيمَ الْقَاعِدَةِ فِي السُّعُودِيَّةْ. ?almalik fahd jatawaʕʕadu tanðʕiimal qaaʕidati fis suʕuudijjah	1
S15:	أَحْمَدْ هِيكَلْ، وَزِيرُ الثَّقَافَةِ الْأَسْبَقْ، يَحْصُلُ عَلَى جَائِزَةِ الدَّوْلَةِ اَلتَّقْدِيرِيَّةْ. ?aħmad hiikal waziiruθ θaqaafatil ?asbaq jaħsʔulu ʕalaa ʒaa?izatid dawlatit taqdiirijjah	1 & 2
S16:	بَشَّارِ الْأَسَدْ، وَلَخُّودْ، وَمُبَارَكْ فِي قِمَّةٍ ثُلَاثِيَّة. baʃʃaaril ?asad walaħħuud wamubaarak fii qimmatin θulaaθijjah	1 & 2
S17:	نعم naʕam	2
S18:	لا laa	2

Table 2. Word statistics in KSUEmotions corpus

	Phase 1	Phase2
unique words	125	68
Repeated words (unique)	8	4
Total Repeated words	17	5
"questioning" emotion sentences (in Phase 1 only)	16	0
Total	158	73

B. Selection of emotions

In the first phase of KSUEmotions, recording the following emotions were selected: ***neutral, sadness, happiness, surprise, and questioning***. Based on the human perceptual test analyses results, the questioning emotion was the best to be recognized by the listeners, while the happiness emotion was the worst to be recognized by the same listeners. The probable reason for the accuracy of perceptual regarding questioning is related to the presence of the first word /هل/, which is one of the question keywords in Arabic. We also believe that the meaning and content of sentences can play a role in the listener classification. Indeed, particularly in the case of happiness, the sentence text may lead to a less expression of happiness. For instance, we cannot express a sincere happiness when the text subject is death and especially when the speakers are not professional actors. In Phase 2, we have incorporated the "***anger***" emotion and removed the questioning emotion because its recognition affected by the extra question keyword /هل/. So the following emotions: ***neutral, sadness, happiness, surprise, and anger*** have been included in Phase 2 of the corpus subset.

C. Speaker selection

In Phase 1, twenty male and female speakers recorded 16 sentences in the five different emotions. The speakers included 10 males aged between 20 and 37, and 10 females aged between 19 and 30. All speakers were either undergraduate or graduate students, except for one female who was still attending secondary school.

In Phase 2, according to the results of the human perceptual test of the Phase 1, we selected the best seven male speakers among the ten male speakers and the best four female speakers who recorded the Phase 1. In the first phase, almost female speakers were from Syria and only two were from Saudi Arabia; in Phase 2, we added three female Yemeni speakers.

Each speaker was asked to fill the identification card that contains her name, nationality, age, place of birth, location where a part of his/her childhood was spent, current living place, highest level of education achieved, and marital status, etc. The three new female speakers' ages are between 20 and 25 years; two of them have undergraduate education level and the other has reached a secondary school level. Speakers identity cards Table (in DOC directory) shows the identification cards for each speakers who recorded in two Phases.

D. Recording

The KSUEmotions database was not recorded in a studio because our intention is to make it more real-life corpus. Each person in charge of making the recordings was given the required devices to make the recordings, and then they traveled to each of the speakers' homes so that the speakers could complete the recordings in their homes. The 23 speakers (10 males and 13 females) were from Saudi Arabia, Yemen, and Syria. High-quality microphones (SHURE 58 A) and three Dell laptops (model XPS 14Z) running Windows 7 were used to make the recordings. We used the PRAAT software [2] program to control the mono recording processing. 16 KHz was the sampling frequency used. All sentences were printed out for the speakers, and they were asked to read them many times before starting the recording.

E. Filename format

The filename format of ***DxxExxPgxxSxxTxx*** was used, in which each file starts with "Dxx," which indicates the corpus number. (In our lab series of corpora, this corpus is numbered 05, hence "D05.") The next three digits, "Exx," indicate the emotion code (E00, E01, etc.). This is followed by the code "Pgxx," which indicates the speaker gender (0 male, 1 female) and number (01, 02, etc.). The code "Sxx" represents the sentence number (S01, S02…S18), and, finally, "Txx" refers to

the trial number (T01, T02, etc.). For example, D05E03P104S01T01 indicates that sentence 1 was recorded by female speaker number 4, who simulated the sentence using the "Surprise" emotion. Table 3 shows the filename format details.

Table 3. Filename format details

Dxx	**Exx**		**Pgxx**		**Sxx**		**Txx**	
Corpus	**Emotions**		**Persons**		**Sentences**		**Trials**	
05	E00	Neutral	P0xx	Male	S01	Sentence#1	T01	Try # 1
	E01	Happiness	P1xx	Female	S02	Sentence#2	T02	Try # 2
	E02	Sadness	P001	Ali	S03	Sentence#3	T03	Try # 3
	E03	Surprise	P101	Aisha	S04	Sentence#4	T04	Try # 4
	E04	Questioning			S05	Sentence#5	T05	Try # 5
	E05	Anger						

F. Human Perceptual Test (Verification)

After the recordings were completed, we perform a perceptual test aiming at checking whether normal listeners could identify the recorded emotion types. For this test, we applied a set of rules for the listeners to follow. It is important to mention that no training session was provided before conducting the test in order to not influence the listeners. However, we allowed the listeners to ask the supervisor to stop at any time if they wanted to hear a recorded file again before deciding on the emotion type, but we did not allow them to go back and compare a recording with an earlier one spoken by the same speaker. Finally, the listeners could also take a break whenever they desired.

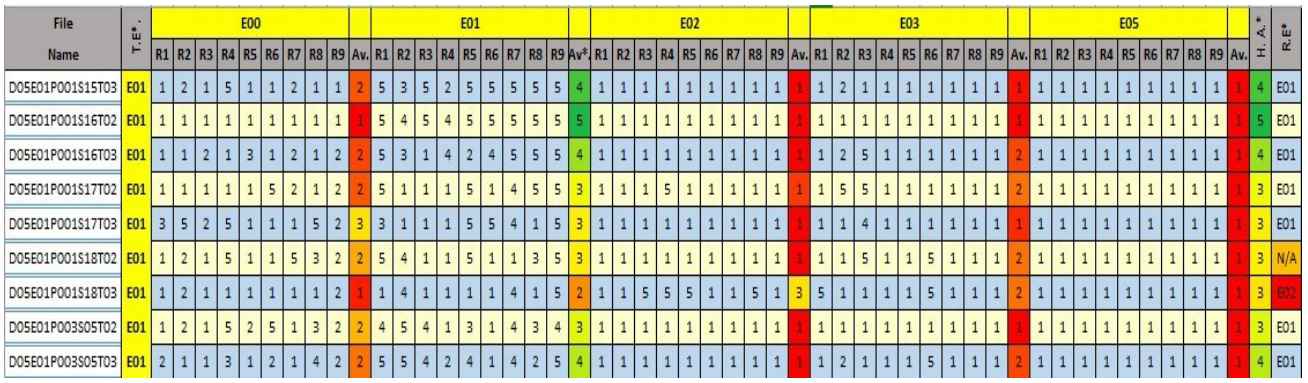
The test files were constructed as follows. First, all the filenames representing all of the recordings from all the speakers (3280 total files) were listed in Excel spreadsheets file and linked to their source audio files using hyperlinks. Then, the filenames were reordered randomly. The nine listeners were six males and three females, all Arabic native speakers except for one male who was fluent in both written and spoken Arabic. All of the listeners were undergraduates in their 20s, except for the non-native listener (Indian nationality), who was in his 40s. An example of the details for each listener's responses, including their ratings by percentage, are given in Table 4. The next step was to convert the collected data into Mean Opinion Score (MOS) [84], as shown in Table 5, by dividing all percentages by 20 and rounding off the result to obtain the final results as shown in Figure 1.

Table 4. An example of Listener Table

Listener No. 6							
File #	**Neutral (%)**	**Happiness (%)**	**Sadness (%)**	**Surprise (%)**	**Questioning (%)**	**Male/Female**	**Notes**
D05E00P105S08T01	80	0	20	0	0	F	
D05E01P006S07T01	30	70	0	0	0	M	noisy

Table 5. Mean Opinion Score (MOS) [3]

MOS	Quality	Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable



File Name	T.E*.	E00										E01										E02										E03										E05										H.A.*	R.E*
		R1	R2	R3	R4	R5	R6	R7	R8	R9	Av.	R1	R2	R3	R4	R5	R6	R7	R8	R9	Av*.	R1	R2	R3	R4	R5	R6	R7	R8	R9	Av.	R1	R2	R3	R4	R5	R6	R7	R8	R9	Av.	R1	R2	R3	R4	R5	R6	R7	R8	R9	Av.		
D05E01P001S15T03	E01	1	2	1	5	1	1	2	1	1	2	5	3	5	2	5	5	5	5	5	4	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	E01
D05E01P001S16T02	E01	1	1	1	1	1	1	1	1	1	1	5	4	5	4	5	5	5	5	5	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	5	E01
D05E01P001S16T03	E01	1	1	2	1	3	1	2	1	2	2	5	3	1	4	2	4	5	5	5	4	1	1	1	1	1	1	1	1	1	1	1	2	5	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	4	E01
D05E01P001S17T02	E01	1	1	1	1	1	5	2	1	2	2	5	1	1	1	5	1	4	5	5	3	1	1	1	5	1	1	1	1	1	1	1	5	5	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	3	E01
D05E01P001S17T03	E01	3	5	2	5	1	1	1	5	2	3	3	1	1	1	5	5	4	1	5	3	1	1	1	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	E01
D05E01P001S18T02	E01	1	2	1	5	1	1	5	3	2	2	5	4	1	1	5	1	1	3	5	3	1	1	1	1	1	1	1	1	1	1	1	1	5	1	1	5	1	1	1	2	1	1	1	1	1	1	1	1	3	N/A
D05E01P001S18T03	E01	1	2	1	1	1	1	1	1	2	1	1	4	1	1	1	1	4	1	5	2	1	1	5	5	5	1	1	5	1	3	5	1	1	1	5	1	1	1	1	2	1	1	1	1	1	1	1	1	3	E02
D05E01P003S05T02	E01	1	2	1	5	2	5	1	3	2	2	4	5	4	1	3	1	4	3	4	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	E01
D05E01P003S05T03	E01	2	1	1	3	1	2	1	4	2	2	5	5	4	2	4	1	4	2	5	4	1	1	1	1	1	1	1	1	1	1	1	2	1	1	5	1	1	1	1	2	1	1	1	1	1	1	1	1	4	E01

***T.E.:** Target Emotion; **Av**.: Average; **H.A.:** Highest Average; **R.E.:** Recognized Emotion

Figure 1. Human Perceptual Test Final Result

G. KSUEmotions Statistics

Tables, 6, 7 and 8 show the KSUEmotions corpus Phase 1 and Phase 2 speakers' gender, utterances, emotion type, and reviewer statistical.

Table 6. Number of Utterances According to Speakers' Gender

	Phase 1		Phase 2	
Gender	*Number of utterances*	*Percentage*	*Number of utterances*	*Percentage*
Male	800	50%	840	50%
Female	800	50%	840	50%
Total	1600	100%	1680	100%

Table 7. Number of Reviewers According to Gender in Phase 1 & Phase 2

Gender	Number of reviewers	Percentage
Male	6	67%
Female	3	33%
Total	9	100%

Table 8. Number of Utterances According to Emotion Type

	Number of Utterances	
Emotions type	*Phase 1*	*Phase 2*
Neutral	320	336
Happiness	320	336
Sadness	320	336
Surprise	320	336
Question	320	0*
Anger	0*	336
Total	1600	1680

* This emotion was excluded from the Phase

H. Verification Result

Table 9 shows the human perceptual test final results for the KSUEmotions Corpus for two phases. As shown in this Table Phase 2 is more accurate than Phase 1 and also male speakers' results are better than females in two phases.

Table 9. Human Perception Test Results

		Total No. of files	**Recognized files**	**%**
Male	**Phase 1**	800	646	80.75
	Phase 2	840	757	90.12
	All	1640	1403	**85.55**
Female	**Phase 1**	800	633	79.13
	Phase 2	840	727	86.55
	All	1640	1360	**82.93**
Phase 1		1600	1279	79.94
Phase 2		1680	1484	88.33
Phase 1+ Phase 2		3280	2763	84.24

I. Data Content

- **SPEECH**

The SPEECH directory contains two sub directories Phase1 and Phase 2. Each subdirectory contains recorded emotions in this Phase in five subdirectories: E00 (Neutral), E01 (Happiness), E02 (Sadness), E03 (Surprise), and E04 (Questioning) in Phase1 and E00, E01, E02, E03, and E05 (Anger) in Phase 2.

- **LABELS**

This directory contains timeless label file. All labels done by KACST symbols.

- **ALIGNMENT**

This directory contains files automatic time segmentation. All labels done by KACST symbols.

- **DOC**

This directory contains, speakers and reviewers' identity cards, KSUEmotions statistics, missing files, and KACST symbols and IPA table.

Note: Most of the above material was published in past team works [A, B, C, D]

J. References

[1] King Abdulaziz City for Science and Technology (KACST), "KTD Corpus," *Unpubl. Tech. Report.*

[2] D. Boersma, Paul & Weenink, "Praat: doing phonetics by computer." 2014.

[3] F. Ribeiro and D. Florêncio, "Crowdmos: An approach for crowdsourcing mean opinion score studies," Acoust. Speech Signal Process. (ICASSP), 2011 IEEE Int. Conf. on. IEEE, pp. 2416–2419, 2011.

K. KSUEmotions past works

A. Mefiah, Y. A. Alotaibi and S. A. Selouani, "Arabic speaker emotion classification using rhythm metrics and neural networks," 2015 23rd European Signal Processing Conference (EUSIPCO), Nice, 2015, pp. 1426-1430.

B. A. Meftah, S. A. Selouani and Y. A. Alotaibi, "Preliminary Arabic speech emotion classification," 2014 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Noida, 2014, pp. 179-182.

C. Ali Meftah, Yousef A Alotaibi, Sid-Ahmed Selouani, "Designing, Building, and Analyzing an Arabic Speech Emotional Corpus: Phase 2," The 5th International Conference on Arabic Language Processing (CITALA 2014),Oujda, Morocco,November 26, 2014, Pages 181-184.

D. Ali Meftah, Yousef Alotaibi, Sid-Ahmed Selouani," Designing, Building, and Analyzing an Arabic Speech Emotional Corpus," Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme, Reykjavik, Iceland, 2014, pp. 22-29.