*Avrupa Bilim ve Teknoloji Dergisi*
*Sayı 14, S. 241-244, Aralık 2018*

***Araştırma Makalesi***

www.ejosat.com ISSN:2148-2683

*European Journal of Science and Technology*
*No. 14, pp. 241-244, December 2018*

***Research Article***

# Investigation of the Relation between Emotional State and Acoustic Parameters in the Context of Language

Turgut Özseven[1*]

[1] Tokat Gaziosmanpaşa University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, 60200 Tokat

(First received 26 July 2018 and in final form 27 November 2018)

(DOI: 10.31590/ejosat.448095)

## Abstract

Acoustic analysis is the most basic method used for speech emotion recognition. Speech records are digitized by signal processing methods, and various acoustic features of speech are obtained by acoustic analysis methods. The relationship between acoustic features and emotion has been investigated in many studies. However, studies have mostly focused on emotion recognition success or the effects of emotions on acoustic features. The effect of spoken language on speech emotion recognition has been investigated in a limited number. The purpose of this study is to investigate the variability of the relationship between acoustic features and emotions according to the spoken language. For this purpose, three emotions (anger, fear and neutral) of three different spoken languages (English, German and Italian) were used. In these data sets, the change in acoustic features according to spoken language was investigated statistically. According to the results obtained, the effect of anger on the acoustic features does not change according to the spoken language. For fear, change in spoken language shows a high similarity in Italian and German, but low similarity in English.

**Keywords:** Speech emotion recognition; Emotion and language; Acoustic analysis.

## 1. Introduction

The emotional state is in can cause many physiological changes, especially the voice. Different emotional states affect respiratory patterns and muscle tension, causing changes in voice structure. These changes are used in Speech Emotion Recognition (SER) studies. Subjective methods are performed by an expert listening to the speech recording. This method may vary according to the expert's experience. Acoustic analysis is an objective evaluation method.

Acoustic analysis is basically a digital signal processing process. The features used in SER studies are obtained by acoustic analysis. This features contain relevant voice descriptive attributes. Acoustic features mostly used in studies; fundamental frequency, formant frequency, jitter, shimmer, signal noise ratio and energy parameters.

One of the difficulties in SER studies is the determination of the features that are effective on individual emotions due to individual differences in the voice [1]. Another difficulty in SER is the acquisition of emotionally triggered data. For this reason, databases such as EMO-DB [2], EMOVO [3], SAVEE [4] and SUSAS [5], which are validated in the literature, are used or data gathered by researchers are used. For acoustic analysis, ready-made tools such as Praat [6] and OpenSMILE [7] or codes developed by researchers are used.

When the status of current studies is examined; according to the results of working with ZCR, pitch and MFCC parameters and emotional speech recognition on smartphones, time-frequency parameters can only be used for angry and happy [8]. A new acoustic feature set for emotion recognition was used in the perceptual analysis. Emotion recognition rate was improved between 7-11% with new feature set [9]. In a study involving heuristic evaluation of emotions using pitch, energy, speech rate, and spectral properties, SER performance was achieved at 83.5% [10]. A set of energy, pitch, formant and MFCC features were used in the study of emotion recognition in the Indian language [11]. As can be seen from the literature studies given above, existing studies are mostly focused on increasing the SER ratio. Although success rates achieved in SER studies have been at acceptable levels, these achievements have been obtained in spoken language dependent. That is, when speech or data set change, success rate changes. For this reason, the examination of differences in voice production between spoken language and emotion will contribute to the literature and future studies.

The aim of this study is to investigate the differences of the emotions according to the spoken language. For this purpose, the emotional speech database of three different spoken languages was analyzed by acoustic analysis. The features obtained by acoustic analysis were analyzed by statistical analysis and the change of sensation according to spoken language was examined.

[1] Corresponding Author: Tokat Gaziosmanpaşa University, Department of Computer Engineering, turgut.ozseven@gop.edu.tr

In the second part of this study, material and method, in the third part, experimental results were given. The results obtained in the last section are interpreted.

## 2. Material and Method

In this study, Berlin Database of Emotional Speech (EMO-DB), Surrey Audio-Visual Expressed Emotion (SAVEE) Database and Italian Emotional Speech Database (EMOVO) were used for the German, English and Italian spoken languages respectively.

EMO-DB was obtained by expression of different emotions by actors. Voice records are 16 bit mono and have a sampling frequency of 16 kHz [2]. EMOVO is a database built from the voices of up to 6 actors who played 14 sentences simulating emotional states. The recordings were performed with a sampling frequency of 48 kHz, 16 bit stereo, wave format [3]. SAVEE Database recorded an audio-visual emotional database from four native English male speakers, one of them was postgraduate student and rest were researchers at the University of Surrey [4].

The anger, fear and neutral emotions were selected from the emotional speech data sets collected in English, German and Italian spoken languages. The distribution of the used data is given in Fig. 1.

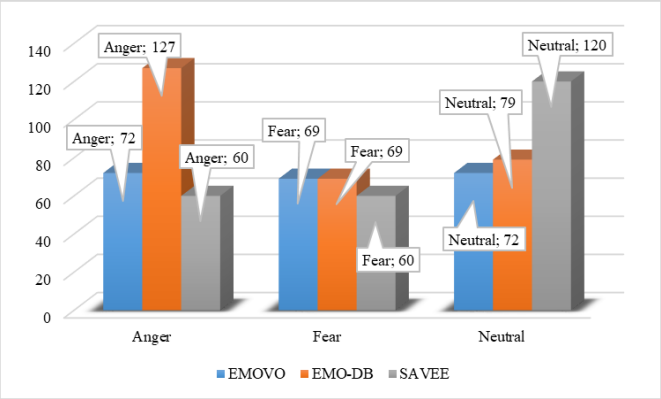


*Figure 1. Distribution of used data.*

Detailed information on the data sets obtained from the three sources is given in Table 1.

*Table 1. Detailed information of used data set.*

| Data Set | Emotion | Length | Sampling Rate | Number of Subjects | | Age |
|---|---|---|---|---|---|---|
| | | | | Male | Female | |
| **EMO-DB** | **Anger** | 1 sec – 5 sec | 16 kHz | 60 | 67 | 21 to 34 years |
| | **Fear** | 1 sec – 5 sec | | 36 | 33 | |
| | **Neutral** | 1 sec – 5 sec | | 39 | 40 | |
| **EMOVO** | **Anger** | 2 sec – 4 sec | 48 kHz | 36 | 36 | 23 to 30 years |
| | **Fear** | 2 sec – 5 sec | | 33 | 36 | |
| | **Neutral** | 2 sec – 5 sec | | 36 | 36 | |
| **SAVEE** | **Anger** | 2 sec – 6 sec | 44.1 kHz | 60 | 0 | 21 to 31 years |
| | **Fear** | 2 sec – 6 sec | | 60 | 0 | |
| | **Neutral** | 2 sec – 6 sec | | 120 | 0 | |

Speech recordings were analyzed by acoustic analysis and the fundamental frequency, formant frequencies, jitter, shimmer, intensity and zero-crossing rate acoustic features were obtained from each speech recording via Praat [6].

Fundamental frequency (F0) is the frequency of the laryngeal stimulation and depends on the sound folds and the lower laryngeal air pressure. This value is around 220-240 Hz in pre-adolescent girls and males, and between adult and male in average between 100-150 Hz and 150-250 Hz [12]. Formant (F1, F2, F3) is resonant in the sound path, and provides spectral information about the quantitative characteristics of the sound path. Jitter is the parameter that changes between periods. It includes irregularities that occur in F0 [13]. This features is defined as the variation of the fundamental frequency between successive oscillatory cycles [14]. The periodic variation between amplitude peaks is called shimmer. This features is defined as the amplitude changes of the laryngeal flow between consecutive oscillatory cycles [15]. Intensity is the energy of speaking. ZCR indicates the rate of signal changes in the signal. It is known as the number passing from zero of voice signal. Since a sinusoidal signal is a transition from two zeroes in each period, the signal frequency is calculated as the half of the number of transitions from zero.

In order to investigate the variability of the obtained acoustic features according to the spoken language, the neutral emotional state is taken as the basis. For each spoken language, the anger and fear change with the neutral were statistically examined. The data used does not have the normal distribution. Also, the variation of the mean values of each acoustic feature will be analyzed depending on the emotions. For this reason, the Mann-Whitney U Test was used for the analyzes. The Mann-Whitney U Test is used to measure the relationship between two independent variables. It compares the medians of those groups. It transforms continuous variables into sequential values in two groups. Thus, it evaluates whether the order between the two groups is different [16]. Statistical analyzes were performed separately for each database to determine the effective parameters to distinguish between neutral-anger and neutral-fear emotions. In the data set, emotion is an independent variable and acoustic properties are dependent variables.

## 3. Results

In this study, emotion-based variability of the acoustic features according to the neutral condition was investigated statistically while the effect of spoken language on emotions was investigated. Statistical analysis was performed by Mann-Whitney U test. Statistical analyzes were performed separately for each database to determine the features that were effective in distinguishing neutral-anger and neutral-fear emotions. The

statistical significance value (p) was 0.05 in the analyzes. After statistical analysis, parameters affecting anger and fear emotion according to neutral status in three databases are given in Table 2. The results given in Table 2 can be summarized as in Table 3.

*Table 2. The impact of acoustic features on recognizing fear and anger emotions according to statistical analysis results.*

| Acoustic Features | p (significance value) | | | | | |
|---|---|---|---|---|---|---|
| | EMO-DB | | EMOVO | | SAVEE | |
| | Fear-Neutral | Anger-Neutral | Fear-Neutral | Anger-Neutral | Fear-Neutral | Anger-Neutral |
| **F0** | .000* | .000* | .009* | .000* | .000* | .000* |
| **F1** | .000* | .000* | .022* | .009* | .444 | .000* |
| **F2** | .000* | .003* | .005* | .767 | .011* | .000* |
| **F3** | .009* | .086 | .000* | .500 | .416 | .000* |
| **Jitter** | .000* | .000* | .879 | .019* | .000* | .001* |
| **Shimmer** | .151 | .647 | .216 | .072 | .100 | .687 |
| **Intensity** | .118 | .000* | .830 | .000* | .000* | .000* |
| **ZCR** | .000* | .000* | .785 | .022* | .000* | .000* |

*p≤0.05

*Table 3. Features affecting emotion and speech language*

| Emotion | Database | F0 | F1 | F2 | F3 | Jitter | Shimmer | Intensity | ZCR |
|---|---|---|---|---|---|---|---|---|---|
| **Anger** | **EMOVO** | √ | √ | × | × | √ | × | √ | √ |
| | **EMODB** | √ | √ | √ | × | √ | × | √ | √ |
| | **SAVEE** | √ | √ | √ | √ | √ | × | √ | √ |
| **Fear** | **EMOVO** | √ | √ | √ | √ | × | × | × | × |
| | **EMODB** | √ | √ | √ | √ | √ | × | × | √ |
| | **SAVEE** | √ | × | √ | × | √ | × | √ | √ |

When the results given in Table 2 and Table 3 are examined;

- The F0 is able to distinguish between anger and fear in all databases, regardless of spoken language.
- The F1 feature is independent of spoken language, distinguishing anger emotion. However, it differs in English spoken language to distinguish fear from emotion.
- The F2 feature is independent of the language of speech in the emotion of fear. However, it differs in Italian spoken language to differentiate emotion of anger.
- The F3 feature is not effective in German and Italian spoken language in the emotion of anger. It is only effective in English. However, fear is not effective only in English spoken language to distinguish emotion.
- The jitter feature can be used independently of the spoken language for the emotional state of anger. However, it is ineffective for Italian in the emotion of fear.
- The Shimmer feature cannot be used to detect these two senses because it has no emotion and no effect on any database.
- The Intensity feature can be used independently of the spoken language for anger. However, except for English, it cannot be used for fear.

The ZCR feature can be used independently of the speech language for the anger. However, except for Italian, it cannot be used for fear.

# 4. Conclusions

In this study, the relationship between acoustic features and emotion recognition was investigated in the context of spoken language. According to the results obtained, the features affected by the anger are largely similar for the three languages. The similarity in Italian and German languages for Fear emotion is higher than that in the English language. The use of acoustic features to detect emotional states in these findings can be used independently of the scene. Furthermore, in the database to be used in the speech emotion recognition studies, it can be reached that the participant's spoken languages do not have any effect on the results. The important thing is to trigger the relevant sensation correctly.

# References

[1] B. Zupan, D. Neumann, D. R. Babbage, and B. Willer, 'The importance of vocal affect to bimodal processing of emotion: implications for individuals with traumatic brain injury', J. Commun. Disord., vol. 42, no. 1, pp. 1–17, 2009.

[2] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, 'A database of German emotional speech.', in Interspeech, 2005, vol. 5, pp. 1517–1520.

[3] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, 'EMOVO Corpus: an Italian Emotional Speech Database.', in LREC, 2014, pp. 3501–3504.
[4] P. Jackson, S. Haq, and J. D. Edge, 'Audio-Visual Feature Selection and Reduction for Emotion Classification', in In Proc. Int'l Conf. on Auditory-Visual Speech Processing, 2008, pp. 185–190.
[5] J. H. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom, 'Getting started with SUSAS: a speech under simulated and actual stress database.', in Eurospeech, 1997, vol. 97, pp. 1743–46.
[6] P. Boersma and D. Weenink, Praat: doing phonetics by computer [Computer program], Version 5.1. 44. 2010.
[7] F. Eyben, M. Wöllmer, and B. Schuller, 'Opensmile: the munich versatile and fast open-source audio feature extractor', in Proceedings of the international conference on Multimedia, 2010, pp. 1459–1462.
[8] W. Tarng, Y.-Y. Chen, C.-L. Li, K.-R. Hsie, and M. Chen, 'Applications of support vector machines on smart phone systems for emotional speech recognition', World Acad. Sci. Eng. Technol., vol. 72, pp. 106–113, 2010.
[9] M. C. Sezgin, B. Gunsel, and G. K. Kurt, 'Perceptual audio features for emotion detection', EURASIP J. Audio Speech Music Process., vol. 2012, no. 1, pp. 1–21, 2012.
[10] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, 'Primitives-based evaluation and estimation of emotions in speech', Speech Commun., vol. 49, no. 10–11, pp. 787–800, Oct. 2007.
[11] D. D. Joshi and M. B. Zalte, Recognition of Emotion from Marathi Speech Using MFCC and DWT Algorithms. IJACECT, 2013.
[12] S. Sarıca, 'Ses Analizinde Kullanılan Akustik Parametreler', Tıpta Uzmanlık Tezi, Kahramanmaraş Sütçü İmam Üniversitesi Tıp Fakültesi, Kahramanmaraş, 2012.
[13] K. M. Okur E., 'CSL ve Dr. Speech ile ölçülen temel frekans ve pertürbasyon değerlerinin karşılaştırılması', KBB Ihtis. Derg., vol. 8, pp. 152–157, 2001.
[14] M. Farrus and J. Hernando, 'Using jitter and shimmer in speaker verification', Signal Process. IET, vol. 3, no. 4, pp. 247–257, 2009.
[15] O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, and J. Singh, 'Use of temporal information: Detection of periodicity, aperiodicity, and pitch in speech', Speech Audio Process. IEEE Trans. On, vol. 13, no. 5, pp. 776–786, 2005.
[16] Y. Karagöz, 'Nonparametrik tekniklerin güç ve etkinlikleri', Elektron. Sos. Bilim. Derg., vol. 9, no. 33, pp. 18–40, 2010.