EUROPÄISCHE UNION DER HÖRAKUSTIKER e. V.

2023 European Phoniatrics Hearing EUHA Award

Turkish Audio-Visual Speech Emotion Recognition Test (TR-AVSER): The dynamic, multimodal, face and vocal expression set

Authors: Çağıl Gökdoğan**, Ozan Gökdoğan**, Tugay Rıfat Duyar, Bulut Tanyeri, Filiz Çevik Tan, Menteşe (Muğla), Turkey

EUHA
Europäische Union der Hörakustiker e.V.

Published by: Europäische Union der Hörakustiker e. V.
Saarstraße 52, 55122 Mainz, Germany
Phone+49 (0)6131 28 30-0
Fax +49 (0)6131 28 30-30
E-mail: info@euha.org
Website: www.euha.org

Turkish Audio-Visual Speech Emotion Recognition Test (Tr-AVSER): Development of the Dynamic, Multimodal, Face and Vocal Expression Set

ABSTRACT

Introduction: Emotion recognition is a process which involves both facial and vocal expressions in order to maintain healthy interpersonal communication. Difficulties experienced in emotion recognition can be observed along with numerous health problems. Since tests for emotion recognition clinically allow early diagnosis and follow-up of some diseases, it is important in terms of clinical practice that dynamic test models which reflect the closest model to our daily life are created.

Purpose: The purpose of the study is to create a standard and reliable test battery which can be used in researches in different scientific areas.

Method: 17 students, who are receiving professional acting training, were asked to portray 6 basic emotions and two sentences determined for vocal expressions were read aloud by them to create the emotion recognition test battery. The visual and audio recordings done simultaneously were played to the healthy participants and the 12 emotion expressions which reflect 6 emotions in the best manner were selected. The acoustic analyses of these expressions were done and FO and intensity analyses were carried out in terms of prosodic features.

Findings: No difference was found between the emotion expressions of 6 actors whose recordings were used in the emotion recognition test in terms of gender. It was determined that the actor performance evaluations of the listeners were consistent.

Conclusion: It is considered that TR-AVSER is a valid and reliable test method for studies which are planned to be carried out on emotion recognition.

Key Words: Emotion recognition, communication, acoustic sound analysis

INTRODUCTION

Humans are social creatures and need interpersonal communication. In communication, the person's intent, motivation, emotional state and knowledge on the context of communication are important. In addition, the individual is expected to change his/her reactions according to decisions on the context and use his/her social cognition skill. (1-4)

It is possible to provide communication in different ways. The most effective one among these is the communication of emotions and it is at the center of social life. In order to be able to maintain healthy emotional communication, emotions should be transmitted accurately and the transmitted emotions should be perceived accurately. (5)

Communication of emotions is a non-verbal method of communication which allows the transmission of the desired emotions to be expressed to the listener through frequent changes in facial expressions, body posture and/or gestures and speech tone. Visual and auditory data are used for accurate communication of emotions. While clues such as facial expressions and body language are evaluated in visual input, vocal expressions or clues in speech give an idea about the emotion wished to be expressed in auditory data. In this respect, it is possible to classify communication of emotions as facial and vocal emotion communication. (5)

In studies on vocal communication of emotions, Brunswik's dual social perception lens model has been adopted. (6) The lens model defines vocal communication of emotions with two way variables: expression (coding – producing) and impression (decoding – recognition). The coder of the emotion is the speaker and the speaker reflects the desired emotion to his/her speech. The decoder is the listener and is responsible for interpreting the mood wished to be transmitted. In order to be able to interpret the emotion, the listener needs to make use of the speaker's prosody or acoustic characteristics of his/her voice. It is considered that the speaker's ability to express his/her intent vocally is related to the listener's ability to interpret the clues. (7)

Emotions have been defined and separated into two groups in the studies of Feidakis, Daradoumis and Cabella. According to these studies, there are 66 emotions and 10 of these constitute basic emotions (anger, expectation, distrust, fear, happiness, joy, love, sadness, surprise and trust), whereas the remaining 56 emotions constitute secondary emotions. (8) Since it has been determined that similar emotions display similar acoustic changes through the prosodic changes while expressing emotions, it is difficult to distinguish these from each other. Therefore, mostly Russell's emotional state model is used to evaluate emotions. (9,10)

It has been reported in different studies that prosodic changes in speech do not display cultural differences in 6 basic emotions (fear, happiness, anger, disgust, sadness, surprise) as

different from facial expressions and that they are universal. (11-13) However, it has been shown that recognizing emotions accurately only using speech tone in emotion recognition (66%) is more difficult compared to recognizing emotions from facial expressions (75-80%). Therefore, it we think of each information we use in emotion recognition as a direction, it becomes possible to state that reliable emotion recognition is multi-directional. (14)

Difficulties in recognizing emotions weaken communication skills and create interpersonal communication problems. In particular the difficulties experienced in the use of social cognition skill and the breaking down of the judgment skill as a result negatively affects individuals' moods (anxiety, depression, etc.) As a result of the weakening of the social cognition skill, negative effects are not only seen in the social and personal relationships of individuals but also in terms of the occupations of those who actively continue their work life. Problems in work life decrease motivation and work performance and develop economic problems. As a result, these individuals experience more social isolation. Therefore, the effects of emotional recognition on social behavior, quality of life and communication skills have started to be evaluated in a detailed manner. It should be remembered that among the groups which experience the most difficulty in recognizing emotions in social life is the group consisting of individuals with hearing loss. It is considered that the current auditory amplifications used by these individuals do not have sufficient an algorithm to recognize and distinguish emotions. Therefore, recent studies in this area have gained speed. (15-18)

The tendency in emotion recognition studies has been the use of emotional stimulants which define emotions only through facial expressions. However, emotional communication in daily life is multi-faceted. Studies have emphasized the importance of multi-emotional integration while processing emotional stimulants. The insufficiency of reliable and valid multi-faceted evaluation sets have led researchers to create their own tests. However, it is considered that the problems faced in the comparison of data in studies depending on the test method used decrease study reliability. Therefore, the need for test methods in which dynamic facial expressions and vocal expressions are used together that will make it possible for more reliable and valid studies to be done has risen. (19-21)

In order to be able to create a standard method which we think will help researchers in different scientific areas in using a valid and reliable evaluation method for emotion recognition, diagnosis of health problems which may develop in particular due to hearing loss in the area of audiology (dementia, Alzheimer's disease, psychiatric disorders, etc.), showing the capacity of devices which amplify hearing loss and determining strategies for new hearing device and cochlear implant programming, it was aimed at developing the Turkish Audio-

Visual Speech Emotion Recognition Test (Tr-AVSER): Dynamic, Multimodal, Face and Vocal Expression Set.

METHOD

Ethical committee approval was obtained from local University Clinical Research Ethical Committee with number 02-I on 07.02.2019 date.

Creation of Tr- AVSER Stimulants-Identification of Actors

In order to develop the speech emotion recognition test, 17 actors aged 20-25, who are senior students in our university's Fine Arts Faculty, Performing Arts Department were selected for the recordings of dynamic audio-visual stimulants (12 females and 5 males). It was given importance that the native language of these actors was Turkish and that they did not have any accent, intonations, etc. It was asked that they did not have any tattoos, '*piercings'*, beards, mustaches, etc. on their faces and define the identified text and emotions with facial expression-speech from a 1,5 meter distance without wearing glasses.

Stimulants

As for speech text, two sentences were used ("The children are talking by the door" and "The dogs are sitting by the door"). Sentence selection is mostly done using psycho linguistic data bases in the studies in foreign countries. However, since we do not have such data bases in Turkey, the sentences were formed as a result of the determined criteria by asking for the views of an expert in the linguistics area. The criteria consisted of not having the same number of syllables in the sentences and not having an emotional meaning.

Selection and Creation of Emotions

The six emotions accepted internationally and stated as standard in studies were used (happy, sad, angry, frightened, surprised and disgusted). It was considered that the use of six basic emotions which are accepted culturally as universal in different countries was suitable. (19) The actors were supervised by an expert using training techniques to induce the desired moods. The actors were told that they could use any technique they wished to reflect the determined moods and they were given enough time to be able to get into these moods.

Recording and Structuring of the Audio-Visuals

The actors' audio and visual recording were done separately in the Recording Studio of our university's Fine Arts Faculty, Performing Arts Department. The techniques stated in Livingstone and Russo's studies were used in the recordings and structuring of the recordings and permission was received from the study writers in question. (19)

Evaluation of the Recordings

For the validity and reliability studies of the recordings, 50 healthy adult participants aged 18-25 were played 204 vocal expressions at a sound level they were comfortable with in a quiet room, using supra-aural headphones (17 speakers x 2 sentences x 6 expressions). The participants were asked to mark which expressions the 6 emotions reflected on a preprepared questionnaire. After the scoring of the questionnaire, the actor recordings which reflected the emotions the best were identified and these recordings were played to 20 participants again (6 actors x 2 sentences x 6 expressions) and they were asked to score these expressions from bad to very good (using the Likert scale). For test-retest reliability, the recordings were played to 10 participants again a week later and they were asked to evaluate them. Then, the scores were recorded. Thus, the actor recordings which received the best scores from the desired 6 emotions were added to the final version of the test.

Sound analyses were performed using the sentence recordings of 6 actors (3 female, 3 male). The sounds were extracted from the video recordings using Adobe Audition 13.0.6.38 Version (recorded in wav format with Sampling frequency: 48000 Hz). The extracted sounds were separated and saved as "subject.name_first.word_emotion." F0 and intensity data were obtained over the screen images of the separated sounds with Praat version 6.1.03 and "a" and "e" vowels in the emphasized last syllable of the first words, by determining a point located approximately in the center of the time and intensity curve in which the vowels were produced. In the achieved Spectogram, the values were adjusted as View Range (dB) 40-90, Averaging Method= median, Spectrogram Window length (s): 0.05, Pitch Range (Hz) 75-500, Formant dot size (0.4). These adjustments were done with the purpose of making screen images more understandable. The other settings were left as Praat's factory settings.

Statistical Evaluation

NCSS (Number Cruncher Statistical System) 2007 (Kaysville, Utah, USA) software was used for statistical analyses. In the evaluation of the study data, besides descriptive statistics (Mean, Standard Deviation, Median, Frequency, Ratio, Minimum, Maximum), Mann Whitney U test was used in the comparison of the two groups which did not display normal distribution in quantitative data. Significance was evaluated in the levels $p<0,01$ and $p<0,05$. For test-retest reliability analysis, Kendall's coefficient of concordance was used.

Findings

For the audio-visual recognition set, 204 recordings made by 17 actors were evaluated to identify the actors who produced the best visual and audio portrayals. Firstly, 6 actors (3 females and 3 males) who received the best portrayal score (over 80%) were selected from the

17 actors. The 2 sentence recordings for 6 different emotions made with these actors were played once again to the healthy group consisting of 20 participants. Then, the scoring results were compared. No statistically significant difference was found between the scoring of the 1st and 2nd sentences of 6 actors who portrayed the emotions of happiness, fear, surprise and disgust. No statistical difference was found between the test-retest scores. In the emotion of happiness, the best score for the 1st and 2nd sentences was received by female actor No: 1. The best scores were as follows: male actor No: 2 for the emotion of fear; female actor No: 3 for the emotion of surprise and female actor No: 1 for the emotion of disgust. While no statistical difference was found between the scoring of the 1st and 2nd sentences of 5 actors, a statistical difference was found between the 1st and 2nd sentence portrayal of female actor No: 2 for the emotion of sadness and between the 1st and 2nd sentence portrayal of female actor No: 3 for the emotion of anger. The best score for the emotions of sadness and anger were received by male actor No: 2. The data of the actors who received the best scores in 6 different emotions and 2 sentences and had a high value of test-retest were added to the Tr-AVSER test data and the test was given its final version. Ten listeners were asked to evaluate the emotions in the 1st and 2nd sentences from 1 to 4. The differences in the views of the listeners were evaluated with Kendall's coefficient of concordance after they made their evaluations. Kendall's W coefficient of concordance test is a coefficient which determines the concordance level between scores given by many judges. (Sencan 2005) As a result of the Kendall W Test, the values between the judges were (Kendall's W= .089, ss= 5, p= 0.64) for the first sentence and (Kendall's W= .261, ss= 5, p= 0.13) for the second sentence. No statistically significant difference was found between these values. These results mean that there is no difference in the views of the listeners. 20 listeners were asked to score the emotions in the first and second sentences from 1 to 4. The differences in the views of the listeners were evaluated with Kendall's coefficient of concordance after they made their evaluations. As a result of the Kendall W Test, the values between the judges were (Kendall's W= .170, ss= 5, p= 0.50) for the first sentence and (Kendall's W= .215, ss= 5, p= 0.21) for the second sentence. No statistically significant difference was found between these values. These results mean that there is no difference in the views of the listeners. F0 and intensity parameters were also determined over the phonemes taken from the sentences vocalized by the actors (Tables 1and 2).

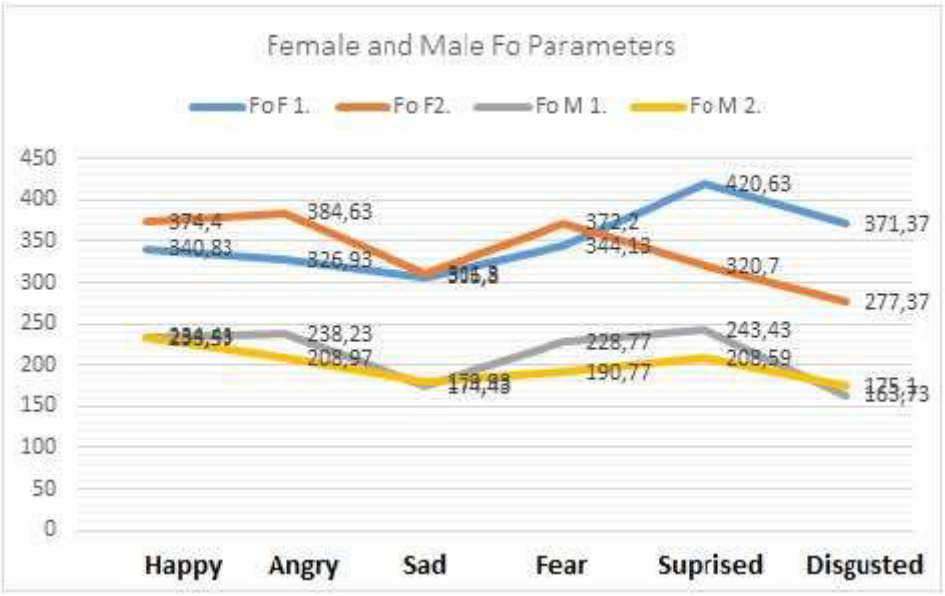
	1st Sentence		2nd Sentence	
	Basic frequency	Loudness	Basic frequency	Loudness

	(Fo)	(dB)	(Fo)	(dB)
Happiness	340,83	70,58	374,40	68,55
Anger	326,93	66,03	384,63	68,50
Sadness	306,80	61,46	311,30	63,47
Fear	344,13	60,95	372,20	63,47
Surprise	420,63	63,27	320,70	64,08
Disgust	371,37	61,88	277,37	60,36

Table1. 1st and 2nd Sentence Fo –Intensity Values in Female Actors

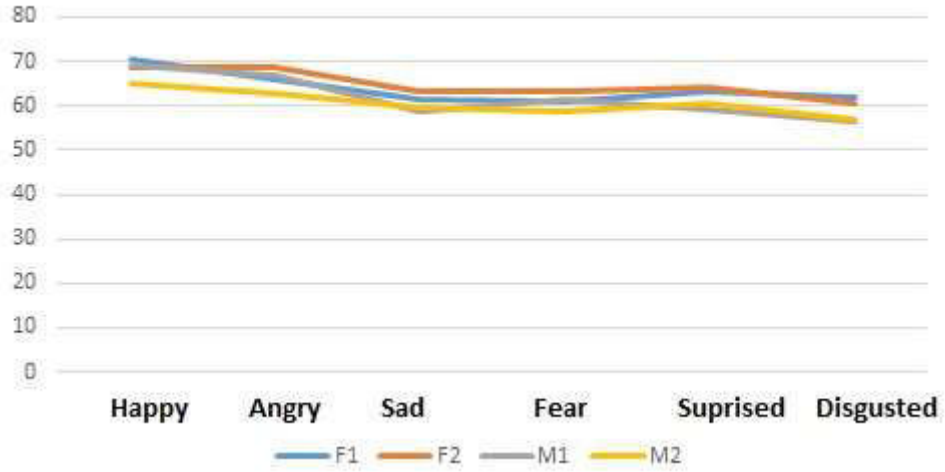
	1st Sentence		2nd Sentence	
	Basic frequency (Fo)	Loudness (dB)	Basic frequency (Fo)	Loudness (dB)
Happiness	233,53	69,35	234,43	65,01
Anger	238,23	66,93	208,97	62,98
Sadness	174,43	58,67	179,93	59,48
Fear	228,77	61,24	190,77	58,78
Surprise	243,43	59,15	208,59	60,77
Disgust	163,73	56,47	175,10	57,15

Table2. 1st and 2nd Sentence Fo –Intensity Values in Male Actors



Graphic 1. Average F0 Values of Female and Male Actors in Different Emotions

FoF1: The average F0 values of female actors in the 1st sentence; FoF2: The average F0 values of female actors in the 2nd sentence; FoM1: The average F0 values of male actors in the 1st sentence; FoM2: The average F0 values of male actors in the 2nd sentence



Graphic 2. Average Intensity Values of Female and Male Actors in Different Emotions

F1: The average intensity values of female actors in the 1st sentence; F2: The average intensity values of female actors in the 2nd sentence; M1: The average intensity values of male actors in the 1st sentence; M2: The average intensity values of male actors in the 2nd sentence

Discussion

Emotion recognition in establishing communication in daily life appears in a multifaceted manner. In recognition of emotions in communication, use of facial and vocal expression and perception of both gain importance. In this respect, separate and/or joint evaluation of both parameters in emotion recognition tests provides more realistic information to the researcher. In addition, it is considered that the use of dynamic expressions is an aspect that helps accurate recognition of emotional expressions and that multi emotion integration is achieved in this manner. It is known through studies that neuronal activation is different in dynamic expressions compared to static expressions. It has been shown that dynamic stimulants create a greater imitation response in the facial muscles of the observers compared to static expressions (Sato et al., 2007). Therefore, it is considered that the use of dynamic facial expressions and vocal expressions together will allow more reliable and valid studies to be done compared to static facial expressions. Therefore, it was preferred to use vocal expressions together with dynamic facial expressions in our study, which we think will reflect daily life in the best manner.

When content of tests on emotion recognition were analyzed, it was seen that mostly voice recordings and static facial expressions have been used. The tendency in emotion recognition researches has been to use emotional stimulants which depict emotions only with facial expressions. In this respect, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) developed by Livinstone and Russo is an example of dynamic stimulant models. Temporary solutions have been presented in different studies to create dynamic sets. However, since these approaches are different in terms of technical quality, it is difficult to compare the findings of these studies and this decreases study reliability.

In the developed tests, it has been stated that non-professional individuals were also preferred in the selection of professional actors and speakers in the portrayal of facial expression and vocal expression recordings. In our study, drama students contributed to facial and vocal expression recordings. These actors were preferred because their expression of emotions was close to being natural and they had no difficulty in expressing emotions in different circumstances.

In our study, both male and female actors portrayed facial and vocal expressions for the determined sentences and emotions. Although it was stated that female actors were more successful in portraying emotions, both genders' success in portraying different emotions was the same according to the scores given by our participants; this was taken into consideration and both genders' portrayals were given place in the test.

Emotion production/coding develops in line with the speaker's cultural background and learning skills. Based on this, moods in speaking emerge as a result of unconscious reactions.

As a result of physiological changes (for instance, changes in respiration rate due to sympathic nerve activation, etc.), there may be changes in the acoustic feature of speech voice. These changes which take place during speech constitute the suprasegmental characteristics of the speech. Suprasegmental characteristics point out to the prosodic structure of speech such as emphasis, intonation. Prosody appears as a characteristic where three important acoustic information is processed as resonance (F0, basic frequency), intensity (amplitude) and temporal information (rhythm, speech rate) that make it possible for us to perceive speech and identify emotional differences. Emotional prosody plays an important distinguishing role in the production and definition of emotion in speech. When an angry emotional expression is used in speech, it is observed that resonance increases to higher frequencies, whereas it falls to lower pitches in happiness. (6,9) However, while a decrease has been observed in F0 value in negative emotions such as sadness in some studies, an increase has been observed in the same value in positive emotions such as happiness. Similarly, it was observed in our study that F0 values decreased in particular in the vocalization of the emotion of sadness among the negative emotions, whereas F0 values increased in other emotions. There are no apparent differences between males and females in terms of emotions other than sadness. Therefore, it is considered that the decrease in the rate of accurately recognizing emotion only through vocal emotional expression might be due to the close results in F0 values. Although this change in F0 values is similar in the female and male actors' vocalization, it is known that the differences in F0 values is due to anatomical structure based on gender. There are studies which showed that intensity increased in emotions such as anger, resentment, joy in the vocal expression of emotions. However, it was determined in our study that although there was no apparent difference, there was a decrease in the intensity of emotions of sadness, fear and disgust compared to other emotions.

There are limited studies in Turkish on emotion recognition in terms of emotional prosody which we can do comparisons with and vocal expressions used for emotions are different as well. Meaningless sound and/or word production was used in some studies, whereas sound recording samples taken from films were used in others. In this respect, it is considered that aspects such as sentences used in our study will contribute to the literature since there are no studies which make use of sentences in Turkish.

In the recent years, studies on distinguishing the differences in emotion recognition have advanced as they received great interest from researchers in the areas of neuroscience, psychology, psychiatry, audiology and software creation. The most important part of these

studies is the use of verified and reliable emotional expressions. Stimulant sets on emotion recognition have increasingly become useable to meet this need.

Conclusion

Tests developed for emotion recognition are methods which will help in carrying out studies in different areas. It is considered that structuring tests by taking the changes in different languages due to prosody into consideration will make it possible to identify the difficulties in recognition emotions in vocal expression.