applied sciences

MDPI

Article

Call Redistribution for a Call Center Based on Speech Emotion Recognition

Milana Bojanić [1,*], Vlado Delić [1] and Alexey Karpov [2]

[1] Faculty of Technical Sciences, University of Novi Sad, 21000 Novi Sad, Serbia; vlado.delic@uns.ac.rs
[2] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, SPIIRAS, 14th Line 39, 199178 St. Petersburg, Russia; karpov@iias.spb.su
* Correspondence: milana.bojanic@uns.ac.rs

Received: 26 April 2020; Accepted: 25 June 2020; Published: 6 July 2020

Featured Application: A call redistribution method for a call center based on speech emotion recognition is proposed. The research goal is efficiency improvement in emergency call centers based on automatic recognition of more urgent callers.

Abstract: Call center operators communicate with callers in different emotional states (anger, anxiety, fear, stress, joy, etc.). Sometimes a number of calls coming in a short period of time have to be answered and processed. In the moments when all call center operators are busy, the system puts that call on hold, regardless of its urgency. This research aims to improve the functionality of call centers by recognition of call urgency and redistribution of calls in a queue. It could be beneficial for call centers giving health care support for elderly people and emergency call centers. The proposed recognition of call urgency and consequent call ranking and redistribution is based on emotion recognition in speech, giving greater priority to calls featuring emotions such as fear, anger and sadness, and less priority to calls featuring neutral speech and happiness. Experimental results, obtained in a simulated call center, show a significant reduction in waiting time for calls estimated as more urgent, especially the calls featuring the emotions of fear and anger.

Keywords: emotion recognition; intelligent speech signal processing; affective computing; human computer interaction; supervised learning

1. Introduction

Spoken language processing combines knowledge from the interdisciplinary area of natural language processing, cognitive sciences, dialogue systems, and information access. Speech Emotion Recognition (SER) and text-to-speech synthesis (TTS), including voice and style conversion, as part of human–machine spoken dialogue systems correspond to certain cognitive aspects underlying the human language processing system [1]. In the last few decades, there has been growing interest in developing human–machine interfaces that are more adaptive and responsive to a user's behavior [2]. In that sense, the use of emotion in speech synthesis and recognition of emotion in speech takes an important place in attempts to improve naturalness of human–machine interaction (HMI) [3]. As to TTS, different applications such as smart environments, virtual assistants, intelligent robots, and call centers have set requirements for different speech styles identified with corresponding emotional expressions [4]. Recognition of emotions in HMI is not restricted to speech analysis only, but also image analysis (facial expression recognition, eye-tracking data) and physiological signals (pulse rate, skin conductance, facial electromyography, electroencephalography (EEG) signal) [5]. Emotion recognition in spoken dialogue systems such as call centers provides a possibility to respond to callers according to the detected emotional state or to pass control over to human operators [2,6–8].

In the SER research, two main approaches are utilized in describing the emotional space. The first approach describes the emotional space with a finite number of prototypical emotions according with categorical emotion model. The second approach uses dimensions (typically arousal and valence) to determine possible emotional states in the space defined by chosen dimensions. The latter approach corresponds to dimensional emotion models. Dimensional emotion models mostly use two or three dimensions (e.g., valence, arousal, and sometimes dominance) to describe the emotional space in which the emotional variability is to the greatest extent determined by the first two dimensions and thus used as a basis for research in the field of SER [9]. The valence dimension describes the pleasantness of emotion and ranges from positive (e.g., joy) to negative (e.g., anger). The arousal dimension indicates the level of activation during some emotional experience and it ranges from passive (e.g., sleepiness) to active (e.g., high excitement). The position of some basic, categorical emotions in the valence–arousal space is shown in Figure 1.

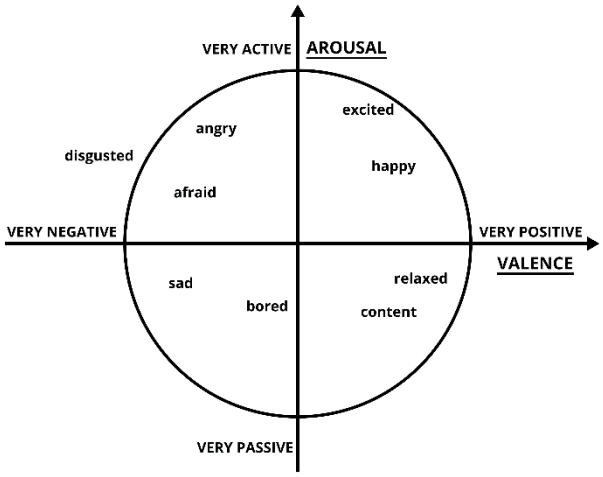

Figure 1. The circumplex model of affect in the valence–arousal space. Adapted from [10].

The dimensional models allow using emotional categories (appropriately positioned in a two-dimensional emotional space) among which it is possible to determine a distance metric [10]. Goncalves et al. utilized four dimensions (namely, valence, arousal, sense of control, and ease in achieving a goal) to describe user's emotional state while interacting with an electronic game [11]. Landowska proposed a procedure to obtain new mappings with mapping matrices for estimating the dimensions of a valence-arousal-dominance model from Ekman's six basic emotions [12]. The procedure, as well as the proposed metrics, might be used, not only in evaluation of the mappings between representation models, but also in a comparison of emotion recognition and annotation results. Emotion valence classification in self-assessed affect challenge is reported in [13]. Detection of a degree of speaker's sleepiness can help recognizing his/her emotional arousal as well [14]. Sometimes both approaches, emotion category and valence-arousal classification, are utilized for comparison, as in the INTERSPEECH Emotion Sub-Challenge on acted speech corpus [15].

In a situation when all call center operators are busy and unable to answer a new call, the system puts that call on hold regardless of its urgency. By way of illustration, if a call is the fifth call in the queue in a given moment, a caller which is terrified, angry, or upset would be left to wait for a certain period of time before his/her call is considered. This period of time is equivalent to the time in which all the preceding calls are answered. This classical approach in call centers does not take into consideration the urgency of a call and calls are processed in the order in which they are received. Petrushin utilized the emotion recognition as a part of a decision support system for prioritizing telephone voice messages in a call center and assigning a proper agent to respond the message [6]. His goal was to recognize two possible states: "agitation" which includes anger, happiness and fear, and "calm" which includes normal state and sadness. The average recognition accuracy was in the range of 73–77%.

In this research, the first presumption was that there are some calls which are more urgent and which should be processed faster. The second presumption was that the urgency of the call correlates with a caller emotional state reflected through speech. The motivation behind this research was to improve the effectiveness of call center service through giving the first level priority to the callers who are experiencing a negative valence emotional state (fear and anger), the second level priority to a sad or neutral emotional state, and the third level priority to a joyful emotional state. The proposed approach consists of recognition of caller's emotional speech and redistribution of the calls according to the proposed emotion ranking. Thus, faster processing and the decrease in waiting time for callers estimated as more urgent, is achieved.

The paper is organized as follows: Section 2 covers related works including acoustic modeling of emotional speech and the underlying emotional speech corpus, as well as methods for emotion classification. The proposed algorithm for redistribution of calls is described in detail in Section 3. The simulation and experimental results are reported and discussed in Section 4. Finally, conclusion remarks and future research directions are summarized in Section 5.

2. Materials and Methods

2.1. Emotional Speech Corpus

The GEES (Corpus of Verbal Expressions of Emotions and Attitudes—in Serbian: *Korpus Govorne Ekspresije Emocija i Stavova*) is the first corpus of acted emotional speech recorded in Serbian [16]. Six actors (3 female, 3 male) were recorded while verbally expressing semantically neutral textual material into five basic emotions: anger, joy, fear, sadness, and neutral. The underlying textual material included 32 isolated words, 30 short sentences, 30 long sentences, and one passage of 79 words. The corpus was evaluated by human listeners and reported recognition accuracy was 94.7% [16]. In this study, a part of corpus containing short and long sentences was taken into consideration because it better reflects a real conversation scenario. The isolated words and the passage were omitted from the research. Aiming to have each speaker equally represented, 58 recorded utterances from every speaker in each emotion class were used for the feature extraction. The total number of utterances used in experiments was 1740. It has been pointed out that acted emotions are more exaggerated than real ones [17] and discussed that acted emotions have limited application in real-life situations. Still, by studying the acoustic features of emotional speech on the acted emotion corpus, one can analyze acoustic variations and get insight into acoustic correlates of emotional speech. Those acoustic correlates are, to a greater extent, present in emotions occurring in real life situations or in elicited emotional speech. In that sense, the relationships between the acoustic features and the acted emotions, as well as between the acoustic features and the real life emotions, do not contradict [18]. Using acted emotions in emotional speech recognition is a way to obtain and study generic (maybe universal) expressions of emotions [19]. Additionally, our research setting was to recognize more intensive emotional states which are reflecting more urgent callers. These intensive vocal emotional expressions are more frequent in acted emotional speech corpora than in natural speech corpora.

2.2. Acoustic Modeling

The most commonly used acoustic features for SER are: prosodic features (pitch, intensity, duration), cepstral features (MFCC), spectral features (formant position and bandwidth), and occasionally voice quality features (harmonic-to-noise ratio, jitter, shimmer), in line with the studies [19–23]. The task of finding a robust feature set has led to the idea of applying statistical functionals to low-level descriptors (LLD) and resulted in very large feature vectors containing up to a few thousands of prosodic and spectral features [19]. Recently, new trends in machine learning have been directing research of automatic affect recognition towards end-to-end technique that combines deep, convolutional and recurrent neural networks trained directly on underlying raw audio signal [24,25]. A proposal of multilevel model based on a combination of LLDs and convolutional recurrent neural

network model is given in [26]. Still, a lot of research in the area is based on hand-crafted features that have shown to be robust in many computational paralinguistics tasks such as emotion, autism, accent, addressee, deception, cognitive and physical load detection, and so on [20–22] (list of the INTERSPEECH Paralinguistics Challenge tasks up to 2019 is available at http://www.compare.openaudio.eu/tasks/). Schuller et al. introduced the INTERSPEECH 2013 ComParE feature set [20]. It contained 6373 features including energy, spectral, cepstral (MFCC) and voicing related LLDs (pitch, voicing probability, jitter, shimmer), as well as a few LLDs including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity and psychoacoustic spectral sharpness, etc. This set of hand-crafted acoustic features is still state-of-the-art now [27]. Another more minimalistic feature set proposed in [23] includes prosodic, excitation, vocal tract, and spectral descriptors, obtained by applying functionals to 18 LLDs that give a total of 52 utterance-level features only. Kaya et al. used the ComParE feature set for the proposed cascaded normalization. The proposed normalization approach, combining speaker level, value level and feature vector level normalization, has shown a superior performance in the task of cross-corpus acoustic emotion recognition on five corpora recorded in five languages [28]. Utterance level features, obtained through the statistical analysis of prosodic features (pitch, energy), spectral information (formants, spectrum centroid and spectrum cut-off frequency) and cepstral information (mel-frequency bands energy), are extracted to recognize seven basic emotions in Mandarin [29]. While in some studies SER relies on the prosodic and voice quality feature set only [30], and in others on cepstral features only [31], our previous study showed that a combination of both spectral and prosodic features has a higher discrimination capability for speech emotion recognition than prosodic or spectral features used separately [32]. Wagner et al. compared and discussed the advantages and usability of hand-crafted and learned representations (an end-to-end system that learns the data representation directly from the raw waveforms) [33]. Their research suggests that hand-crafted features can better generalize to unseen data and can also provide a better robustness to various acoustic conditions in comparison to purely end-to-end systems.

The proposed approach to acoustic modeling is based on the statistical analysis of the acoustic feature contours and it is performed in three steps. The openSMILE toolkit [34], used as official baseline for the series of INTERSPEECH Computational Paralinguistics challenges, is used to extract the acoustic feature set. The first step includes the extraction of short-term pitch, energy and 12 MFCC values on a frame basis. Additionally, the voicing probability and the zero crossing rate are calculated for every frame. Sequences of those short-term pitch, energy and MFCC values form feature contours. In the second step, the first derivative of the acoustic features is calculated in order to model the dynamics of speech parameters. The third step of the feature extraction process involves a statistical analysis of the feature contours. The proposed set of 12 statistical functionals has been chosen from three groups of functionals which are the most frequently used [19]:

1. The first four moments (mean, standard deviation, skewness and kurtosis);
2. Extrema and their positions (minimum, maximum, range, the relative position of minimum and the relative position of maximum);
3. Regression coefficients (the slope and the offset) and the mean squared regression error.

Finally, the extracted feature set results in 384 features for each of the processed utterances.

2.3. Classification Methods

A recent survey in the field of SER provided an overview of traditional classifiers and deep learning algorithms applied for SER [35]. Among traditional classifiers, they listed Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Support Vector Machines (SVM), and Artificial Neural Networks (ANN), Decision Trees (DT), k-Nearest Neighbor (kNN), k-means, and Naive Bayes Classifiers, concluding that there is no generally accepted machine learning algorithm used in this field. Recently, the focus on research changed direction towards Deep Neural Networks (DNN), with most widely used Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). For the

purpose of speech emotion classification in this study, Linear Discriminant Classifiers (LDC) and kNN are taken into account due to their simplicity, efficiency and low computational requirements. LDCs and kNN classifiers have been used since the very first studies and turned out to be quite successful for both acted and spontaneous emotional speech [19]. Zbancioc et al. used a weighted kNN classifier for the classification task of four emotions (anger, sadness, joy and neutral) contained in the SROL emotion corpus utilized in their research [36].

In all our experiments, 10-fold partitioning of the data set was used to estimate the recognition accuracy of a particular classifier. Training and test sets included utterances from all six speakers, so these results belong to speaker-dependent experiments. Although "speaker-independent" experiments (e.g., leave-one-speaker-out) are possible on the GEES with 6 speakers (for example, the results reported by [37]), we decided to perform speaker-dependent tests in order to train classifier with more samples belonging to different speakers. In such a way, the acoustic variability present in the feature space is better modelled providing better prediction ability even when tested with an unknown speaker. Indeed, the accuracies obtained in speaker-independent cross-validation tend to be lower than the accuracies obtained in speaker-dependent cross-validation [37], but not significantly [38].

The first considered classifier is the linear Bayes classifier with the underlying assumption that classes are modeled by Gaussian densities and equal covariance matrices. Maximum likelihood estimates of Gaussian density parameters are used. As to the linear Bayes classifier, the average recognition accuracy achieved in our emotion classification experiments on the GEES corpus was 91.5% [32]. Joy was recognized with 84.2% and anger with 88.8% recognition rate. Class recognition rates for fear, neutral and sadness in the case of linear Bayes classifier were 92.5%, 97.1% and 94.8%, respectively. Table 1 shows a normalized confusion matrix for the linear Bayes classifier applied on the GEES corpus. From Table 1, it could be noted that sadness is misrecognized as a neutral state in 4% of test samples and fear is confused with neutral state in 3.2%. Neutral state has the highest recognition rate, thus its misclassification as fear and joy is about 1%. Anger and joy have lower recognition rates due to the problem of mutual misclassification, about 11% of anger test samples is recognized as joy and almost 15% of joy is misrecognized as fear.

Table 1. Normalized confusion matrix for linear Bayes classifier.

	Recognized Emotion Class (%)				
True Emotion Class	**Anger**	**Fear**	**Joy**	**Neutral**	**Sadness**
Anger	88.8	0	11.2	0	0
Fear	1.4	92.5	1.4	3.2	1.4
Joy	14.9	0.6	84.2	0.3	0
Neutral	0	1.1	1.2	97.1	0.6
Sadness	0	1.2	0	4	94.8

For the second classifier, the kNN classifier is used as a very intuitive method that classifies unlabeled examples based on their similarity to examples in the training set. It implicitly involves non-parametric density estimation, which leads to a very simple approximation of the linear Bayes classifier. Employing high dimensionality feature vectors, dimensionality reduction is sometimes applied in order to improve classification results, as in [39], where a speaker-penalty graph learning is proposed to penalize the impact of different speakers. Due to the fact that the recognition accuracy of the kNN classifier is affected by the high dimensionality of feature set, linear discriminant analysis (LDA) feature reduction has been applied on feature set [40]. In the five-class emotion classification task on the GEES corpus, the kNN classifier achieved the average recognition accuracy of 91.3% after LDA feature reduction and with k = 9. The lowest class recognition rate was obtained for joy (83.6%) and anger (86.8%). Regarding fear, neutral and sadness, higher class recognition rates were achieved—93.7%, 95.9% and 96.3%, respectively. Employing LDA, kNN achieved the average accuracy almost equal to the best result in our experiments (91.5%). In the case of linear Bayes classifier,

there were no improvements after LDA feature reduction probably due to good linear separability between classes in the original feature space. Using both classification methods in our SER experiments, lower recognition results obtained for joy and anger may be explained with the observed tendency in human perception tests to misclassify anger and joy from the GEES corpus [16].

In our earlier study, the SER experiments on the same GEES corpus were performed using a multilayer perceptron (MLP) with one hidden layer [41]. The number of neurons in the input layer was equal to the number of extracted features (same feature vector as described in Section 2.2), and the number of neurons in the output layer was equal to the number of emotion classes (5). MLP was trained using standard backpropagation (BP) algorithm with varying number of neurons in the hidden layer. The highest recognition rate was achieved with 15 neurons in the hidden layer. Further increase in neurons in the hidden layer resulted in insignificant improvement of the recognition rate at the cost of increased computational complexity and thus longer processing time.

The average recognition accuracy achieved in emotion classification experiments with MLP was 90.4%. Joy was recognized with 82.5% and anger with 86.5% recognition rate. In the case of these two emotions, results of MLP underperform results of the linear Bayes classifier approximately by 2%. Sadness and neutral are emotions with the highest recognition rates of 97.7% and 93.7%, respectively. Fear is recognized with 91.7%.

It can be noted that two emotions with the lowest recognition rates, namely joy and anger, have a lower recognition accuracy compared to the experimental results with the linear Bayes classifier. This is an additional reason why we decided to use the linear Bayes classifier in the proposed system, besides it is a fast classification method.

2.4. Comparison of the SER Results with Other Studies

In general, it is a difficult task to objectively compare results of one SER research with other results reported in literature. This is due to a high diversity of research approaches to SER, regarding speech emotion corpora, the extracted feature set, classification methods and additional experimental settings (e.g., speaker-dependent or speaker-independent tests, cross-validation method applied). Regarding acted speech, two corpora have been used in plenty of research: Berlin Emotional Speech Database (Emo-DB) containing the total of 535 sentences uttered by 10 actors (5 male, 5 female) in seven emotional states, and Danish Emotional Speech Database (DES) containing the total of 419 utterances portrayed in five emotional states by 4 actors (2 male, 2 female) [28,37]. In the research by Hassan et al. the proposed 3DEC classification was tested on all three corpora (Emo-DB, DES, GEES), and the best results were achieved on the GEES corpus [37]. It can be explained by the fact that the GEES contains more samples available for training the classifier than other two corpora, and by the fact that the overall human recognition accuracy reported for GEES is 94.7%, against 86.1% for Emo-DB and 67.3% for DES. The overall human accuracy reflects the distinction degree of the acoustic representation of basic emotions in a corpus, which is very high for the GEES corpus.

In our earlier study [42], the comparison of basic emotion classification in valence-arousal space was made on the Emo-DB and the GEES corpora. The mapping of basic emotions into three classes along the valence axis (positive, neutral, and negative), and three classes along the arousal axis (high, neutral, and low) was performed. The recognition results along the arousal axis were above 90% for both corpora. The average recognition results along the valence axis were 83.2% for the GEES and 76.9% for Emo-DB. It is in line with the findings showing that arousal discrimination tasks, based on acoustic features, achieve higher recognition rates than valence discrimination tasks [28].

We consider the GEES with 1740 utterances portrayed in 5 emotional states by 6 actors as a suitable and adequate basis for SER research. Also, taking into account that Serbian is still an under-resourced language, there are far less available emotional speech data and the corresponding research for Serbian (GEES is the only one emotional speech corpus accessible for research purposes) even in comparison with other Slavic languages like Russian [43], Czech [44], etc.

A comparative analysis of our results and the results of some other SER studies conducted on the GEES corpus was performed. Due to the fact that this is a rather small corpus in Serbian, it was not a subject of much research. Two SER researches on the GEES corpus were found to be compared with our results.

The first study for comparison is by Hasan et al. [37], who proposed a hierarchical classification technique using SVM for binary emotion classification on every level. As to feature extraction they decided to apply a "brute force" approach and 6552 acoustic features for each utterance. The extracted feature vector included 56 low-level descriptors (among which is pitch, energy, spectral energy, MFCCs) and 39 statistical functionals applied to these LLDs and their first and second derivatives. In the experiments three acted databases were used: the Danish Emotional speech (DES), the Berlin database (Emo-DB) and the Serbian GEES database, and one spontaneous database (Aibo corpus). The proposed hierarchical classification, called 3DEC, is based on input data in such a way that input data and their confusion plots determine the hierarchy of the proposed classification scheme. They used both speaker-dependent and speaker-independent approaches for SVM-based model training and testing. We present only results of speaker-dependent tests as to be able to make a comparison with our results. For the speaker-dependent test, 10-fold cross-validation for the whole corpus is applied, as in our case.

The reported [37] average recognition accuracy on the GEES corpus is 94.1%. It is achieved with the proposed 3DEC combination of SVM classifiers in the speaker-dependent test. Recognition accuracy in their research is obtained as an unweighted average accuracy (UA), i.e., accuracy per class is averaged by the total number of classes. It should be noted that in the case of the GEES corpus, UA accuracy is identical to weighted average accuracy (WA) due to equally balanced emotion classes. Comparing our result with the result of Hasan et al. [37], it can be seen that our average recognition accuracy is lower by 3%. It should be noted that our result is obtained with a significantly smaller feature vector (384 features against 6552 features in [37]). Additionally, the classification methods applied are different. In our experiments, the linear Bayes classifier is used as a simple and fast method for training and test stages, and their proposed 3DEC combination of SVMs requires training of four SVMs. We consider that our proposed SER achieves a slightly lower result compared to the best recognition accuracy reported for the GEES (94.1% in [37]), but having significantly smaller feature vectors and computationally less demanding classification method.

One more study on the GEES corpus, by Shaukat et al. [45], applied the multistage (hierarchical) emotion categorization with SVM. In their research, the extracted utterance-level vectors of 318 features, among which are pitch, energy, MFCC, formants and their statistical functionals (e.g., mean, variance, maximum, minimum, etc.). In the experiment on the GEES corpus, they reported the average emotion recognition rate of 90.63%.

Comparing our result with the result of Shaukat et al. [45], it can be seen that our average recognition accuracy is higher by 1%. They applied a hierarchical classification techniques with 4 SVMs, thus training of all 4 SVMs is necessary. It should be noted that feature vector set used in [45] is smaller, but an important difference is that their experiments were performed on individual speaker sub-corpora and overall recognition accuracy was calculated as an average value of recognition accuracies obtained for each individual speaker. Our recognition accuracy is evaluated after 10-fold cross-validation on the whole corpus, like in the study of Hasan et al. [37], which we consider as a more objective measure of recognition performance.

3. Algorithm for Call Redistribution Based on Speech Emotion Recognition

As mentioned earlier, in this research, one presumption was that the urgency of the call correlates with the caller's emotional state reflected through speech. Our focus was on emergency call centers and health care centers for elderly people. Aiming to recognize more urgent callers among them, we have proposed the ranking of five basic emotions.

So, basic emotions with negative valence (fear, anger and sadness) reflect unpleasantness of the speaker and our presumption was that those speakers have a health, or any other, more urgent problem.

On the other hand, there are positive valence emotions (e.g., joy) and neutral valence (neutral state) that are supposed to reflect less urgent speaker's state and those calls could be processed later.

The proposed ranking of five basic emotions is:

1. Fear (f)
2. Anger (a)
3. Sadness (s)
4. Neutral (n)
5. Joy (j).

In the proposed ranking, fear is put first because it is an emotion that people experience when facing a serious problem (serious injuries, heart attack, accidents, etc.). In the research conducted on the CEMO corpus containing dialogues recorded in a real-world medical call center, it was pointed out that patients had often expressed stress, pain, fear of being sick or even real panic [8]. Fear is the most common emotion in the CEMO corpus, with different levels of intensity and many variations [7]. Anger is the second negative and high arousal emotion, expressed in various stressful and disturbing situations. Sadness is in third place. It is an emotion with negative valence which is typical for elderly and lonely people. Holmen et al. reported that experiencing loneliness had a negative influence on the state of mood, so loneliness and sad mood prevailed especially among elderly subjects with cognitive difficulties [46]. Joy is in last place because it is considered to reflect full satisfaction and good mood, which are not indicators of urgent states.

The research setting is explained using an example of five calls received at the same moment—while all operators are busy. For each call, the initial part of the caller's speech is recorded. This recording is further processed and the feature vector x^i is extracted. The feature vector is forwarded to a classifier which gives one of the five emotion labels (anger, joy, fear, sadness, and neutral) to input speech. Finally, after SER, those five calls are redistributed according to the recognized emotions and the proposed emotion ranking. The proposed framework of call processing is shown in Figure 2. For example, in the scenario shown in Figure 2, the original call order was neutral, joyful, sad, afraid, and angry; after SER and proposed call redistribution, the system will firstly process the call featuring fear, then the call featuring anger, afterwards a sad caller, then neutral, and at last the call featuring joy.

The proposed algorithm, whose block diagram is shown in Figure 3, has the following steps:

1. When a call is received while all operators are busy, the system asks for the reason of the call and records the caller's speech for about 5–8 s. This recording contains about 1–2 sentences, depending on the dialogue strategy, which will be processed quickly by SER while the call is put on hold. For each recording, the feature vector x consisting of 384 features is extracted.
2. The extracted feature vector is input to our trained SER classifier. The classifier outputs one of the five emotion labels (fear, anger, sadness, neutral and joy) to the input speech. Keywords recognized by automatic speech recognition (ASR) can also be used for sentiment analysis, but it depends on both language and type of the call center.
3. If there are several calls on hold at the same time, they are redistributed based on the associated emotion label. Redistribution is done according to the introduced priority vector p:

$$p = [p_1 = f,\ p_2 = a,\ p_3 = s,\ p_4 = n,\ p_5 = j]^T, \tag{1}$$

where f represents fear, i.e., it denotes the speaker recognized as being in a state of fear, a denotes the speaker recognized as angry, s marks the speaker recognized as sad, n refers to neutral, and j to a joyful state of the speaker. The introduced priority vector, i.e., emotion ranking, represented in Equation (1), is proposed considering application in emergency call centers and health care centers for elderly people. It should be noted that the proposed algorithm is not restricted to the aforementioned priority vector only. Regarding a specific domain of application, a new emotion ranking can be adopted.

4. Calls are processed in the new order which is obtained after their emotion labeling based on SER (and ASR) and redistribution according to the proposed emotion ranking, i.e., the priority vector. Firstly, all callers that are recognized as afraid are processed, after them angry callers and so on. In the end, the callers recognized as joyful are processed. The final goal of the redistribution is reduction in waiting time for the callers recognized as the priority. Let us denote the waiting time t_{1i} of a caller i without SER and call redistribution, where $i = 1, \ldots, C$ and C is the number of calls received at the same moment. Then, t_{2i} denotes the waiting time of a caller i after SER and call redistribution (after application of the proposed algorithm). The objective function is:

$$\max \sum_{i=1}^{C} t_{1i} - t_{2i}, \tag{2}$$

according to the priority vector p. The objective function is formulated as to maximize waiting time reduction for the callers recognized as the priority regarding the priority vector p. So, the goal of call redistribution is to maximize waiting time reduction for the caller i, if the caller i is set as priority regarding the vector p. In our experiments this is the case for the caller recognized as being afraid—fear is in first place in the priority vector p. Afterwards, the objective function maximizes the waiting time reduction for the caller recognized as being angry, since anger is in second position in the priority vector p.

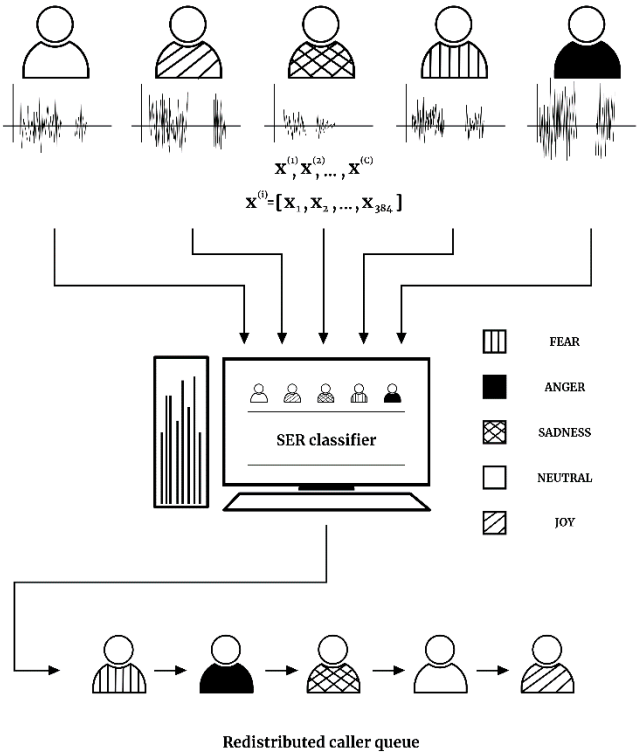

Figure 2. Proposed call redistribution based on SER.

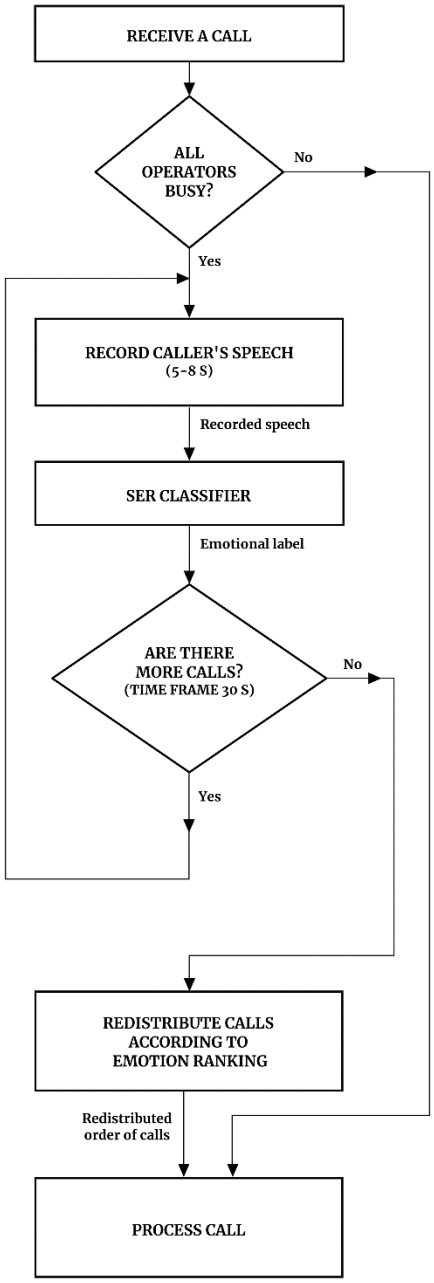

Figure 3. Block diagram of the proposed algorithm.

The call processing and the proposed algorithm are intended to be a part of a client-server application based upon computer-telephone integration (CTI). The main part of the application is running on the server side located on a remote computer. A client side is located at a call center. When a call is received, if there is at least one free operator, the call is answered immediately. In the case all operators are busy at that moment, the client side initiates a connection with the server which is waiting for clients. After the connection is established, a new session is started and the client sends a recorded speech sample of the call. On the server side, the feature vector is extracted for a received speech record and it is forwarded to SER module which classifies it into one of the predefined emotion categories. If a new call is received at a call center within a period of 30 s, the client sends the recorded speech sample of the second call and steps 1 and 2 of the algorithm are performed. The established session lasts as long as there are new calls within the time frame of 30 s, which is chosen as the overlapping time between two consecutive calls. When there are no more calls within the specified time period, all calls processed during the current session of client-server communication are redistributed according to the

proposed emotion ranking. The call redistribution is intended to be applied on a finite number of calls received in a short period of time while all the operators are busy. In the experiments, the situations of three, five and seven simultaneously received calls were considered. Let us denote them as group of calls. For example, when a call center simultaneously receives seven calls, those seven calls will be redistributed according to the SER system output and processed in a new order. While operators are answering those seven calls, if there is a new incoming call, it will be put in a new group of calls to be redistributed. The proposed emotion ranking can be specified after the connection establishment, so that the server adapts the system response to the specific type of a call center (a client). At the end of the session, the server sends the client the list of redistributed calls which are then processed according to the redistributed order.

4. Simulation and Experimental Results

The research was designed as a set of experiments in a simulated call center receiving a different number of calls simultaneously, i.e., during a short period of time when all operators are busy. The experiments focused on: (i) the redistribution of calls based on emotion label assigned after speech emotion recognition task, and (ii) the evaluation of time period in which the call was put on hold without and after speech emotion recognition was applied for call redistribution. During experiments, the number of simultaneously received calls varied from 3 to 7. In all experiments, prosodic and spectral feature set was used and the linear Bayes classifier and kNN were considered as classification techniques, as described in Section 2.

An average waiting time, without and with the redistribution, for each emotional state is evaluated as an average value of waiting time obtained for 50 experimental iterations in the simulated call center with one ideal active human-operator (a human factor is not considered). This procedure is repeated for each experimental setting (3, 5, and 7 simultaneously received calls). Waiting time reduction estimate is made under assumptions about underlying distribution of emotions in input calls and distribution of call duration. We assumed that all the emotions had a uniform distribution as well as that call durations were uniformly distributed across the chosen range (from 30 s to 3 min 50 s). The specified range was chosen with the assumption that it is wide enough to take into consideration the duration of shorter, medium, and longer phone calls as well. Thus, the evaluated waiting time after call redistribution may be shorter for every caller proportionally to the number of active operators in the call center.

A pseudo-random number generator is used for the generation of emotion labels (random choice of emotion for input call) as well as the generation of input call duration. The order of the calls in queue (the order of the call arrival) has, as in simulation as in real-world call center, the biggest influence on the estimated waiting time which a caller could spend in callers' queue. In our simulations, the order of the call arrival featuring specific emotion is also unknown and thus determined by generated pseudo-random number. Thus, regarding every iteration in simulation, the random number of occurrences of each emotion class with the associated random call duration, and finally random order of calls (emotions) in callers' queue, jointly influence the variations of estimated average waiting time, before and after call redistribution. Additionally, the recognition rate of some emotional state has an influence on the average waiting time after call redistribution.

Simulation of call redistribution in a call center is explained on an experimental example for three simultaneously received calls. Each call is represented by one utterance in the GEES corpus. Firstly, the vector of randomized emotion labels for three input calls was generated. According to the input emotion label vector, three utterances belonging to chosen emotion classes were randomly (regarding a speaker) selected from the corpus and provided as an input to SER. As an initial part of the simulation, duration of a call, generated as a random value between 30 s and 3 min 50 s, was appended to each of these utterances. Knowing the initial order of the simulated calls (determined with input emotion label vector), the initial waiting time in the caller's queue is calculated for each caller as a sum of call duration for all preceding callers in the queue. Every caller is represented with input utterance determined with input emotion label. Thus, initial waiting time for every emotion class is evaluated. Secondly,

based on the classifier output each input utterance gets one of the five emotion labels, thus output emotion label vector is obtained. Given the output emotion label, calls are redistributed according to the priority vector. New waiting time is calculated for each caller based on the new position in redistributed caller's queue. Accordingly, new waiting time for every emotion class is evaluated.

Table 2 shows the average waiting time which a caller will spend if his/her call is among three calls received at the same moment while all operators are busy, before and after application of SER and call redistribution. It can be observed that there is a significant waiting time reduction for callers recognized as being in a state of fear: initially, they were waiting for about 2 min 40 s, and after SER and the proposed call redistribution they had to wait only 8 s. In the case of an angry caller, there is also a noticeable waiting time reduction: the initial waiting time was 2 min 19 s and after redistribution only about 1 min. In the case of a sad caller, there is little time saving expressed in few seconds: the initial waiting time was 2 min and after redistribution reduced to 1 min 45 s. Regarding neutral and joyful emotional states of the caller, there is an increase in waiting time after SER and call redistribution: about 1 min increased waiting time for a neutral caller and about 2 min for a joyful caller. This increase was expected as the redistribution always places callers with these emotions at the end of callers' queue.

Table 2. Average waiting time when 3 calls are received simultaneously while all operators are busy.

Emotion	without the Proposed Algorithm [min]:[s]	after Application of SER and Call Redistribution [min]:[s]
fear	2:43	0:08
anger	2:19	1:01
sadness	2:00	1:45
neutral	2:09	3:07
joy	1:56	4:07

The average waiting time which a caller will spend if his/her call is among five calls received in a short period of time while all operators are busy, before and after the application of the proposed algorithm, is shown in Table 3. In the case of fear as the first in emotion ranking, there is the biggest and significant decrease in waiting time: from 4 min 17 s to 25 s after SER and redistribution. There is also a significant decrease in waiting time for angry callers: from 4 min 36 s to 1 min 57 s.

Table 3. Average waiting time when 5 calls are received simultaneously while all operators are busy.

Emotion	without the Proposed Algorithm [min]:[s]	after Application of SER and Call Redistribution [min]:[s]
fear	4:17	0:25
anger	4:36	1:57
sadness	4:49	3:56
neutral	4:07	5:13
joy	3:24	7:34

Unlike the experiment with three calls at the same time, in the experiment with five calls, callers recognized as being sad have achieved nearly 1 min shorter waiting time after SER and redistribution. In the case of a neutral state, the waiting time is increased for about 1 min. For callers recognized as being joyful, the increase is larger and amounts to about 4 min.

Table 4 shows the average waiting time which a caller will spend if his/her call is among 7 calls received simultaneously, i.e., in a short period of time while all operators are busy. As in two previous experimental settings, three emotions ranked as the priority one (fear, anger and sadness) have a significant decrease in waiting time. Calls featuring fear have the biggest waiting time reduction: it amounts to about 5 min 40 s. Calls featuring anger have achieved 2 min 20 s reduction in waiting time. In the case of a sad caller, the achieved decrease in waiting time is about 1 min. It can be observed that neutral and joyful callers have an increase in waiting time: 2 min 37 s and 5 min 30 s, respectively.

Table 4. Average waiting time when 7 calls are received simultaneously while all operators are busy.

Emotion	without the Proposed Algorithm [min]:[s]	after Application of SER and Call Redistribution [min]:[s]
fear	6:39	0:54
anger	6:33	4:11
sadness	7:16	6:24
neutral	6:12	8:49
joy	6:10	11:40

The comparative results of average waiting time in all three experimental settings (3, 5, and 7 simultaneously received calls) regarding the callers in all five emotional states, are shown in Figure 4. As the callers in the state of fear have the highest priority, their average waiting time is significantly reduced in all experimental settings, even up to twenty times in the case of three simultaneously received calls, ten times in the case of five simultaneously received calls, and six times reduced in the case of seven calls. Angry callers are given the second priority in redistribution, so in all experiments the decrease in their average waiting time is achieved. In the case of three and five simultaneously received calls, the waiting time after redistribution is reduced to less than half of the waiting time before redistribution. In the case of seven simultaneously received calls, the waiting time is reduced by one third of the initial waiting time.

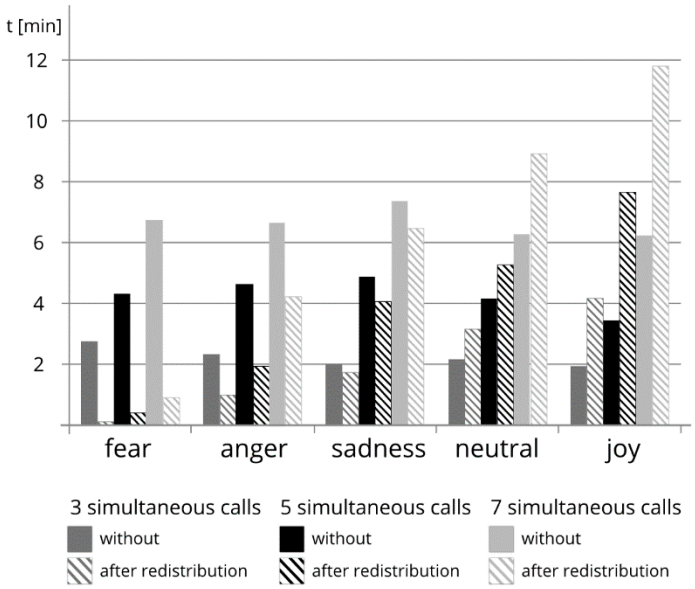

Figure 4. Average waiting time for five emotional states, without and after application of the proposed call redistribution.

From experimental results shown in Figure 4, it can be noticed that sad callers will have a moderate decrease in waiting time after the proposed call redistribution. The absolute value of waiting time reduction is the biggest in the case of seven simultaneously received calls, but the relative value of reduction is the biggest in the case of five calls and it amounts to about 18% of the initial waiting time.

As can be observed from Figure 4, callers in a neutral state have increased waiting time after call redistribution, about 1 min increase in the case of three and five simultaneously received calls, and about 2 min increase in the case of seven simultaneously received calls. Joy is marked as the emotion with the lowest priority, which is why callers featuring joy are put at the end of the caller's queue. It causes a significant increase in waiting time for the caller in a state of joy, about twice as longer waiting time after the proposed call redistribution in all experimental settings.

Table 5 shows the waiting time reduction for five emotional states after SER and the proposed call redistribution is applied, in all experimental settings (with 3, 5 and 7 simultaneously received calls) in a simulated call center. Time reduction is calculated as difference between the average waiting time

without the call redistribution and the average waiting time after application of SER and the proposed call redistribution:

$$\Delta t_e = \bar{t}_{1e} - \bar{t}_{2e}, \tag{3}$$

where $\bar{t}_{1e}$ denotes the average waiting time for a caller in the emotional state e without SER and call redistribution, e denotes one of the five emotional states (fear, anger, sadness, neutral and joy), and $\bar{t}_{2e}$ denotes the average waiting time for a caller in the emotional state e after application of SER and call redistribution.

Table 5. Waiting time reduction after the proposed call redistribution is applied. Time is expressed in [min]:[s].

Emotion	3 Calls Simultaneously	5 Calls Simultaneously	7 Calls Simultaneously
fear	2:35	3:52	5:45
anger	1:18	2:39	2:22
sadness	0:15	0:53	0:52
neutral	−0:58	−1:06	−2:37
joy	−2:11	−4:10	−5:30

The positive values of waiting time reduction in Table 5 indicate the real reduction in waiting time after call redistribution, which is the case of the callers recognized as being in a state of fear, anger or sadness. Negative values of waiting time reduction indicate that waiting time after call redistribution is actually increased, which is the case of the callers recognized as being in a neutral or joyful state. From the results presented in Table 5, it can be observed that as the number of simultaneously received calls grows, the calls featuring three recognized emotions considered as indicators of more urgent caller's state, namely fear, anger and sadness, show the tendency to have a decreased waiting time after the proposed call redistribution. On the other hand, the calls featuring recognized neutral speech and joy show tendency of increased waiting time as the number of simultaneously received calls grows, but it is considered justified as long as more urgent calls are processed instead of less urgent one.

To examine the results in the case of larger number of iterations, the simulations were performed using 200, 500, and 1000 iterations in all three experimental settings (3, 5, and 7 simultaneously received calls). For each experimental setting, obtained results are presented in Tables 6–8, respectively. Regarding initial average waiting time, even with 1000 iterations there are differences in evaluated initial average waiting time across five emotional states due to combination of random order of emotions in callers' queue and random duration of each call in the queue. Similar to the experiments with 50 iterations, after application of SER and call redistribution, calls featuring fear and anger have achieved significant reduction in waiting time. Unlike the simulation with 50 iterations, calls featuring sadness achieved in some cases slight increase and in some cases slight decrease in waiting time after call redistribution. This can be explained with the fact that neutral callers are put in the middle of callers' priority, so it was expected that their waiting time after increased number of iterations is evaluated as slightly changed initial average value. As can be observed from Tables 6–8, callers recognized as being in neutral and joyful states will have increased waiting time, similar to the results obtained in the simulation with 50 iterations.

Experimental results show the decrease in waiting time of the prioritized emotions. Indeed, there is a minor probability of misrecognizing anger as joy (because both are characterized by a high arousal, but opposite valence poles), and placing that caller at the end of the callers' queue, but possible negative effect depends on the position of such a call in original queue and emotional states of other callers in it. Overall experimental results show an essential decrease in waiting time of the prioritized emotions with negative valence.

In real-world emergency call centers, it is unlikely to expect all emotions equally distributed, as it was case in our simulation experiments. It is more likely to receive more calls featuring fear and less calls featuring joy, as it is reported for the CEMO corpus recorded in a real-world medical call center [7].

Although the results of the proposed SER might be to a certain extent lower in real-world emergency call center, we consider, based on high recognition accuracy for fear, sadness, and neutral that the proposed approach to SER and call redistribution based on it would improve effectiveness of such call center service.

Table 6. Average waiting time when 3 calls are received simultaneously while all operators are busy.

	without the Proposed Algorithm [min]:[s]			**after Application of SER and Call Redistribution [min]:[s]**		
Emotion\iterations	**200**	**500**	**1000**	**200**	**500**	**1000**
fear	2:03	2:05	2:07	0:13	0:15	0:13
anger	2:10	2:00	2:13	1:12	1:21	1:06
sadness	2:08	2:02	2:17	1:49	2:14	2:13
neutral	2:14	2:07	2:11	3:14	3:07	3:09
joy	2:29	2:19	2:07	4:09	3:59	4:07

Table 7. Average waiting time when 5 calls are received simultaneously while all operators are busy.

	without the Proposed Algorithm [min]:[s]			**after Application of SER and Call Redistribution [min]:[s]**		
Emotion\iterations	**200**	**500**	**1000**	**200**	**500**	**1000**
fear	4:21	4:33	4:25	0:23	0:27	0:30
anger	4:27	4:03	4:04	2:14	2:24	2:30
sadness	4:14	4:16	4:19	4:24	4:21	4:22
neutral	4:28	4:21	4:30	6:16	6:21	6:17
joy	4:26	4:17	4:11	8:18	7:56	8:07

Table 8. Average waiting time when 7 calls are received simultaneously while all operators are busy.

	without the Proposed Algorithm [min]:[s]			**after Application of SER and Call Redistribution [min]:[s]**		
Emotion\iterations	**200**	**500**	**1000**	**200**	**500**	**1000**
fear	6:24	6:30	6:32	1:00	0:54	0:48
anger	6:18	6:35	6:22	3:30	3:40	3:42
sadness	6:41	6:09	6:30	6:31	6:32	6:26
neutral	6:12	6:34	6:29	9:07	9:19	9:22
joy	6:28	6:31	6:14	12:04	12:05	12:08

5. Conclusions

The presented research has addressed the problem occurring in emergency call centers when there are several incoming calls in a short period of time while all operators are busy. The proposed solution takes into account a caller's emotional state, by recognizing emotion in speech and giving priority to the caller with negative valence emotion (fear, anger and sadness). The research aims to improve efficiency of emergency call centers based on recognition of more urgent callers. Utilizing the proposed emotion ranking and call redistribution, there is a significant reduction in waiting time for the callers recognized as being in the state of fear. A noticeable waiting time reduction is also achieved in the case of callers recognized to be angry, and a slight reduction in the case of callers recognized to be sad. On the other hand, the algorithm puts neutral and joyful callers at the end of the call queue, so those callers will have an increased waiting time. This is the price to be paid, and it has been considered that less urgent callers are more capable of bearing a longer waiting time.

Additionally, the waiting time for the most urgent calls can be shortened by giving the signal to operators who process lower priority calls that there is an emergency call on hold. Depending on

the dialogue strategy in a call center, the current call will be ended faster or put on hold, so that an emergency call would be received immediately.

Although there are evident differences between the emotional speech corpus recorded in a real call center and the acted emotional speech corpus recorded under controlled conditions, the experimental results in the simulated call center give a promising sign that the proposed approach to SER and call redistribution based on it would improve effectiveness of a real call center service. The proposed algorithm is a basis for detecting critical users in the specific type of call centers considered in the research.

Other SER techniques can be used instead of the proposed one, with similar results related to the improvement of a call center effectiveness. The proposed SER based on hand-crafted features (like at the OpenSMILE toolkit) could be faster and more robust in real conditions than any DNN or end-to-end based SER system, particularly in the case of a rather small GEES corpus, i.e., the only one available in Serbian that was suitable for the presented research. Due to the lack of available data, any DNN- or end-to-end-based SER system for Serbian could not be trained well, and there is a high risk of model over-fitting. In the only emotional speech corpus for under-resourced Serbian (GEES), there are just 1800 utterances, which is definitely not enough for state-of-the-art NN-based approaches.

Further research should consider "in the wild" recordings from real-world call centers (emergency call centers or health care centers for elderly people), so that the proposed approach could be tested on realistic data and its efficiency verified. Further research may also be directed toward combining paralinguistic and linguistic information. Recordings of the initial part of a call (1–2 sentences with duration of 5–8 s) in human–machine dialogue can be used as input not only into SER, but also into ASR. After ASR, recognized keywords can be used as an additional indicator of certain emotional states and thus priorities. It could increase reliability of the emotion estimation and utility of the proposed algorithm, even in the case of a lower arousal, i.e., more passive levels of emotion activation. Of course, a possible fusion of SER and ASR depends on the dialogue strategy, and the language and vocabulary expected in particular human–machine interactions.

Author Contributions: Conceptualization, M.B. and V.D.; methodology, M.B., V.D. and A.K.; formal analysis, M.B.; investigation, M.B.; writing—original draft preparation, M.B.; writing—review and editing, V.D. and A.K.; visualization, M.B.; supervision, V.D. and A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work has resulted from cooperation between researchers from two institutions at the project HARMONIC (ERA.Net RUS Plus, 2017-2021) related in part to human–machine interaction, as well as supported by the Russian Science Foundation project #18-11-00145 (Section 2.2).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Delić, V.; Perić, Z.; Sečujski, M.; Jakovljević, N.; Nikolić, J.; Mišković, D.; Simić, N.; Suzić, S.; Delić, T. Speech technology progress based on new machine learning paradigm. *Comput. Intel. Neurosc.* **2019**, *2019*, 4368036:1–4368036:19. [CrossRef]
2. Lee, C.M.; Narayanan, S.S. Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech Audio Process.* **2005**, *13*, 293–303. [CrossRef]
3. Ten Bosch, L. Emotions, speech and ASR framework. *Speech Commun.* **2003**, *40*, 213–225. [CrossRef]
4. Suzić, S.; Delić, T.; Pekar, D.; Delić, V.; Sečujski, M. Style transplantation in neural network-based speech synthesis. *Acta Polytech. Hung.* **2019**, *16*, 171–189. [CrossRef]
5. Wrobel, M. Applicability of Emotion Recognition and Induction Methods to Study the Behavior of Programmers. *Appl. Sci.* **2018**, *8*, 323. [CrossRef]
6. Petrushin, V. Emotion in speech: Recognition and application to call centers. In Proceedings of the Conference on Artificial Neural Networks in Engineering (ANNIE), St. Louis, MO, USA, 7–10 November 1999; pp. 7–10.

7. Vidrascu, L.; Devillers, L. Five emotion classes detection in real-world call center data: The use of various types of paralinguistic features. In Proceedings of the International Workshop on Paralinguistic Speech–between Models and Data (PARALING'07), Saarbrucken, Germany, 3 August 2007; DFKI: Saarbrucken, Germany, 2007; pp. 11–16.
8. Devillers, L.; Vaudable, C.; Chastagnol, C. Real-life emotion-related states detection in call centers: A cross-corpora study. In Proceedings of the INTERSPEECH 2010, Makuhari, Chiba, Japan, 26–30 September 2010; pp. 2350–2353.
9. Nicolaou, M.A.; Gunes, H.; Pantic, M. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. Affect. Comput.* **2011**, *2*, 92–105. [CrossRef]
10. Russell, J. A circumplex model of affect. *J. Pers. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
11. Gonçalves, V.P.; Costa, E.P.; Valejo, A.; Filho, G.; Johnson, T.M.; Pessin, G.; Ueyama, J. Enhancing intelligence in multimodal emotion assessments. *Appl. Intell.* **2017**, *46*, 470–486. [CrossRef]
12. Landowska, A. Towards New Mappings between Emotion Representation Models. *Appl. Sci.* **2018**, *8*, 274. [CrossRef]
13. Montacié, C.; Caraty, M. Vocalic, Lexical and Prosodic Cues for the INTERSPEECH 2018 Self-Assessed Affect Challenge. In Proceedings of the INTERSPEECH 2018, Hyderabad, India, 2–6 September 2018; pp. 541–545. [CrossRef]
14. Gosztolya, G. Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds. In Proceedings of the INTERSPEECH 2019, Graz, Austria, 15–19 September 2019; pp. 2413–2417. [CrossRef]
15. Gosztolya, G.; Busa-Fekete, R.; Toth, L. Detecting Autism, Emotions and Social Signals Using AdaBoost. In Proceedings of the INTERSPEECH 2013, Lyon, France, 25–29 August 2013; pp. 220–224.
16. Jovičić, S.T.; Kašić, Z.; Djordjević, M.; Rajković, M. Serbian emotional speech database: Design, processing and evaluation. In Proceedings of the 9th International Conference Speech and Computer—SPECOM'2004, St. Petersburg, Russia, 20–22 September 2004; pp. 77–81.
17. Williams, C.; Stevens, K. Emotions and speech: Some acoustical correlates. *J. Acoust. Soc. Am.* **1972**, *52*, 1238–1250. [CrossRef] [PubMed]
18. Ayadi, M.E.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
19. Schüller, B.; Batliner, A.; Steidl, S.; Seppi, D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.* **2011**, *53*, 1062–1087. [CrossRef]
20. Schuller, B.; Steidl, S.; Batliner, A.; Vinciarelli, A.; Scherer, K.; Ringeval, F.; Chetouani, M.; Weninger, F.; Eyben, F.; Marchi, E.; et al. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In Proceedings of the INTERSPEECH 2013, Lyon, France, 25–29 August 2013; pp. 148–152.
21. Schuller, B.; Steidl, S.; Batliner, A.; Epps, J.; Eyben, F.; Ringeval, F.; Marchi, E.; Zhang, Y. The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. In Proceedings of the INTERSPEECH 2014, Singapore, 14–18 September 2014; pp. 427–431.
22. Schuller, B.; Steidl, S.; Batliner, A.; Bergelson, E.; Krajewski, J.; Janott, C.; Amatuni, A.; Casillas, M.; Seidl, A.; Soderstrom, M.; et al. The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring. In Proceedings of the INTERSPEECH 2017, Stockholm, Sweden, 20–24 August 2017; pp. 3442–3446. [CrossRef]
23. Eyben, F.; Scherer, K.R.; Schüller, B.W.; Sundberg, J.; Andre, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput* **2016**, *7*, 190–202. [CrossRef]
24. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), Shanghai, China, 20–25 March 2016; pp. 5200–5204. [CrossRef]
25. Papakostas, M.; Spyrou, E.; Giannakopoulos, T.; Siantikos, G.; Sgouropoulos, D.; Mylonas, P.; Makedon, F. Deep Visual Attributes vs. Hand-Crafted Audio Features on Multidomain Speech Emotion Recognition. *Computation* **2017**, *5*, 26. [CrossRef]
26. Zheng, C.; Wang, C.; Jia, N. An Ensemble Model for Multi-Level Speech Emotion Recognition. *Appl. Sci.* **2020**, *10*, 205. [CrossRef]

27. Schuller, B.; Batliner, A.; Bergler, C.; Messner, E.M.; Hamilton, A.; Amiriparian, S.; Baird, A.; Rizos, G.; Schmitt, M.; Stappen, L.; et al. The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks. In Proceedings of the INTERSPEECH 2020, Shanghai, China, 25–29 October 2020.
28. Kaya, H.; Karpov, A. Efficient and effective strategies for cross-corpus acoustic emotion recognition. *Neurocomputing* **2018**, *275*, 1028–1034. [CrossRef]
29. Chen, L.; Mao, X.; Wei, P.; Xue, Y.; Ishizuka, M. Mandarin emotion recognition combining acoustic and emotional point information. *Appl. Intell.* **2012**, *37*, 602–612. [CrossRef]
30. Fernandez, R.; Picard, R. Recognizing affect from speech prosody using hierarchical graphical models. *Speech Commun.* **2011**, *53*, 1088–1103. [CrossRef]
31. Nwe, T.; Foo, S.; De Silva, L. Speech emotion recognition using hidden Markov models. *Speech Commun.* **2003**, *41*, 603–623. [CrossRef]
32. Delić, V.; Bojanić, M.; Gnjatović, M.; Sečujski, M.; Jovičić, S.T. Discrimination capability of prosodic and spectral features for emotional speech recognition. *Elektron. ir Elektrotechnika* **2012**, *18*, 51–54. [CrossRef]
33. Wagner, J.; Schiller, D.; Seiderer, A.; André, E. Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant? In Proceedings of the INTERSPEECH 2018, Hyderabad, India, 2–6 September 2018; pp. 147–151. [CrossRef]
34. Eyben, F.; Weninger, F.; Groß, F.; Schuller, B. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In Proceedings of the 2013 ACM Multimedia Conference, Barcelona, Spain, 21–25 October 2013; pp. 835–838. [CrossRef]
35. Akçay, M.B.; Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *116*, 56–76. [CrossRef]
36. Zbancioc, M.; Feraru, S. The Analysis of the FCM and WKNN Algorithms Performance for the Emotional Corpus SROL. *Adv. Electr. Comput. Eng.* **2012**, *12*, 33–38. [CrossRef]
37. Hassan, A.; Damper, R.I. Classification of emotional speech using 3DEC hierarchical classifier. *Speech Commun.* **2012**, *54*, 903–916. [CrossRef]
38. Rybka, J.; Janicki, A. Comparison of speaker dependent and speaker independent emotion recognition. *Int. J. Appl. Math. Comput. Sci.* **2013**, *23*, 797–808. [CrossRef]
39. Xu, X.; Huang, C.; Wu, C.; Wang, Q.; Zhao, L. Graph learning based speaker independent speech emotion recognition. *Adv. Electr. Comput. Eng.* **2014**, *14*, 17–22. [CrossRef]
40. Bojanić, M.; Delić, V.; Sečujski, M. Relevance of the types and the statistical properties of features in the recognition of basic emotions in the speech. *Facta Univ. Ser. Electron. Energetics* **2014**, *27*, 425–433. [CrossRef]
41. Bojanić, M.; Crnojević, V.; Delić, V. Application of neural networks in emotional speech recognition. In Proceedings of the 11th Symposium on Neural Network Applications in Electrical Engineering, Belgrade, Serbia, 20–22 September 2012; pp. 223–226. [CrossRef]
42. Bojanić, M.; Gnjatović, M.; Sečujski, M.; Delić, V. Application of dimensional emotion model in automatic emotional speech recognition. In Proceedings of the 2013 IEEE 11th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 26–28 September 2013; pp. 353–356. [CrossRef]
43. Verkholyak, O.; Kaya, H.; Karpov, A. Modeling Short-Term and Long-Term Dependencies of the Speech Signal for Paralinguistic Emotion Classification. *SPIIRAS Proc.* **2019**, *18*, 30–56. [CrossRef]
44. Partila, P.; Tovarek, J.; Voznak, M.; Rozhon, J.; Sevcik, L.; Baran, R. Multi-Classifier Speech Emotion Recognition System. In Proceedings of the 26th Telecommunications Forum TELFOR'18, Belgrade, Serbia, 20–21 November 2018; pp. 1–4. [CrossRef]
45. Shaukat, A.; Chen, K. Emotional State Categorization from Speech: Machine vs. Human. *arXiv* **2010**, arXiv:1009.0108.
46. Holmen, K.; Ericsson, K.; Winblad, B. Quality of life among elderly: State of mood and loneliness in two selected groups. *Scand. J. Caring Sci.* **1999**, *13*, 91–95. [CrossRef]