

Fear emotion classification in speech by acoustic and behavioral cues

Shin-ae Yoon¹ · Guiyoung Son¹ · Soonil Kwon¹

Received: 26 September 2017 / Revised: 8 June 2018 / Accepted: 25 June 2018 /

Published online: 7 July 2018

© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract Machine-based emotional speech classification has become a requirement for natural and familiar human-computer interactions. Because emotional speech recognition systems use a person’s voice to spontaneously detect their emotional status and take subsequent appropriate actions, they can be used widely for various reason in call centers and emotional based media services. Emotional speech recognition systems are primarily developed using emotional acoustic data. While there are several emotional acoustic databases available for emotion recognition systems in other countries, there is currently no real situational data related to the “fear emotion” available. Thus, in this study, we collected acoustic data recordings which represent real urgent and fearful situations from an emergency call center. To classify callers’ emotions more accurately, we also included the additional behavioral feature “interjection” which can be classified as a type of disfluency which arises due to cognitive dysfunction observed in spontaneous speech when a speaker gets hyperemotional. We used Support Vector Machines (SVM), with the interjections feature, as well as conventionally used acoustic features (i.e., F0 variability, voice intensity variability, and Mel-Frequency Cepstral Coefficients; MFCCs) to identify which emotional category acoustic data fell into. The results of our study revealed that the MFCC was the best acoustic feature for spontaneous fear speech classification. In addition, we demonstrated the validity of behavioral features as an important criteria for emotional classification improvement.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11042-018-6329-2>) contains supplementary material, which is available to authorized users.

✉ Soonil Kwon
skwon@sejong.edu

Shin-ae Yoon
sgy1017@sejong.ac.kr

Guiyoung Son
sikwon@sejong.ac.kr

¹ Department of Software, College of Software and Convergence Technology, Sejong University, 209, Neung-dong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea

Keywords Emotional speech classification · Emergency situation · Behavioral cue · Disfluency(interjection) · Speech signal processing

1 Introduction

The goal of this study is to develop an automated emotional speech recognition system for inclusion in the Korean emotional database, under the category of “real emergency situation.” Emotional speech recognition systems are an important topic in the Human-Computer Interaction research field and have reached a level of accuracy which has allowed widespread application in media services such as call centers and voice mail, etc. [14, 27, 39, 55]. Emotional speech recognition systems continue to be researched in order to develop more accurate systems. Because emotional speech recognition systems are developed based on human communication, this technology continues to require a deeper understanding of the emotion expression in human communication for further development.

In human communication, emotion is one of the most important factors for conveying intent in the connoted way. Thus, emotional information, defined as emotional intelligence, is an important part of natural human communication [52]. Studies have shown that during communication emotion can be expressed both deliberately and subconsciously through different channels including the utterance itself or paralinguistic features such as pitch, energy, or facial expressions, or a combination of both. With an increasing interest in emotional intelligence in order to provide better interaction between humans and computers, emotional intelligence is now required for accurate emotion perception [36]. For this reason, emotional speech recognition systems which aim to automatically identify a person’s emotional state in the Human-Computer Interaction(HCI) field have become a significant research topic [40]. Emotion speech recognition systems are especially beneficial for the management of service quality in media services such as call centers [43]. Automated telephone systems are already used in many areas such as consumer service center, emergency operations center, etc. because its applicability has substantial potential for dealing with users’ requirements by detecting their emotions through the spoken language [34]. For example, if the emotion speech recognition system detects a caller’s emotion to be too negative for the automated system to deal with, the call can be switched to a human service provider who is better equipped to provide adequate reactions [4, 33]. Moreover, if an emotional speech recognition system can respond to the user naturally, the system can make the user feel less daunted when they interact with computers. Existing emotion-based speech recognition systems in call centers and medical centers are currently used to monitor callers’ or patients’ emotions and take appropriate measures [47].

Emotional speech recognition systems are developed by extracting features from emotional acoustic data. To this end, there have been great deal of studies relating to emotional speech recognition systems using existing corpora [5, 59] (i.e., Berlin Database of Emotional Speech (BDES; from the German Corpus published by the Department of Acoustic Technology at Berlin Technology University), Danish Emotional Speech (DES; from the Danish Corpus published by Aalborg University, Denmark), and telephone messages [5, 59] which have been widely used in emotional speech recognition system development. However, there is no emotional database available in Korea. Thus, a prerequisite to the development of a Korean the emotional speech recognition system is the building of an emotional speech database in Korean, which requires the collection of emotional speech data. Recorded calls from emergency centers were considered as part of these efforts. Emergency call centers aim to offer

precise and suitable instructions to callers in emergency situations to minimize potential risks. Because calls are made by people in emergency situations requiring proper instruction, the caller typically exhibits panic or fear. Thus those emotions are naturally represented in the callers' voices. It is legislated that in Korean emergency centers, each interaction between caller and call recipient shall be recorded. Hence the acoustic data from emergency call centers can be considered an optimal database for negative emotional speech.

One reason that building an emotional database using acoustic data from emergency call centers is important is that it provides real-life acoustic data. Although there are considerable emotional corpora, much of these corpora use artificially recorded data which is recorded either by professional actors making use of Wizard of Oz, or non-professionals reading sentences under instructions to read with specific emotions. Artificially produced emotional data are very clear and discernable, Thus the emotional speech recognition systems using those corpora lack the finesse of those which use acoustic data produced in real-life contexts. This is because unlike corpora recorded by an actor, emotions are not always so obviously expressed, and can be represented in much more complicated ways in the real-world [60]. Hence, studies on emotional speech recognition systems using acoustic data from real-life situations are becoming increasingly important. There have been few studies using conversations which occur in real life, especially telephone conversations. Existing real-world acoustic data is mostly amassed at telecommunications service centers [43]. However, this data does not include urgent situations in which the caller's normal cognitive function may be effected by stress. Emotional speech recognition systems which use acoustic data from the real world, including urgent situations, need to be further studied [57]. Therefore, we use considerable real-life speech which was recorded under real emergency situations at an emergency call center in this study. Also, we anticipate that this data will be valuable for finding an emotional and behavioral feature to recognize emotional speech patterns in real-life situations with higher accuracy, specifically in Koreans.

To verify whether the acoustic data from the emergency call center was valid for the negative emotional database, our priority was to identify which category the callers' emotions fell into. To this end, we looked into emotion studies in the field of psychology, because emotions have been an important psychology research topic for a long time. Of the many emotion theories, two classic emotional distinction theories stand out: the discrete emotion theory and the dimensional theory. The discrete emotion theory devised initially by Tolkins (1962) and further developed by many researchers, argues that individual emotions (e.g., surprise, interest, joy, etc.) have biological and neurological inherent profiles [18, 41, 65] and basically, this theory identifies eight basic discrete emotions (surprise, interest, joy, rage, fear, disgust, shame, and anguish). On the other hand, the dimensional theory [46] suggests that emotions are distinguished by only two emotional dimensional spaces: 'valence' which is defined as judging how positive or negative an emotion is, and 'arousal' which is defined as the intensity of an emotion. According to this theory, it is by these emotional dimensional spaces that the basic emotions are defined [46, 49]. Therefore, emotions can be distinguished as either falling under the eight basic discrete emotions, or as three simple emotions: positive, neutral, or negative. In emotional speech recognition system studies, there has been much focus on the discrete emotion theory for emotion distinction [14, 28, 37, 43, 47, 59, 61, 64]. However, in real life, emotional utterances are far more complicated, and it is therefore important to consider these two theories together as a mutual supplementary view [1, 12, 16, 31]. Yik et al. (1999) demonstrated the eight basic discrete emotion clusters based on a dimensional theory called the circumplex model of affect. This theory was advanced in 2011 to

encompass a 12-point affect circumplex model (Fig. 1) which demonstrates discrete emotion-clusters based on the dimensional theory [31, 65]. According to this model, a person who calls the emergency services’ emotional state would fall into the negative (i.e., ‘displeasure’), high-activation (i.e., ‘activation’), and tense or jittery (boxed in orange). The current study is based on Yik et al.’s affect circumplex model for machine-based emotional speech recognition systems. We concluded that, under this model, callers to emergency call centers’ emotional states were in real-life situations corresponding to high arousal and negative valence, falling into the emotional state of the activated displeasure clusters (e.g., frenzied, jittery).

There are many acoustic features which are widely used in emotional speech recognition [48]. The relationship between acoustic features and their representation in the specific discrete-dimensional emotional accounts have been well reviewed (Table 1) [19, 21]. Considering both the circumplex model of affect (i.e., negative valence and high arousal emotional states) and vocal expression of emotions (Table 1), ‘the voice under an emergency’ is represented with significant variations in voice intensity as well as Formant frequency of speech (i.e., F0, pitch). Among the acoustic features, we analyzed variability in F0 (i.e., pitch), variability in voice intensity (i.e., energy), and Mel-Frequency Cepstral Coefficients (MFCC).

In this study, we adopted ‘disfluency’ as a prosodic parameter (e.g., pitch, energy, speaking rate), although this particular behavioral feature is seldom used by most emotional speech recognition systems [34, 43]. Even though disfluency is a key feature which refers to both an emotional state and cognitive dysfunction, there has been little research on the inclusion of this feature as a factor for speech emotion pattern recognition. Disfluency is defined as “the hesitation phenomena in utterance” [6] or “all kinds of speech features which contrast with

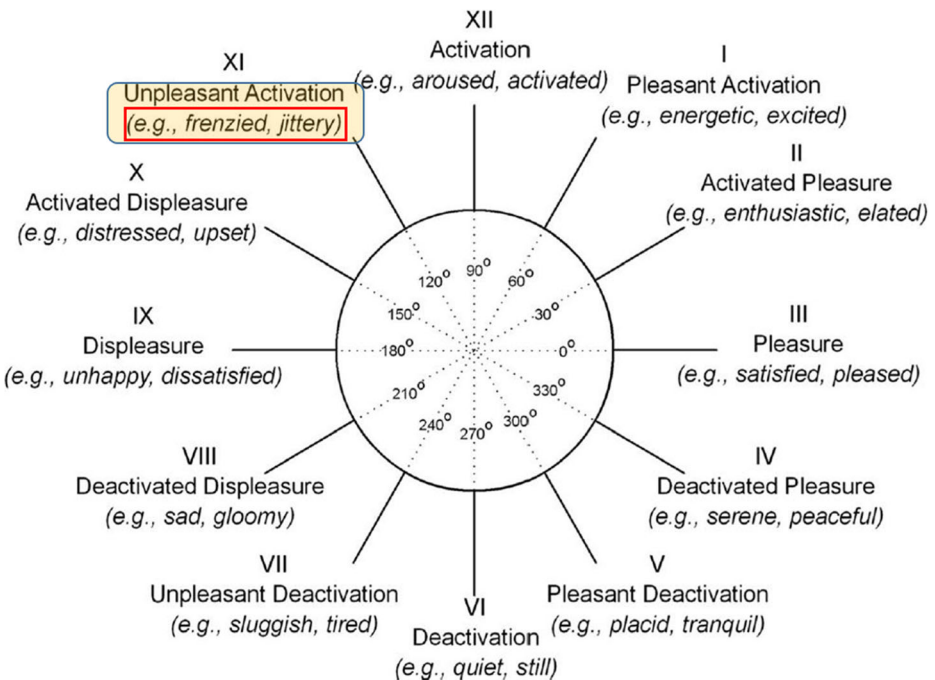


Fig. 1 Circumplex model of affect [65]. This model demonstrated the discrete emotions clustered based on the dimensional theory. According to this model, the caller’s emotional state falls into the negative (i.e., displeasure in this figure) and the high-activation (i.e., activation in this figure); Tense or Jittery (Boxed in orange)

Table 1 Acoustic features and their representation in the specific discrete-dimensional emotional accounts summarized and adapted from Juslin & Laukka (2003) and Lukka & Juslin (2005)

Vocal expression (categories)	Discrete emotion					Dimensional emotion	
	Anger	Fear	Happiness	Sadness	Tenderness	High-arousal	Positive-valence
Speech rate	Fast	Fast	Fast	Slow	Slow	Slow	Fast
Voice intensity	High	High	High	Low	Low	High	Low
Voice intensity variability	High	High	High	Low	Low	High	Low
High-frequency energy	High	High	High	Low	Low	High	Low
F0 (<i>M</i>)	High	High	High	Low	Low	High	Low
F0 variability	High	Low	High	Low	Low	High	Large
F0 contours	Up	Up	Up	Down	Down		
Voice onset	— ^a	—	Fast	—	Slow	Slow	Fast
Microstructural regularity	Irregular	Irregular	Regular	Irregular	Regular		
Proportion of pauses	Small	Small	Small	Large	Large	Small	
Precision of articulation	High	—	High	Low	Low	High	Low
Formant 1(<i>M</i>)	High	Low	High	Low	—	High	Low
Formant 1 (bandwidth)	Narrow	Wide	Narrow	Wide	—	Narrow	
Jitter	High	—	High	Low	—		
Glottal waveform	Steep	Rounded	Steep	Rounded	—		

^a “—” in table indicates that there was no specific or frequent finding for acoustic cues. Empty spaces indicate there was no vocal studies reviewed in from Juslin & Laukka (2003) and Laukka & Juslin (2005)

fluency” [63]. There are various forms of disfluency phenomena which are mainly categorized into either stuttered or non-stuttered forms. Stuttered disfluencies include sound, syllables, or one syllable/word repetition (e.g., “r-r-really”, “I-I-I”), and sound prolongations (e.g., “mmmmeans”) [15], where as non-stuttered disfluencies are interjections between utterances (e.g., “uh”, “umm”), word or phrase repetition (e.g., “turn, turn left” whereby the word *turn* is repeated), deletion (also known as false starts; which occur when a speaker starts saying something they mean to say but then starts a new sentence; e.g., “No it’s* do you see the sign over there”; whereby the speaker started with “no it’s” and then restarted with “do you -”), word or sequence substitution (e.g., “Do you have a caravan?”* whereby *caravan* is substituted for *summer cottage*), insertion (when a speaker restarts utterances with more than one word after the initial utterance; e.g., “I want to* I really want to” whereby the speaker restarts the utterance with *really* inserted), and articulation errors (e.g., “turn lift”* there whereby the word *left* displayed an articulation error to *lift*) [26].

These disfluencies appear when the speaker searches for the next word or phrase or to occur at the beginning of an utterance, due to either a greater demanding in cognitive load or low cognitive control [6]. In Lindström et al. (2008) in which participants were asked to speak while they were handling a steering wheel (i.e., vehicle spoken dialogue task), disfluency such as interjections (e.g., “uh”, “umm”), and filler words (e.g., “like”, “you know”) were observed more when the speaker was performing tasks that required higher cognitive load than tasks that required less cognitive load [26]. In addition, Lindsey et al. (1995) and Dibble et al. (2013) both argue that people also tend to show disfluency in their utterances when they convey information that may arouse negative emotions [10, 25]. In order to test the emotional impact on cognitive processing, Coffman and Davison (1997) asked subjects to imagine specific situations from audio-recordings and express their thoughts and emotional reactions [7]. They

found that subjects showed a higher proportion of interjection or filler words in emotionally provocative scenarios than in neutral ones. Thus, disfluency can be thought of as a reasonable and theoretically grounded paralinguistic criterion which determines if a voice is highly emotional or not. Although there are some interjections in normal situations, it can be concluded that the appearance and pattern of disfluency are distinctive and more frequent in emergency situations than in normal situations. Therefore, interjections in the non-stuttered disfluency category may be taken into consideration as the feature for emotional speech recognition systems because they are quantifiable features that can be analyzed.

Thus, we aimed to develop an automatic classification of negative emotional states in a speech from emergency calls using the classification method, the Support Vector Machine (SVM) with several acoustic features, and the behavioral feature ‘interjection.’ We believe this may initiate the collection and building of a Korean emotional speech database for Korean emotional recognition speech classification, in order to encourage and enable many emotional-based automated service fields.

This paper is divided as follows; in Section 2, we present related studies that investigated emotional speech recognition systems. In Section 3, we explain the research method composing of the survey procedure, defining speech emotional voice, data acquisition and annotation, and pattern classification. In Section 4, we describe the experiment procedure. In Section 5, we present the results and discussion. Our conclusion forms the final section.

2 Related work

2.1 Emotion recognition in real-life spontaneous speech

Previous emotional recognition classification studies used induced emotional corpora which were mostly recorded in unnatural surroundings at high quality [11]. The characteristics of databases used for emotional recognition in speech are summarized in Table 2.

Because there are many differences between acted corpora and natural acoustic data, there are some dissimilarities between them in terms of classification results. The induced emotional corpora recorded in at high-quality unnatural surroundings tends to have a high level of accuracy. Sato and Obuchi (2007) reported that recognition results for acted emotions are

Table 2 Database used in the speech emotional recognition system revised from M. El Ayadi et al. (2011)

Corpus	Language	Source	Emotions
LCD emotional prosody speech and transcripts	English	Professional actors	11 emotions
Berlin emotional database	German	Professional actors	7 emotions
Danish emotional database	Danish	Nonprofessional actors	5 emotions
Natural	Mandarin	Call centers	2 emotions
ESMBS	Mandarin	Nonprofessional actors	6 emotions
INTERFACE	English, Slovenian, Spanish, French	Actors	8 emotions
MPEG-4	English	U.S. American movies	7 emotions
Beihang University	Mandarin	Nonprofessional actors	5 emotions
FERMUS III	German, English	Automotive environment	6 emotions
KES	Korean	Nonprofessional actors	4 emotions
CLDC	Chinese	Nonprofessional actors	6 emotions

much better than those in voice portal data [53]. On the other hand, Morrison et al. (2007) concluded that emotionally induced corpora fail to model subtle real-world emotions. This supports the argument that studies on emotional recognition in speech should be conducted using databases acquired in a natural context [34]. With recent databases containing natural emotional states occurring in real-life (mainly call-center interactions), more sophisticated classification studies on emotional recognition systems have become possible [9]. Several initial studies on emotion recognition using real-life utterances have been conducted, focusing on the detection of negative emotion (i.e., anger). This is because of the significant importance of negative emotions in call center interactions [14, 37]. Neiberg and Elebius (2008) studied the automatic detection of negative emotion (i.e., anger) which occurred in real-life using the corpora of the Swedish Telephone Company. They demonstrated 83% accuracy using SVM. Galanis et al. (2013) also used a corpus consisting of 135 customer service call center real-life conversations from a Greek telecom company for emotion classification using SVM [37]. Devillers et al. (2005) further established several classification methods (SVM, decision trees) for the recognition of non-basic emotion and blended emotion (e.g., anger or fear/ anxiety or annoyance) using two sets of corpora, which were all real-life dialogues (i.e., from financial call centers and medical emergency centers) [9]. Considering these studies concerning real-life acoustic data, we concluded that using acoustic data from real-world situations leads to more accurate classification in emotion recognition systems, so we used real-world data consisting of both urgent and non-not-urgent situations.

2.2 Behavioral features in emotional voice recognition

While previous studies related to emotion recognition have been limited to acoustic information [9, 43], many behavioral features have been widely used for emotional recognition in speech. This may be due to the increasing database of real-life situations. Unlike induced or acted emotional corpora, emotional utterances in real-life have high a probability of accompanying behavioral features such as laughter, sighs, disfluency, etc. The more natural utterances there are in a database, the more consideration has to be taken into account of these behavioral features. A study by Lee and Narayana (2005) used three sources of information (i.e., acoustic, lexical, and discourse) for emotion recognition, focusing on negative and negative emotions. Results showed that using both acoustic and language information improved emotion classification [22]. Polzehl et al. (2011) elaborated emotion recognition modeling anger classification in recorded emotional corpora using both behavioral and acoustics features; the fusion of both features showed a slight improvement in overall scores (with a baseline of 60% accuracy to a maximum of 79% accuracy) [47]. However, the features used in those studies were phrase or word usage frequencies (e.g., Emotional Saliency, Bag-of-Words, Term Frequency), not paralinguistic features derived from behavioral features [22, 24, 47]. Devillers et al. (2004) suggest that disfluency, such as the use of fillers or pauses, can be used as a cue for emotion detection [8]. In successive research, Devillers et al. (2005) introduced this feature to detect the emotion anger/fear and showed the detection rate improving to 60%. Thus, Devillers insists that behavioral features such as lexical, prosodic, disfluency, and contextual cues should be taken into consideration for accurate emotion detection in a real-life speech [9]. Because of disfluency's late adoption as a feature for emotional recognition classification, related studies are rare. We hypothesized that the inclusion of the disfluency paralinguistic feature will improve the detection rate of voices under urgent situations. Thus, interjection, a phenomenon represented in disfluency, was identified and selected for classification in this study.

3 Method

3.1 Survey of emotional voices at an emergency call center

To identify which psychologically defined negative emotion (according to the 12-point affect circumplex model) in a caller's voice corresponding to the categorization of a voice under either urgent or emergency situations [65], we conducted a survey of emergency call recipients at the Seoul Emergency Operation Center. Under emergency situations, a caller's cognitive ability to answer the call recipient's questions and describe the situation is extremely important. However, when a caller becomes too emotional, their cognitive capacity often becomes limited, and they have trouble conveying the information required. This results in the call recipient struggling to take appropriate measures. Thus, the purpose of the survey was to identify and categorize which particular emotional states in which callers under urgent circumstances struggle to convey specific or detailed information. The main question in the survey was, "in which emotional situations have you had difficulties understanding the callers' utterances in order to obtain information, and how difficult was it?" In total, 23 different emotional states were identified (e.g., I found it difficult to understand the situation and to give the right instructions when the caller was angry). Situations were precisely depicted using words that implied callers' emotional states. These words were adjectives selected from both the Positive Affect and Negative Affect Schedule (PANAS) [62] and Pleasure-Arousal-Domination (PAD) [32]; the former was initially devised by Watson (1988; translated into Korean by Lee et al., 2003) to measure the positive and negative affective scale [23]; the latter is a systematic framework of emotions which was introduced by Mehrabian and Russell (1974; translated into Korean for this research). Respondents were asked to rate their difficulty from 1 to 5 on a Likert scale (i.e., 1 = not difficult and 5 = very difficult). Questions about specific work experience, the number of calls a day they received, and which gender/age group they struggled with most followed. There were a total of 68 respondents (male = 60) (Table 3).

3.2 Defining emotional voice

The inquiries used in the survey first needed to be justified. To test the survey's credibility, the internal consistency reliability (Cronbach alpha) was calculated. To identify which emotional states represented in voices could be grouped as 'fearful voices', a factor analysis was conducted (principal component analysis with OBLIMAX rotation) defining only two factors as valid (fear or non-fear). For more specific definitions and to narrow down the candidates, a regression analysis defining "fear" as the independent variable was conducted, allowing us to

Table 3 Detailed information of respondents in the survey

Category	Information				Total
Sex	Male		Female		68
	60		8		
Age group	30's	40's	50's	Unknown	68
	24	32	8	4	
Duration of Service	6 months	1 year	Over 1 year	Unknown	68
	7	10	49	2	

verify the emotional states that could correlate with the “fear” state. All statistics were calculated using SPSS (ver. 18) software.

Internal consistency reliability analysis showed that the questions used in the survey were highly reliable (Cronbach alpha coefficient=0.8), implying that the adjectives used in the survey correspond well to the emotional states. The survey results revealed that emergency call recipients had difficulty when communicating with people who were ‘excited’ most, then ‘temperamental’, followed by ‘angry’ in a regular sequence (i.e., a difficulty rating of 4.25, 4.16, and 4.01 respectively), followed by other emotional states (Fig. 2).

The factor analysis showed that among the 23 sentences describing the callers’ emotional states, 11 adjectives including ‘fear’ (i.e., terrified, distressed, sad, furious, disgust, misery, tense, scared, unpleasant, angry) were grouped together as the factor defined as ‘fear’ in this study, with 7.068 eigenvalue and 30.729% of explanations for variance; and the other eight adjectives were grouped under the factor defined as ‘non-fear’ in this study, with 3.196 eigenvalues and 13.894% of explanations for all variance. The Kaiser-Meyer-Olkin measure of sampling adequacy was 0.753, which means above average (0.6). Bartlett’s test of sphericity was also significant (χ^2 (253) = 779.555, $p < 0.0005$), (Table 4 and Fig. 3).

In succession, the regression analysis for the cross-validation test indicated that the words ‘terrified’ and ‘unpleasant’ explained statistically significantly ‘fear’ (F (1, 63) = 2.286, $R^2 = 0.414$, $p = 0.004$ for terrified, $p = 0.0001$ for excited). Therefore, we drew the conclusion that fear emotion were similar to ‘terrified’ or ‘excited’ emotions. Considering the statistical results and the psychological, emotional categorical system together, the ‘fear’ represented in callers’ voices fell into the category of ‘high arousal’ as well as ‘negative unpleasant emotional state.’ Accordingly, we were able to set sort of norm in phonological features expressed in a caller’s voice under the emergency situations based on the ‘fear emotion.

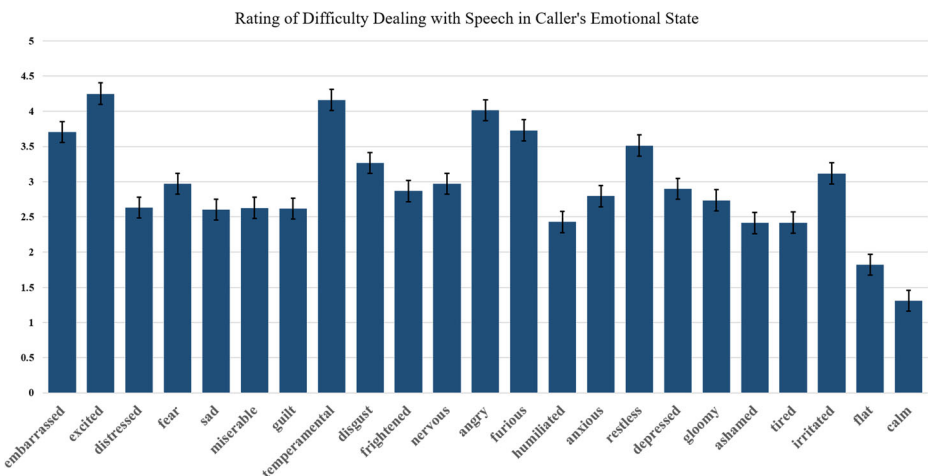


Fig. 2 Rating of survey results. X axis indicates callers’ emotional state (e.g., embarrassed, excited etc.) and y axis indicates degree of difficulty in obtaining information from the caller in order to take appropriate action or provide instructions to the caller to alleviate the situation. Mean and standard error are displayed

Table 4 Structural matrix from factor analysis result

Variables	Component	
	1	2
Fear	.734	–.259
Distressed	.712	–.320
Embarrassed	.679	.012
Sad	.649	–.265
Furious	.638	–.228
Disgust	.630	–.099
Miserable	.613	–.505
Nervous	.606	–.378
Frightened	.583	–.451
Excited	.557	.235
Angry	.540	–.163
Temperamental	.449	.415
Irritated	.432	–.160
Humiliated	.248	–.806
Flat	.146	–.770
Guilt	.233	–.749
Depressed	.294	–.716
Gloomy	.540	–.690
Tired	.229	–.668
Calm	.043	–.646
Ashamed	.523	–.550
Restless	.380	–.435
Anxious	.198	–.208
Eigen value	7.068	3.196
% of total variance	30.729	13.894
Total variance		44.622%

3.3 Acoustic data acquisition and emotion annotation

We obtained phone calls, defined here as ‘acoustic data’ from an emergency center at a fire station (i.e., a total of 10,319 items of acoustic data). The data was recorded at an 8000 Hz sampling rate on two channels, while the call recipient communicated with the caller during the emergency. Initially, 10,319 recordings were selected from among these with a specific focus on the emotional state. The total duration of the data is alternatively 187.7 h, with each single call’s duration ranging from 3 s to 20 min. Recorded conversations occurred between the call recipient and caller; each conversation was from different individuals while the operator was the same person for several calls. The distribution of recording duration and the number of recordings per call are shown in Fig. 4.

The data was annotated into the categories of fear and non-fear, by majority vote (3 out of 5 researchers). Researchers heard each recording and excluded those with too much background noise to identify the voice accurately, and annotated them as either fear or non-fear. Of the total data, 860 sets of data fell into the fear voice category and 9452 into the non-fear voice category during initial annotation; the remaining data was tagged as neither of the two categories due to mistaken calls. After initial tagging focusing on only the caller’s voice itself and not the background noise (e.g., other people’s voices), the emotional state obtained from factor analysis results in the survey were matched to the initially annotated acoustic data. As a result, of the 860 sets of data initially tagged as fear, only 162 of these were labelled as valid fearful voice.

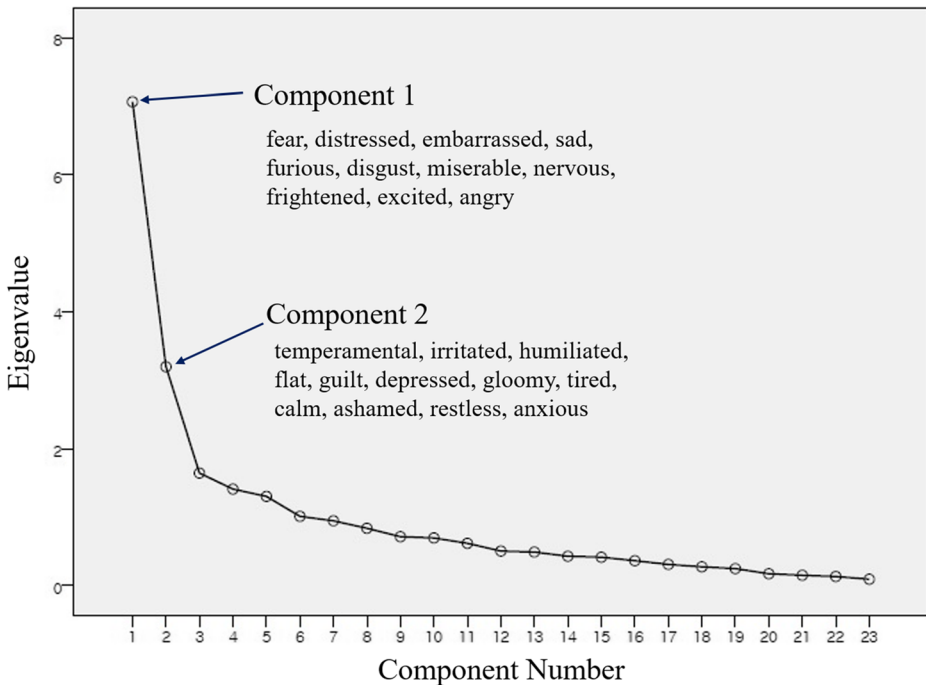
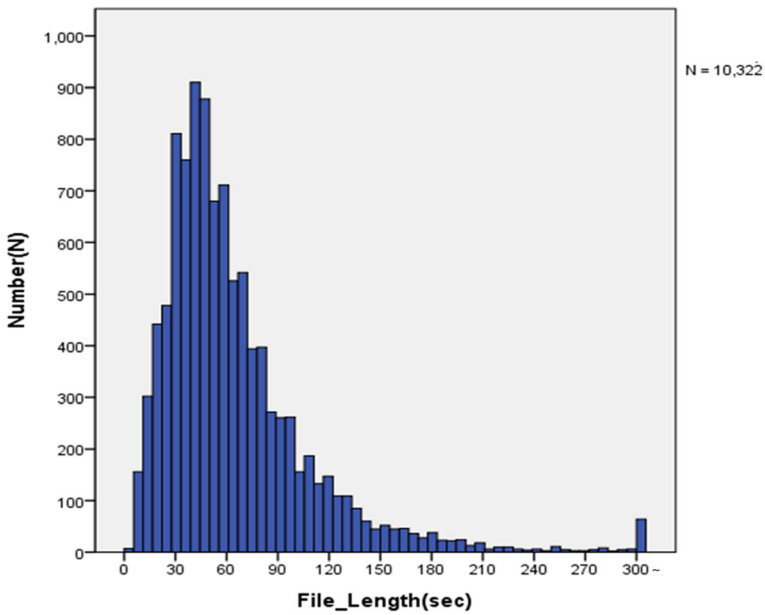


Fig. 3 Scree plot graph in eigenvalue against factor number. Values were grouped into two components according to pre-defined factor number (two); Component 1 (Fear) - fear, distressed, embarrassed, sad, furious, disgust, miserable, nervous, frightened, excited, and angry; and component 2 (Non-fear) - temperamental, irritated, humiliated, flat, guilty, depressed, gloomy, tired, calm, ashamed, restless, and anxious

3.4 Classification method

There are several well-known pattern classification methods; the Artificial Neural Network (ANN), the Gaussian Mixture Model (GMM), the Hidden Markov Model (HMM), the Maximum Likelihood Bayesian classifier (MLC), the Kernel Regression, K-nearest Neighbors (KNN), the Support Vector Machines (SVM), and many others [30, 54]. Because each classifier has its own benefits and limitations, there is no specific agreement on what the best classifier in emotion recognition is. Petrushin (1999) developed an emotional recognition system with an ensemble of neural networks [43]. Xiang et al. (2012) devised a speech emotion recognition system with SVM classification, suggesting that SVM is a simple and efficient computation of a machine learning algorithm with the limited training of data [40]. Tin Lay et al., (2003) proposed an HMM-based system for six archetypal emotion recognition systems. A study by Neiberg and Elenius (2008) indicated that the GMM method has a reasonable accuracy of 83%. Recently, to increase accuracy, there has been a tendency to combine two or more classifiers [34]. Of these various methods, SVM is known to be a simple and efficient computation of machine learning algorithms, and it is widely used for pattern recognition and problem classification. Especially in cases of limited training data, SVM provides very good classification performance compared to other classifiers [42], particularly for binary classification (i.e., fear or non-fear for this study). Thus, we selected SVM for our classification method.



Call_Length_all(sec)	675990
Length_Min/call(sec)	3
Length_Max/call(sec)	1156

Fig. 4 Distribution and detailed information of recording duration per call. Most frequent call duration was the 30–60-s range

Taking the total acoustic data as both training set (S_{train}) and testing set (S_{test}), S_{train} and S_{test} are represented as $i \in S$, where i is a tuple (l_i, f_i) , $l_i \in [63]$ is a class label and $f_i \in \mathbb{R}^p$. Optimizing the hyperplane separating between two categories (e.g., fear/non-fear) is as

$$w \cdot x - b = 0 \tag{1}$$

where w is the normal vector in hyperplane (see Eq. 1). To maximize the distance between two parallel hyperplanes which separate the data, the hyperplanes can be defined as (Eqs. 2 and 3)

$$w \cdot x - b = 1 \tag{2}$$

$$w \cdot x - b = -1 \tag{3}$$

with the distance as $\frac{1}{\|w\|}$. Thus, maximizing the distances between two hyperplane means minimizing $\|w\|$, satisfying the constraints (Eq. 4).

$$l_i(w \cdot x - b) \geq 1 \tag{4}$$

since there should be no data points between hyperplanes and $l_i \in \{1, -1\}$ [44].

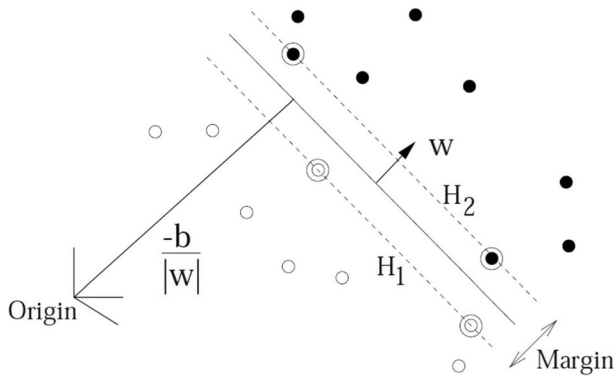


Fig. 5 Linear separating hyperplanes for the separable case [3]

An example view of the method is shown in Fig. 5, which is typical two dimensional case. The support vectors are circled.

Figure 6 is the classifier architecture for the real-time classifier. Its main components are described below. The classifier architecture describes the following process. It consists of the ensemble of linear SVM. First, load the model $T_{i,j}$ and the class labels are assigned. And then, we segmented the audio only over the threshold which was empirically preset. From each segmented audio sample, we extracted the feature $C(i,j)$. After that, $n(n-1)/2$ pairwise SVM for

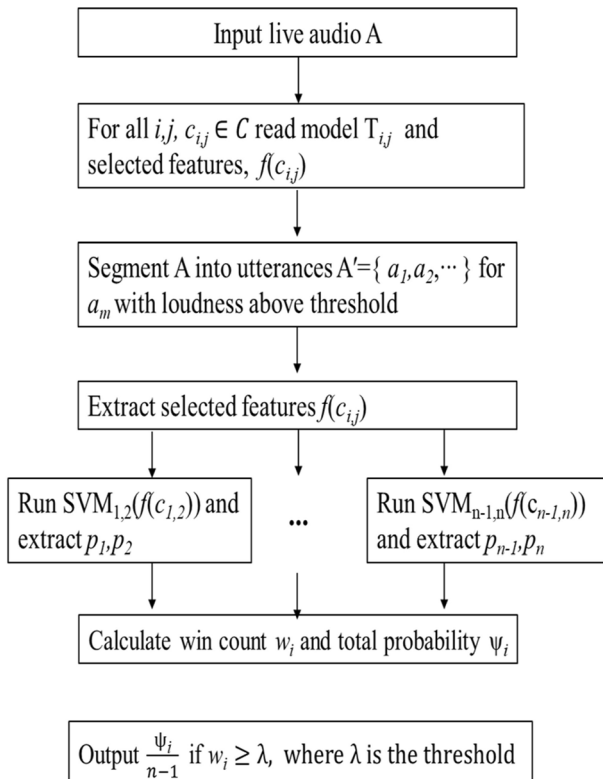


Fig. 6 The classifier architecture. $SVM_{i,j}$ computes the probabilities p_i and p_j for labels i, j , using features $f(c_{ij})$ [44]

n classes classify the features into 1 or 0 in parallel. Finally, the average of probability ψ_i of the binary SVMs is determined.

4 Experiments

4.1 Acoustic feature selection

There have been many studies demonstrating the correlation between emotional voices and acoustic features [17, 20, 28, 29, 48, 56, 66, 67]. Widely known characteristics of acoustic features in several high-arousal and negative discrete emotions are summarized in Table 5.

Among these acoustic features, for fear/non-fear voice recognition we used F0 variability which is the variability in the fundamental frequency of speech, corresponding to pitch variability (i.e., the energy or loudness in voice) and voice intensity, [19]. Because the acoustic data was phone calls from real-life situations, the callers ranged in age from children to the elderly and included both male and female. Thus, the data inevitably had significant individual variance, making normalization more difficult. For this reason, we measured the intra-speaker variability in both F0 and voice intensity for each set of acoustic data (one set = one call). Each acoustic feature extraction was based on the smaller partitioning of speech signals into small intervals of 20 msec frames. In addition to these acoustic features, MFCCs were used in this study. We empirically set 19 filters in the Mel filter bank and used the first 16 coefficients as feature vectors [51].

4.2 Behavioral feature selection

Disfluencies are one of the main behavioral features enabling us to distinguish fear from non-fear voices. There have been many disfluency studies that show correlations with emotional speech recognition [22]. Among many disfluency phenomena (e.g., repetition, hesitation, pause, laughter, inspiration, expiration), we used the interjection phenomena which provides quantifiable parameters and reveals significant results enabling the distinguishing fear and non-fear voices. Hence, the specific interjection features represented in fearful voices were extracted for emotional voice pattern recognition. To this end, the interjection occurrence pattern in both fear-defined data and non-fear defined phonological data in the sample data were observed and compared. There were three commonly observed components of the

Table 5 Selected acoustic features and their representation in the specific discrete (negative) and dimensional (high-arousal) emotional accounts from Juslin & Laukka (2003) and Laukka & Juslin (2005)

	Anger	Fear	High arousal
Mean F0	High	High	High
F0 variability	Much	Little	Much
F0 contour	Rising	Rising	
Voice intensity	High	Low (except in panic fear)	High
Voice intensity variability	Much	Much	Much
High-frequency energy	Much	Little	Much
Mean F1	High	Low	High
Speech rate	Fast	Fast	

interjection features in both fear and non-fear voices: pre-pause (a), quick phonological segment (b), and post-pause (c). We empirically set the average duration of two components (the pre-pause and the quick phonological segment) as 20–160 msec and 60–220 msec, respectively (Fig. 7).

There was a general difference in the duration of the three components between the interjection in fear and non-fear voices. Also, interjection occurred less often in non-fear voices than in fear voices. The average duration times for the three components in fear and non-fear voices were different (in fear voices and non-fear voices, 96.8 msec vs. 231.6 msec for pre-pause, 160.8 msec vs. 249.1 msec for the quick phonological segment, and 224.7 msec vs 288.3 msec for post-pause, respectively). To discover significant differences in duration between fear and non-fear voices, the duration of the pre-pause, quick phonological segment, and post-pause were statistically compared. Paired t-test results showed that the duration of the interjection of the two components (pre-pause and quick phonological segment) were statistically different between fear and non-fear voices ($t(9)=1.86$, $p < 0.05$ and $t(21)=2.99$ $p < 0.005$ respectively; Fig. 8). Therefore, we set these duration times as standards for distinguishing fear and non-fear voices in an interjection.

Disfluencies such as interjection and word repetition, etc., occur even during normal situations, but these phenomena appear more frequently in situations where the cognitive demand is greater [6, 63]. The greater cognitive load can be caused by negative arousal emotions [10]. Thus, we expected that interjection would be more frequently observed in fear voices than non-fear voices. This was confirmed by behavioral feature selection. In addition, we also found a distinctive feature of interjections for fear compared with non-fear voices: the duration of the pre-pause and the quick phonological segments were shorter in fear voices than in non-fear voices. We speculated these differences might be due to the contradiction between the greater cognitive load elicited by both the callers' negative arousal emotion and the callers' distressed feelings of responsibility for the urgent situation. Under emergency circumstances, callers' cognitive function temporarily become limited; thereby it takes longer for them to utter

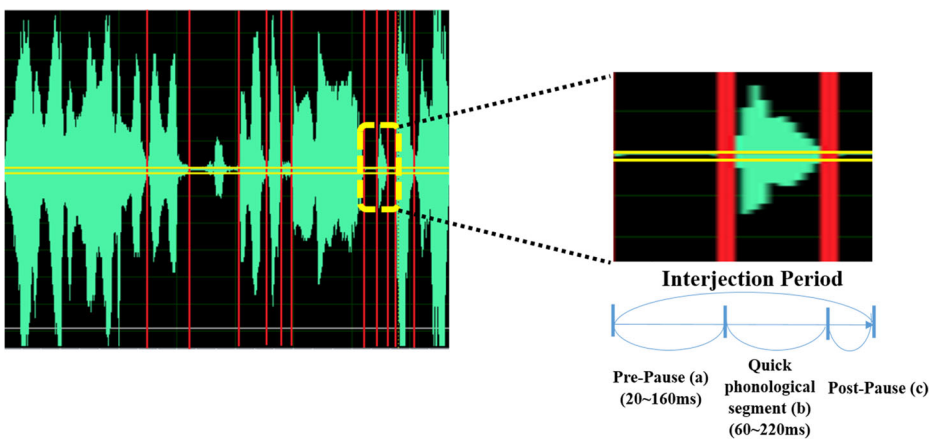


Fig. 7 Interjection spectrum during utterances and specific phase represented in fear situations (Left; specific utterance section in which interjection occurred, Right; magnified interjection from the left). An interjection is divided into three phases; pre-pause, quick phonological segment, and post-pause. The average duration of each phase were empirically set at 20–1760 msec for pre-pause, 60–220 msec for the quick phonological segment, and no specific average duration for post-pause

Duration of interjection

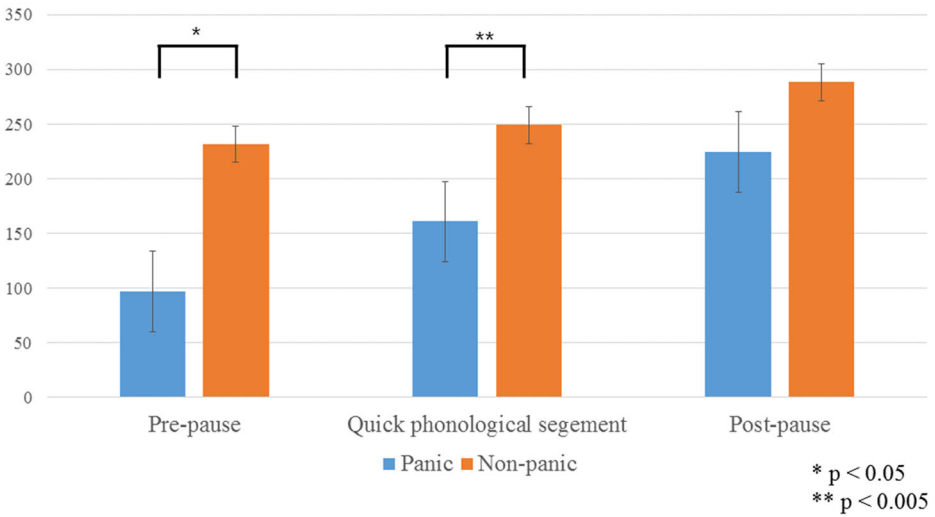


Fig. 8 Duration of interjection for each phase for fear (in blue) VS non-fear (in orange) voices (msec). Mean and standard error are displayed. Paired t-tests showed statistically significant differences in duration of pre-pause (left) and quick phonological segments (middle) between fear and non-fear voices. (* $p < 0.05$, ** $p < 0.005$)

the appropriate words or sentences needed to convey the required information. At the same time, the caller may feel pressure to speak as quickly and accurately as possible. This contradiction may give rise to the shorter average duration of the pre-pause and the quick phonological segment typical in fear voices.

4.3 Classification procedure

Six kinds of features (e.g., F0 variability only, Voice intensity variability, MFCC, interjection, the combination of Voice intensity variability with MFCC, and the combination of Voice intensity variability, MFCC, and interjection) were produced, and then SVM was used to classify the emotion categories. The results are shown in a confusion matrix in Table 6. “The condition positive in true condition” indicates a fear state; “the condition negative in true condition” indicates a non-fear state. The accuracy counts only true positive and true negative for fear state and non-fear state, respectively (Table 6).

Table 6 Confusion matrix for four-fold cross-validation support vector machines and calculation for accuracy with 162 data sets for each state (i.e., fear/non-fear)

Total population ($N = 324$)		Predicted condition	
		Predicted condition positive	Predicted condition negative
True condition	Condition positive (i.e., Fear, $N = 162$)	True positive	False negative
	Condition negative (i.e., Non-fear, $N = 162$)	False positive	True negative
		$\text{Accuracy(ACC)} = \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	

5 Results and discussion

Table 7 shows the result of the accuracy rate for each model for the 162 sets of test acoustic data used in the current study. The MFCC model displayed the best accuracy rate (80.2%). The voice intensity variability model achieved 68% accuracy, the interjection model achieved 58.9% accuracy, and the F0 variability achieved 48% accuracy. The combination models (which employed a combination of features), accomplished 73% accuracy for voice intensity variability. The best model was the MFCC with the interjection features model and the voice intensity variability which reached an accuracy rate of 66.5% (Table 7).

Our results showed that the MFCC feature model displayed the best accuracy of any feature or any combinations of feature models. The MFCCs feature is widely used for speech recognition and emotional speech recognition system. Also, contradicting our expectation that the interjection behavioral feature would provide a reasonable criterion for greater improvement in emotional speech recognition system, it only reached a moderate accuracy stated above. We attribute the unexpected results to the following reasons: The accuracy rates for all models were quite low compared to results of other studies which classified the emotion of speech, all of which achieved 50 to 95% accuracy [5, 11, 14, 22, 24, 28, 34, 37, 38, 40, 42, 43, 45, 47, 48, 50, 54, 55, 57–59, 61, 66]. Unlike previous emotional speech recognition system studies using databases consisting mostly of individual sentences recorded in unnatural, non-real-life surroundings, using high-technique mikes [43], we used real-acoustic data (i.e., recordings from an emergency call center) from natural surroundings including the background noise. These real emergency situations included naturally induced emotions. In addition, the data length varied from 3 s to 1156 s (Fig. 4) which led to data entailing a significant portion of contextual information and vast fluctuations in callers' emotional state. These distinctive features may have made accurate pattern recognition difficult which may have led to the overall low accuracy rate. This limitation needs to be surmounted.

The simplified classification system (i.e., fear/non-fear) could be another reason for the overall low accuracy rate. Only two categories for efficient and effective operations in an emergency call center were needed ('neutral' was considered as part of the non-fear category). This simple distinction, associated with many acoustic features, led to broader modeling than we expected and may have caused difficulty distinguishing emotions in speech. Nevertheless, the MFCC feature model showed the highest accuracy rate at 80.2%. In fact, among the few studies employing real acoustic data, the MFCC is the most important acoustic feature for emotional speech recognition systems [58]. Some research only used the MFCC feature for emotional speech recognition but yielded high precision [53]. Considering the nature of the data used here, and the simplified categories, we verified the MFCC feature as a vital acoustic feature for emotional speech recognition systems, even for acoustic data from real-life situations in which emotions were

Table 7 Accuracy rate results for each model in the test set data

Model	Specification	Accuracy rate
Model 1	F0 variability	48.0%
Model 2	Voice intensity variability	68.0%
Model 3	MFCC	80.2%
Model 4	Interjection	58.9%
Model 5	Voice intensity variability + MFCC	66.5%
Model 6	Voice intensity variability + MFCC + Interjection	73.0%

naturally induced. Regarding the interjection feature model; although this model only reached slightly above the level of chance for accuracy (i.e., 58.2%), it is very encouraging researchers that a behavioral feature like interjection was introduced for emotional speech recognition. In fact, there has been some research that considered disfluencies as a feature for emotional speech recognition [2, 13, 35]. Inspired by those studies, a few researchers thought about behavioral features such as laughter, stutters, and fillers in emotional speech classification, but the studies are very few in number [9]. This study applied a behavioral feature to emotional speech recognition for real acoustic data, not acted recordings, where real-life callers became emotional. Moreover, even though the result did not show as high accuracy rates as the all models, we showed a better accuracy rate (73%) for the combination of features (i.e., voice intensity variability, MFCCs, and interjection), over the other models (66.5%; i.e., voice intensity variability and MFCCs). This implies that behavioral features are necessary for the improvement of speech recognition systems, especially when using real-world acoustic data from natural situations.

6 Conclusion

In this study, we aimed to determine if the naturally induced emotional states demonstrated in acoustic data obtained from an emergency call center, which entails naturally induced emotions, corresponding to fearful emotions. To do so, we first identified the emotional states of callers under urgent situations and sorted audio files into two emotional categories: fear or non-fear. Then, using SVM, we attempted to automatically classify the speech data into fear/non-fear voices. Widely used acoustic features for emotional speech recognition, including F0 variability, voice intensity variability, and MFCCs, were used for classification. Importantly, we presented interjection, one of the disfluencies in utterances which are often observed when a speaker becomes negatively emotional, as the behavioral feature. We showed an 80.2% accuracy rate for classification results using only MFCCs, verifying that the MFCC is an important feature for emotional speech recognition system. Also, we learned that, even though its rate of accuracy was lower than the MFCC model, the combination of features, especially with the interjection behavioral feature included, improved the rate of accuracy of the classification, compared with models which only used acoustic feature combinations. Because emotion recognition in speech occurring in natural contexts is increasingly required, classification using behavioral features like interjection should be considered for further study.

In this study, we aimed to determine if the naturally induced emotional states demonstrated in acoustic data obtained from an emergency call center, corresponded to fear emotion. Importantly, we presented interjection, one of the disfluencies in utterances which are often observed when a speaker becomes negatively emotional, as the behavioral feature by which to classify the fear/non-fear state. We showed an 80.2% accuracy rate for classification results using only MFCCs, verifying that the MFCC is an important feature for ESRS. Because we confined the classification to either fear or non-fear, this may have led to the lower accuracy rate of our classification system. Even so, dealing with real emotionally aroused acoustic data in the fear state is meaningful. Also, we learned that the combination of features, especially with the interjection behavioral feature included, improved the rate of accuracy of the classification, compared with models which only used acoustic feature combinations. Because emotion recognition in speech occurring in natural contexts is increasingly required, classification using behavioral features like interjection should be considered for further study. In order to further develop emotion recognition system, our work needs to be extend to other real emotional acoustic data.

Acknowledgements This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00189, Voice emotion recognition and indexing for affective multimedia service)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Barrett LF (1998) Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognit Emot* 12(4):579–599
2. Batliner A, Fischer K, Huber R, Spilker J, Nöth E (2000) Desperately seeking emotions or: actors, wizards, and human beings. In: ISCA tutorial and research workshop (ITRW) on speech and emotion
3. Burges CJ (1998) A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc* 2(2): 121–167
4. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of German emotional speech. In: *Interspeech*, pp 1517–1520
5. Chavhan Y, Dhore M, Yesaware P (2010) Speech emotion recognition using support vector machine. *Int J Comput Appl* 1(20):6–9
6. Corley M, Stewart OW (2008) Hesitation disfluencies in spontaneous speech: the meaning of um. *Lang Linguist Compass* 2(4):589–602
7. Davison GC, Vogel RS, Coffman SG (1997) Think-aloud approaches to cognitive assessment and the articulated thoughts in simulated situations paradigm. *J Consult Clin Psychol* 65(6):950–958
8. Devillers L, Vasilescu I, Vidrascu L (2004) Anger versus fear detection in recorded conversations. In: *Proceedings of speech prosody*, pp 205–208
9. Devillers L, Vidrascu L, Lamel L (2005) Challenges in real-life emotion annotation and machine learning based detection. *Neural Netw* 18(4):407–422. <https://doi.org/10.1016/j.neunet.2005.03.007>
10. Dibble JL, Wisner AM, Dobbins L, Cacal M, Taniguchi E, Peyton A, van Raalte L, Kubulins A (2015) Hesitation to share bad news: by-product of verbal message planning or functional communication behavior? *Commun Res* 42(2):213–236
11. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recogn* 44(3):572–587
12. Fontaine JR, Scherer KR, Roesch EB, Ellsworth PC (2007) The world of emotions is not two-dimensional. *Psychol Sci* 18(12):1050–1057
13. Forbes-Riley K, Litman DJ (2004) Predicting emotion in spoken dialogue from multiple knowledge sources. In: *HLT-NAACL*. Citeseer, pp 201–208
14. Galanis D, Karabetos S, Koutsombogera M, Papageorgiou H, Esposito A, Riviello M-T (2013) Classification of emotional speech units in call centre interactions. In: *Cognitive infocommunications (CogInfoCom)*, 2013 IEEE 4th international conference on. IEEE, pp 403–406
15. Goberman AM, Hughes S, Haydock T (2011) Acoustic characteristics of public speaking: anxiety and practice effects. *Speech Comm* 53(6):867–876
16. Hamann S (2012) Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends Cogn Sci* 16(9):458–466. <https://doi.org/10.1016/j.tics.2012.07.006>
17. Iliou T, Anagnostopoulos C-N (2009) Statistical evaluation of speech features for emotion recognition. In: *Digital telecommunications, 2009. ICDT'09. Fourth International Conference on. IEEE*, pp 121–126
18. Izard CE, Libero DZ, Putnam P, Haynes OM (1993) Stability of emotion experiences and their relations to traits of personality. *J Pers Soc Psychol* 64(5):847
19. Juslin PN, Laukka P (2003) Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol Bull* 129(5):770–814. <https://doi.org/10.1037/0033-2909.129.5.770>
20. Kao Y-h, Lee L-s (2006) Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language. In: *InterSpeech*
21. Laukka P, Juslin P, Bresin R (2005) A dimensional approach to vocal expression of emotion. *Cognit Emot* 19(5):633–653
22. Lee CM, Narayanan SS (2005) Toward detecting emotions in spoken dialogs. *IEEE Trans Speech Audio Process* 13(2):293–303
23. Lee H, Kim E, Lee M (2003) A validation study of Korea positive and negative affect schedule: the PANAS scales. *Korean J Clin Psychol* 22(4):935–946

24. Lee F-M, Li L-H, Huang R-Y (2008) Recognizing low/high anger in speech for call centers. In: Proceedings of 7th international conference on signal processing, robotics and automation. World Scientific and Engineering Academy and Society (WSEAS), University of Cambridge, UK, pp 171–176
25. Lindsey AE, Greene JO, Parker RG, Sassi M (1995) Effects of advance message formulation on message encoding: evidence of cognitively based hesitation in the production of multiple-goal messages. *Commun Q* 43(3):320–331
26. Lindström A, Villing J, Larsson S, Seward A, Åberg N, Holtelius C (2008) The effect of cognitive load on disfluencies during in-vehicle spoken dialogue. In: INTERSPEECH, pp 1196–1199
27. Liscombe J, Riccardi G, Hakkani-Tür DZ (2005) Using context to improve emotion detection in spoken dialog systems. In: Interspeech, pp 1845–1848
28. Luengo I, Navas E, Hermáez I, Sánchez J (2005) Automatic emotion recognition using prosodic parameters. In: Interspeech, pp 493–496
29. Lugger M, Yang B (2007) The relevance of voice quality features in speaker independent emotion recognition. In: Acoustics, speech and signal processing, 2007. ICASSP 2007. IEEE International Conference on. IEEE, pp IV-17–IV-20
30. Mao X, Chen L, Fu L (2009) Multi-level speech emotion recognition based on HMM and ANN. In: Computer science and information engineering, 2009 WRI world congress on. IEEE, pp 225–229
31. Mauss IB, Robinson MD (2009) Measures of emotion: a review. *Cognit Emot* 23(2):209–237. <https://doi.org/10.1080/02699930802204677>
32. Mehrabian A, Russell JA (1974) *An approach to environmental psychology*. the MIT Press, Cambridge
33. Metzke F, Englert R, Bub U, Burkhardt F, Stegmann J (2009) Getting closer: tailored human–computer speech dialog. *Univ Access Inf Soc* 8(2):97–108
34. Morrison D, Wang RL, De Silva LC (2007) Ensemble methods for spoken emotion recognition in call-centres. *Speech Comm* 49(2):98–112
35. Narayanan S (2002) Towards modeling user behavior in human-machine interaction: effect of errors and emotions. In: Proc. ISLE workshop on dialogue tagging for multi-modal human computer interaction
36. Narayanan S, Georgiou PG (2013) Behavioral signal processing: deriving human behavioral informatics from speech and language: computational techniques are presented to analyze and model expressed and perceived human behavior-variously characterized as typical, atypical, distressed, and disordered-from speech and language cues and their applications in health, commerce, education, and beyond. *Proc IEEE Inst Electr Electron Eng* 101(5):1203–1233. <https://doi.org/10.1109/JPROC.2012.2236291>
37. Neiberg D, Elenius K (2008) Automatic recognition of anger in spontaneous speech. In: INTERSPEECH, pp 2755–2758
38. Neiberg D, Elenius K, Laskowski K (2006) Emotion recognition in spontaneous speech using GMMs. In: Interspeech, pp 809–812
39. Ostir GV, Markides KS, Black SA, Goodwin JS (2000) Emotional well-being predicts subsequent functional independence and survival. *J Am Geriatr Soc* 48(5):473–478
40. Pan Y, Shen P, Shen L (2012) Speech emotion recognition using support vector machine. *IJSH International Journal of Smart Home* 6(2):101–108
41. Panksepp J (1989) The neurobiology of emotions: of animal brains and human feelings
42. Pao T-L, Chen Y-T, Yeh J-H, Li P-J (2006) Mandarin emotional speech recognition based on SVM and NN. In: Pattern recognition, 2006. ICPR 2006. 18th International Conference on. IEEE, pp 1096–1100
43. Petrushin V (1999) Emotion in speech: recognition and application to call centers. In: Proceedings of artificial neural networks in engineering
44. Pfister T (2010) Emotion detection from speech. 2010
45. Picard RW, Vyzas E, Healey J (2001) Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Trans Pattern Anal Mach Intell* 23(10):1175–1191
46. Plutchik R (1980) A general psychoevolutionary theory of emotion. *Theories of Emotion* 1(3-31):4
47. Polzehl T, Schmitt A, Metzke F, Wagner M (2011) Anger recognition in speech using acoustic and linguistic cues. *Speech Comm* 53(9-10):1198–1209
48. Rao KS, Koolagudi SG, Vempada RR (2013) Emotion recognition from speech using global and local prosodic features. *Int J Speech Technol* 16(2):143–160
49. Russell JA (2003) Core affect and the psychological construction of emotion. *Psychol Rev* 110(1):145–172
50. Rv B (1984) *The characteristics and recognizability of vocal expression of emotions*. Walter de Gruyter, Inc., The Netherlands
51. Sahidullah M, Saha G (2012) Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Comm* 54(4):543–565
52. Salovey P, Kokkonen M, Lopes PN, Mayer JD (2004) Emotional intelligence: what do we know? In: *Feelings and emotions: the Amsterdam symposium, Jun, 2001, Amsterdam, Netherlands*. Cambridge University Press

53. Sato N, Obuchi Y (2007) Emotion recognition using mel-frequency cepstral coefficients. *IMT* 2(3):835–848
54. Schuller B, Rigoll G, Lang M (2003) Hidden Markov model-based speech emotion recognition. In: *Multimedia and expo, 2003. ICME'03. Proceedings. 2003 international conference on. IEEE*, pp 1–401
55. Schuller B, Rigoll G, Lang M (2004) Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: *Acoustics, speech, and signal processing, 2004. Proceedings. (ICASSP'04). IEEE international conference on. IEEE*, pp 1–577
56. Schuller B, Batliner A, Steidl S, Seppi D (2011) Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. *Speech Comm* 53(9-10):1062–1087
57. Tahon M, Delaborde A, Devillers L (2011) Real-life emotion detection from speech in human-robot interaction: experiments across diverse corpora with child and adult voices. In: *Interspeech*
58. Utane AS, Nalbalwar S (2013) Emotion recognition through speech using Gaussian mixture model and hidden Markov model. *IJARCSSE* 3(4)
59. Ververidis D, Kotropoulos C, Pitas I (2004) Automatic emotional speech classification. In: *Acoustics, speech, and signal processing, 2004. Proceedings. (ICASSP'04). IEEE international conference on. IEEE*, pp 1–593
60. Vidrascu L, Devillers L (2005) Annotation and detection of blended emotions in real human-human dialogs recorded in a call center. In: *Multimedia and expo, 2005. ICME 2005. IEEE international conference on. IEEE*, p 4 pp
61. Vidrascu L, Devillers L (2005) Detection of real-life emotions in call centers. In: *INTERSPEECH*, vol 10, pp 1841–1844
62. Watson D, Clark LA, Tellegen A (1988) Development and validation of brief measures of positive and negative affect: the PANAS scales. *J Pers Soc Psychol* 54(6):1063–1070
63. Wingate ME (1984) Fluency, disfluency, dysfluency, and stuttering. *J Fluen Disord* 9(2):163–168
64. Xiao Z, Dellandrea E, Dou W, Chen L (2005) Features extraction and selection for emotional speech classification. In: *Advanced video and signal based surveillance, 2005. AVSS 2005. IEEE conference on. IEEE*, pp 411–416
65. Yik MS, Russell JA, Barrett LF (1999) Structure of self-reported current affect: integration and beyond. *J Pers Soc Psychol* 77(3):600
66. Zhang S (2008) Emotion recognition in Chinese natural speech by combining prosody and voice quality features. In: *International symposium on neural networks. Springer*, pp 457–464
67. Zhu A, Luo Q (2007) Study on speech emotion recognition system in E-learning. In: *International conference on human-computer interaction. Springer*, pp 544–552



Shin-ae Yoon is a Post-doctoral research of the Department of Software at Sejong University, where she has been since December, 2016. She graduated in English Literature from Konkuk University (2006), obtained an M.S. in Second-Language processing from Konkuk University (2009) and a Ph.D. in Cognitive Science from Yonsei University (2016). She has experience in cognitive science, focusing on psycholinguistics, and engaged in the following subjects: neuro-linguistics and language processing.



Guiyoung Son is assistant researcher of the Department of Digital Contents at Sejong University since 2015. She received her M.S. in German Literature from Sungkyunkwan University, Seoul, in 2015. Her research interests are in the cognitive and affective sciences including recognition of affect from video, voice and bio-signals.



Soonil Kwon is an Associate Professor of the Department of software at Sejong University since 2009. He received his M.S. and Ph.D. in electrical engineering from University of Southern California (USC), Los Angeles, CA in 2000, 2005. From 2005 to 2006, he worked at Samsung Electronics as a Senior Engineer. He was also with Intelligent Interaction Research Center in Korea Institute of Science and Technology (KIST), Seoul, Korea as a senior research Scientist from 2006 to 2009. His current research interests include speech signal processing, speech recognition, emotion recognition, and human-computer interaction.