The Investigation the Effects of the Performance of an Independent Emotion Recognition of Model Used in the Dimensioning of Emotions

Turgut Özseven
Turhal Vocational School, Gaziosmanpaşa University, Tokat, Turkey

Muharrem Düğenci
Faculty of Engineering, Department of Industrial Engineering, Karabük University, Karabük, Turkey

Abstract

Emotion recognition aims at determining the state of emotion which is included in the speech or mimics of a person. Emotion recognition from speech is an area related to signal processing and psychology. Acoustic parameters obtained from speech signals through acoustic analysis, which is one of objective evaluation methods, is intensively used in emotion recognition studies. In this paper success in emotion recognition is examined in categorical aspects and the impact of dimensional model on independent emotion recognition success is investigated. Acoustic parameters were subjected to classification with Support Vector Machine in order to determine emotion recognition success. According to the obtained findings, emotion recognition success in categorical structure, dimensional structure and categorical-dimensional structure were 69.5%, 73.3% and 87.1%, respectively. Even if dimensional structure is higher in arousal than in valence, when emotion recognition success is examined on each emotion dimension, valence provided higher success.

Keywords: Dimensional Emotion, Emotion Recognition, Acoustic Analysis, SVM

Introduction

The emotional state of persons causes changes in the physiological structure of persons, which naturally affects voice. Objective and subjective methods are used for emotion recognition from speech. Perceptual recognition, which is a subjective method, is performed by listening by an expert and can show difference depending on the experience and probation of the expert. In order to overcome this problem of subjective evaluation, objective evaluation methods are employed. Acoustic analysis is used with the purpose of objective evaluation of voice disorders by obtaining several acoustic parameters from voice.

Features of voice signal and voice path change depending on the person, age, sex, sound path length, height, weight and emotional state. Therefore the emotional state of a person can be identified depending on feature vectors that will be obtained with sound analysis methods. Speech Emotion Recognition is not a new area of study; it was first developed in mid 1980s using statistical features of some acoustic parameters. In following years, with the development of computer architecture, more complicated emotion recognition algorithms began to be used (Ververidis & Kotropoulos, 2006). With acoustic parameters, language and discourse information were also used for emotion recognition (Chul Min Lee & Narayanan, 2005).

The most important limitation of Speech Emotion Recognition is about obtaining data whose validity is accepted and which can reflect emotions fully. With this purpose, there are ready datasets in the literature (EMO-DB, EMOVA, SAVEE, eINTERFACE, SUSAS) used in several studies whose validity is recognized; in addition, some researchers aimed at gathering their own datasets.
The basic purpose of the process in most emotion recognition studies is classification, and in most of these studies classification of emotional states is performed. Traditional classification techniques have been used in almost all of the proposed emotion recognition systems. Recent studies are focusing on the impact of hybrid classifiers and classifiers on acoustic parameters. The most widely used classifiers in the literature are Support Vector Machine (SVM), Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM).

There are three basic approaches in modeling of emotions in psychology studies (Sezgin, Gunsel, & Kurt, 2012): categorical, dimensional, and appraisal-based approach. Since the appraisal-based approach is not prevalently used because of its complex and sophisticated measurements of change (Gunes, Schuller, Pantic, & Cowie, 2011).

Categorical approach is independent expression of basic emotions. Ekman (Ekman, 1992) defined emotions in 6 basic categories: happiness, sadness, anger, fear, surprise, and disgust. However, categorical approach can fail to reflect the real situation when people are not in a single emotion but in complicated emotions (Gunes et al., 2011).

Due to the deficiencies of categorical approach in reflecting complicated emotions, dimensional approach is employed. Dimensional approach states that emotional state of a person is not depending on a single category but on multiple categories. There are several emotional dimensioning models presented for this purpose (Sezgin et al., 2012). A three-dimensional emotion space is proposed: arousal (activation), potency (power), and valence (pleasure) evaluation (Schlosberg, 1954). Another alternative is simpler two-dimensional emotion space: arousal and valence. Yet, the most widely used dimensional model is based on the assumption of Russell (Russel, 1980) that each basic emotion is represented by a bipolar entity being a part of the same emotional dimension in two-dimensional emotion space.

![Two-dimensional structure of affect](image_url)

Figure 1. The two-dimensional structure of affect, showing the dimensions of Positive Activation and Negative Activation on the diagonals and their relationship to the alternative dimensions of Valence (horizontal) and Arousal (vertical), as well as the 16 affect terms used to measure current mood (Kallinen, 2006). Adapted from Larsen and Diener (Larsen & Diener, 1992) and Watson et al. (Watson, Wiese, Vaidya, & Tellegen, 1999).

In our study EMO-DB dataset is used; therefore, the studies which consist of categorical and dimensional emotion recognition using EMO-DB in the literature were examined. A summary of the literature search is given in Table 1.
<table>
<thead>
<tr>
<th>Emotions/Dimension</th>
<th>Class.</th>
<th>Par.Count</th>
<th>Acc.</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arousal-Negative (boredom, disgust, neutral, sadness)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arousal-Positive (anger, fear, happiness)</td>
<td></td>
<td></td>
<td>96.8%</td>
<td></td>
</tr>
<tr>
<td>Valence-Negative (anger, boredom, disgust, fear, sadness)</td>
<td>SVM</td>
<td>6552</td>
<td>87.0%</td>
<td>(Schuller, Vlasenko, Eyben, Rigoll, &amp; Wendemuth, 2009)</td>
</tr>
<tr>
<td>Valence-Positive (happiness, neutral)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Categorical</td>
<td></td>
<td></td>
<td>84.0%</td>
<td></td>
</tr>
<tr>
<td>High arousal (anger, happiness, anxiety/fear)</td>
<td>ANN</td>
<td>133</td>
<td>83.7%</td>
<td>(Iliou &amp; Anagnostopoulos, 2009)</td>
</tr>
<tr>
<td>Low arousal (sadness, boredom, disgust, neutral)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1…Group 8 (Watson and Tellegen’s model)</td>
<td>HMM</td>
<td>39</td>
<td>81.2%</td>
<td>(Yun &amp; Yoo, 2009)</td>
</tr>
<tr>
<td>High Arousal</td>
<td>ANN</td>
<td>24</td>
<td>47.0%</td>
<td>(Anagnostopoulos &amp; Vovoli, 2009)</td>
</tr>
<tr>
<td>Low Arousal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anger, Disgust, Fear, Happy, Neutral, Sadness</td>
<td>SVM</td>
<td>121</td>
<td>78.3%</td>
<td>(Luengo, Navas, &amp; Hernaez, 2010)</td>
</tr>
<tr>
<td>Anger, Fear, Happy, Neutral, Sadness, Boredom</td>
<td>SVM</td>
<td>~80</td>
<td>79.7%</td>
<td>(Giannoulis &amp; Potamianos, 2012)</td>
</tr>
<tr>
<td>Arousal, Valence</td>
<td>SVM</td>
<td></td>
<td>86.3%</td>
<td>(Sezgin et al., 2012)</td>
</tr>
<tr>
<td>Anger, Disgust, Fear, Happy, Neutral, Sadness, Boredom</td>
<td>SVM</td>
<td>4968</td>
<td>93.8%</td>
<td>(Ivanov &amp; Riccardi, 2012)</td>
</tr>
<tr>
<td>Anger, Disgust, Fear, Happy, Neutral, Sadness, Boredom</td>
<td>SVM</td>
<td>39</td>
<td>67.0%</td>
<td>(Vasuki &amp; Aravindan, 2012)</td>
</tr>
<tr>
<td>Anger, Disgust, Fear, Happy, Neutral, Sadness, Boredom</td>
<td>SVM</td>
<td>95</td>
<td>79.5%</td>
<td>(Hassan &amp; Damper, 2012)</td>
</tr>
<tr>
<td>Arousal, Valence</td>
<td>SVM</td>
<td>~70</td>
<td>90.5%</td>
<td>(Ramakrishnan &amp; El Emary, 2013)</td>
</tr>
<tr>
<td>Anger, Disgust, Fear, Happy, Neutral, Sadness, Boredom</td>
<td>SVM</td>
<td>32</td>
<td>80.6%</td>
<td>(Garg, Kumar, &amp; Sinha, 2013)</td>
</tr>
<tr>
<td>Anger, Disgust, Fear, Happy, Neutral, Sadness, Boredom</td>
<td>SVM</td>
<td>6552</td>
<td>85.2%</td>
<td>(Chiou &amp; Chen, 2013)</td>
</tr>
<tr>
<td>Anger, Disgust, Fear, Happy, Neutral, Sadness, Boredom</td>
<td>SVM</td>
<td>90</td>
<td>80.4%</td>
<td>(Zhao, Zhang, &amp; Lei, 2014)</td>
</tr>
</tbody>
</table>
When the literature search given in table 1 is examined, it can be seen that the highest success obtained with SVM classifier in categorical classification was 93.8 percent. In studies which use the same data, the reason of obtaining different classification successes is the difference in the number and types of acoustic parameters obtained from data. Studies which examine emotional dimensions looked in arousal-valence dimension and high success was obtained with 96.8 percent. In studies found in the literature, emotion recognition is examined in categorical and dimensional aspects and no studies are found which evaluate each dimension categorically within itself.

The purpose of this study is to examine with SVM classifier the emotion recognition success in categorical and dimensional aspect of the seven emotions (anger, boredom, disgust, anxiety/fear, happiness, sadness, neutral) found in the Berlin Database of Emotional Speech (Emo-DB) database. In addition, the emotions in every emotion dimension are evaluated categorically.

### Materials and Method

In our study EMO-DB is used. Emo-DB includes voice recordings consisting of 7 different emotions (anger, boredom, disgust, anxiety/fear, happiness, sadness, neutral) spoken by actors within the project developed by department of communication science at Institute of Speech and Communication of Technical University of Berlin. The database was created by emotional expressions of 10 different sentences by 10 actors in 20-35 age interval. Voice records have 16 kHz sampling frequency and 16 bit mono features (Burkhardt, Paeschke, Rolfes, Sendilmeyer, & Weiss, 2005).

In the study 535 voice records are examined, the distribution of data-specific features of which are given in Table 2.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Male</th>
<th>Female</th>
<th>Age1 (30-35)</th>
<th>Age2 (20-29)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>60</td>
<td>67</td>
<td>90</td>
<td>37</td>
<td>127</td>
</tr>
<tr>
<td>Fear</td>
<td>34</td>
<td>33</td>
<td>50</td>
<td>17</td>
<td>67</td>
</tr>
<tr>
<td>Boredom</td>
<td>35</td>
<td>46</td>
<td>60</td>
<td>21</td>
<td>81</td>
</tr>
<tr>
<td>Disgust</td>
<td>11</td>
<td>35</td>
<td>31</td>
<td>15</td>
<td>46</td>
</tr>
<tr>
<td>Happiness</td>
<td>27</td>
<td>44</td>
<td>53</td>
<td>18</td>
<td>71</td>
</tr>
<tr>
<td>Sadness</td>
<td>25</td>
<td>37</td>
<td>47</td>
<td>15</td>
<td>62</td>
</tr>
<tr>
<td>Neutral</td>
<td>39</td>
<td>40</td>
<td>50</td>
<td>29</td>
<td>79</td>
</tr>
</tbody>
</table>

Table 2. The features belonging to the data used in the study
Acoustic analysis of speech records in both databases was performed with Praat (Boersma & Weenink, 2002) software and 15 acoustic parameters were obtained from each speech record. In the pre-processing of speech records, frame size 25ms, 50% overlap and Hamming windowing were used. Acoustic parameters obtained with acoustic analysis are given in Table 3.

Table 3. The acoustic parameters used in this paper

<table>
<thead>
<tr>
<th>Acoustic Parameters</th>
<th>Descriptive Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>Fundamental Frequency (F0)</td>
<td>✓</td>
</tr>
<tr>
<td>Formant Frequency (F1, F2, F3)</td>
<td>✓</td>
</tr>
<tr>
<td>Jitter (Local, Rap)/ Shimmer (Local, apq3)</td>
<td>✓</td>
</tr>
<tr>
<td>Unvoiced Frame</td>
<td>✓</td>
</tr>
<tr>
<td>Intensity</td>
<td>✓</td>
</tr>
</tbody>
</table>

Fundamental frequency (F0), reflects vibration of speed of vocal fold and determines the individual’s sound (Zupan, Neumann, Babbage, & Willer, 2009). Formant is resonant on the sound path. There are an infinite number of format theoretically, but in practice, only the first 3 or 4 contains important information. Formant are defined with formant numbers as F1, F2 ve F3 (Rezaei & Salehi, 2006). Jitter, indicates the change between periods and contains the resulting involuntary irregularities. The periodic variation between amplitude peak is called as shimmer. Unvoiced frame shows the ratio of pauses during speech. Intensity shows the energy created as a result of the amplitude change in sound signal (Zupan et al., 2009).

In the evaluation of emotion recognition success, Support Vector Machine (SVM) classifier and WEKA (Hall et al., 2009) package program were used. SVM is a method based on statistical learning theory. The basic purpose is based on the principle of defining the function which best distinguished classes, in other words, the hyper-plane. It is one of the most widely used classifiers in SER systems in the literature.

Two groupings were made in order to dimensionally evaluate emotions, which are;

- Arousal-Negative (boredom, disgust, sadness, neutral) ve Arousal-Positive (anger, fear, happiness)
- Valence-Negative (anger, boredom, disgust, fear, sadness) ve Valence-Positive (happiness, neutral)

Results

Emotion recognition successes are realized in three different patterns, namely categorical, dimensional and categorical for each dimension. The categorical emotion recognition successes are given in Table 4.

Table 4. Categorical emotion recognition
The highest success in categorical emotion recognition is sadness and the lowest success is obtained in the emotion of disgust. The reason for such low success in disgust emotion is that its number in the entire data is few. Classification successes obtained for arousal and valence dimensions is given in Table 5.

Table 5. Dimensional emotion recognition

<table>
<thead>
<tr>
<th></th>
<th>Arousal Positive</th>
<th>Arousal Negative</th>
<th>Arousal AVG</th>
<th>Valence Positive</th>
<th>Valence Negative</th>
<th>Valence AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>92.9%</td>
<td>90.3%</td>
<td>91.6%</td>
<td>12.7%</td>
<td>97.9%</td>
<td>74.0%</td>
</tr>
</tbody>
</table>

When the results given in table 5 are examined, it can be seen that the success of arousal dimensioning is higher compared to valence. The very low success in valence positive dimension is due to the fact that it consists of two emotions and the neutral emotion does not lead to essential change on the parameters. Table 6 gives the results obtained when emotions in each dimension of the dimensional structure are categorically classified.

Table 6. The categorical classification of emotions contained within the dimensional structure

<table>
<thead>
<tr>
<th>Arousal Positive</th>
<th>Arousal Negative</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Happiness</td>
<td>Anger</td>
<td>Fear</td>
</tr>
<tr>
<td>Accuracy</td>
<td>56.3%</td>
<td>81.9%</td>
<td>69.6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Valence Positive</th>
<th>Valence Negative</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Happiness</td>
<td>Neutral</td>
<td>Anger</td>
</tr>
<tr>
<td>Accuracy</td>
<td>95.8%</td>
<td>93.7%</td>
<td>92.1%</td>
</tr>
</tbody>
</table>

When the results given in Table 6 are examined, even if the emotion recognition success is higher in arousal dimensional structure, valence is more successful in categorical classification of the dimensional structure. Table 7 gives the emotion recognition success rates according to the categorical and dimension-based categorical result.

Table 7. Emotion recognition accuracy according to the results categorical based on dimension and categorical

<table>
<thead>
<tr>
<th>Emotions</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
According to Table 7, recognition success in some emotions is increasing in arousal-categorical structure whereas some others witnessed decline. In valence-categorical structure, only recognition success in the emotion of fear declines slightly whereas recognition success in other emotions increased considerably. When average success rates are examined, arousal dimensioning increased dimension-based categorical success slightly and valence dimensioning increased dimension-based categorical success considerably. These results show that usage of dimensional structure in emotion recognition provides considerable increase in the success of emotion recognition.

Conclusion

The emotional state of persons can cause changes in their physiological structures, which, therefore, is reflected on their voice. Sometimes the feelings of a person do not consist of a single emotion but a variety of emotions. In such cases effort is paid to perform dimensional analysis of the emotion and to increase the success rate of emotion recognition. Emotion recognition aims at determining the emotional state of a person by examining their speeches through signal processing methods. In this paper, categorical, dimensional and dimensional-categorical emotion recognition success rates in emotion recognition are examined. According to the obtained results, dimensional structure provided higher success compared to categorical structure. Dimensional-categorical structure increased the success rate obtained through each emotion considerably. In the light of these findings, higher success rates will be achieved by using valence dimensional structure followed by categorical structure for each dimension in emotion recognition.

References


