Detection of Negative Emotional States in Real-World Scenario

Theodoros Kostoulas, Todor Ganchev, Iosif Mporas, Nikos Fakotakis
Wire Communications Laboratory, Department of Electrical and Computer Engineering,
University of Patras, 26500 Rion-Patras, Greece
{tkost, tganchev, imporas, fakotaki}@wcl.ee.upatras.gr

Abstract

In the present work we evaluate a detector of negative emotional states (DNES) that serves the purpose of enhancing a spoken dialogue system, which operates in smart-home environment. The DNES component is based on Gaussian mixture models (GMMs) and a set of commonly used speech features. In comprehensive performance evaluation we utilized a well-known acted speech database and real-world speech recordings. The real-world speech was collected during interaction of naïve users with our smart-home spoken dialogue system. The experimental results show that the accuracy of recognizing negative emotions on the real-world data is lower than the one reported when testing on the acted speech database, though much promising, considering that, often, humans are unable to distinguish the emotion of other humans judging only from speech.

Index Terms*: emotion recognition, spoken interaction*

1. Introduction

The increasing use of spoken dialogue systems in the modern society raises the need for more effective and user-friendly interaction between human and machine. To this end, a number of multimodal dialogue systems, whose main modality is speech, have been reported [1, 2, 3]. However, most of these dialogue systems do not take into consideration the emotional state of the user with whom they interact. Indisputably, awareness about the emotional state of the user would contribute significantly for improvement of the performance of dialogue systems. Thus, detecting and modeling the emotional state of the user can provide a feedback to the dialogue flow manager and secure the basis for more successful interaction.

Various approaches for emotion recognition have been reported in the literature [4]. Mixed modeling of human's emotional states from speech and visual information has been used as well. In [5], Wang & Guan used information derived from both speech and facial expressions reporting overall accuracy of 82.1%. Pal et al. [6] reported emotion detection from infant facial expressions and cries. The utilization of the speech signal resulted to accuracy of 74.2% against an overall 75.2% resulting from fusion of image and sound information.

Much work has been done concerning emotion modeling and recognition exploiting only speech. Past research has shown that both acoustic and linguistic information can be used to create classification models for emotion recognition. In [7], Schuller et al. have presented a comparison of several concepts for robust fusion of prosodic and verbal cues in speech emotion recognition, achieving average accuracy of 90.3%. Yacoub et al. [8] has reported emotion recognition results from speech through feature extraction from short utterances typical of Interactive Voice Response (IVR) applications indicating 94.0% accuracy when distinguishing hot anger and neutral utterances. Liscombe et al. [9] have reported an increase on classification accuracy of 2.6% when augmenting standard lexical and prosodic features with contextual features using a corpus from a spoken dialogue system. Devillers & Vidracu [10] have used lexical and paralinguistic cues to detect individual's emotion on spoken dialogues collected from a medical emergency call center. Burkhardt et al. [11] have collected data form a pilot voice portal and examined how to predict anger. Ai et al. [12] have used users' performance features to improve the classification accuracy of emotions in computer tutoring dialogues. Nisimura et al. [13] have reported 98.8% on distinguishing delightful from hateable emotions by analyzing real-world data from the interaction of children with a dialogue system. Rutaru & Litman [14] have used word-level features for predicting emotions during a speech enabled tutoring system showing improvement in emotion prediction over using utterance-turn-level features. Batliner et al. [15] have focused on finding prosodic correlates of children's emotional speech when interacting with a robot.

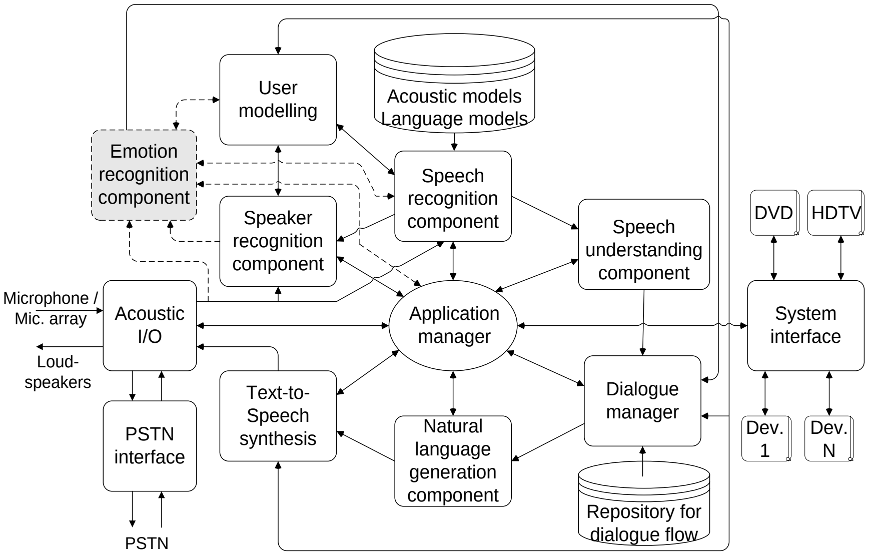


Figure 1. Overall architecture of the smart-home system

Lee & Narayanan [16] have studied data obtained from a call center application and introduced an information-theoretic notion of "emotional salience". Their experiments have proved that the combination of acoustic and language information gives the best results for their data.

The present work reports recent progress in our efforts to develop an emotion recognition component, which is going to be integrated in a smart-home system [17]. Awareness about the emotional state of the users is expected to improve the system's friendliness and would definitely contribute to detecting and avoiding trouble in the user-system communication.

To this end, our emotion recognition component is implemented as detector of negative emotional states (DNES) that is based on evidence detected only from speech. Due to the specifics of our application the DNES operates in a speaker-dependent mode, relying on support from speaker recognition and user modeling components.

In the present work, we mainly place the accent on the evaluation of the DNES on real-world data, which were collected during interactions of naïve users with the smart-home dialogue system outlined in Section 2. For comprehensiveness of evaluation and better understanding the complexity of human performance in near real-world scenario, a direct comparison between the performance of the emotion detection component on acted speech [18] and real-world data is presented.

The remaining of this work is organized as follows: Section 2 offers a brief outline of our smart-home spoken dialogue interaction system, and Section 3 presents the architecture of the developed emotion detection component. The speech data utilized in the experimental evaluation are described in Section 4. Section 5 reports the experimentations and results and Section 6 offers brief discussion about the practical significance of these results. This paper ends with summary and conclusions, presented in Section 7.

2. The smart-home system

The smart-home system [17] utilized in the present study was developed during the FP6 IST INSPIRE project (IST-2001-32746). The scope of this project was to research and develop a home automation system that can control various intelligent appliances either locally (from inside the home) or remotely (through a PSTN).

Figure 1 illustrates the overall architecture of the integrated system, which via speech interaction provides user-friendly access to information, entertainment devices and control of intelligent home appliances installed in smart-home environment. In the Greek prototype, deployed at test site of the Wire Communications Laboratory, University of Patras, the system interface and the intelligent appliances (lamps, air conditioner, TV, radio, etc) were simulated via software graphical interface, which provides animated feedback to the users. The components of smart-home system are distributed over three personal computers, which are linked via local network, plus some specialized hardware.

The smart-home system was designed with open architecture, which allows new components to be added for extending the system's functionality. One such addition is the emotion recognition component (shown in Figure 1 with dashed-line box, which has a gray fill), whose connections to the other components are shown with dashed arrows. To this end, the emotion recognition component is implemented as detector of negative emotional states. By utilizing the feedback from this component, which ideally would capture the emotional state of the users, we intend to improve the dialogue interaction. Specifically, the new DNES component aims at providing awareness about the presence of problem during interaction between the dialogue system and the user.

In brief, the DNES component has the purpose of monitoring the dialogue between the user and the system and detect if the user has negative emotions due to troubles while trying to achieve his/her goals or due to any other reason. When a negative emotional state is identified, the DNES component provides this information to the application manager and the dialogue flow manager (please refer to Figure 1). Once the system is aware about the trouble, it can take a number of actions to address the issue. Generally speaking, these actions would depend on the user's status and preferences, on the state of dialogue, on the circumstances during the spoken interaction, as well as on the availability of other information. Logically, the smart-home system could proceed by offering assistance to the user, or utilize the information for performing automatic adaptation of the speech interaction strategy, etc.

3. Detector of negative emotional states

In the present work, we rely on the assumption that there is one predominant emotion per utterance, and this emotion can be detected from evidence obtainable from speech signal.

The architecture of the DNES component is presented in Figure 2. As illustrated, the training procedure (shown in the gray rectangle) and the operational mode share common preprocessing and feature extraction steps. In brief, during training representative labeled data, i.e. emotionally tagged speech utterances from the speech corpora are utilized for creating the speaker-dependent models. Specifically, the speech recordings are subject to pre-processing, feature extraction and silence-speech separation. The speech feature vector obtained after speech parameterization consists of the fundamental frequency, frame energy, harmonicity (harmonics-to-noise ratio) and the 13 first Mel-frequency cepstral coefficients (MFCCs). The fundamental frequency and harmonics-to-noise ratio (HNR) are calculated using Boersma's algorithm [19], and the MFCCs according to the methodology of Davis and Mermelstein [20]. The MFCC and the frame energy were estimated for time window of 25 ms, with time skip of 10 ms. In order to achieve an accurate estimation of the fundamental frequency, we utilized frame size of 40 ms, and for computing the HNR frame size of 80 ms. Similarly to the MFCC, the fundamental frequency and HNR were computed for sliding frames with time step of 10 ms. The frequency range in which the fundamental frequency is searched was set to [75, 400] Hz.

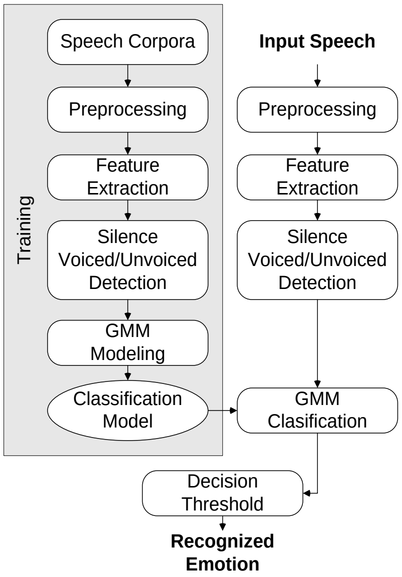


Figure 2. Internal structure of the detector of negative emotional states.

The speech parameters computed so far are appended in a 16-dimensional feature vector, which contains the information required to perform silence/voiced/unvoiced detection. Since the unvoiced speech doesn't carry much information about the emotional state of the speaker, only the feature vectors which represent the voiced speech frames were utilized for creating a GMM for the two classes. Here for each speaker we consider one GMM that represents the negative and one for the non-negative emotional states. Diagonal covariance matrix was used for modeling the underlying distributions of each class. The parameters of the GMMs were adjusted to the training data through the Expectation Maximization (EM) algorithm [21]. Specifically, during the initialization step the k-means clustering algorithm is employed to estimate the initial values of the mixture means. Afterwards, a variable number of EM iterations is performed, where the actual number of iterations depends on the data distribution and on a threshold that specifies the error reduction ratio between two consecutive iterations. When the training of the speaker-dependent models is completed, the DNES component is ready for operation.

During the operational stage each frame of the input speech signal is parameterized. The 16-dimentional feature vectors are next categorized utilizing the models built during training. Specifically, the log-likelihoods computed for the individual GMMs are utilized for making a decision on per frame basis. Afterwards, a final decision is made for the entire speech utterance by applying a predefined speaker-independent threshold on the averaged scores per class. (A speaker-dependent threshold is preferred solution for our application, since it would give to each user the flexibility to adjust the functionality of the DNES component to his/her preferences and needs, but for the purpose of simplicity in the present work we report only results for the scenario where a common speaker-independent threshold is considered for all users). This threshold adjusts the desired balance between false alarm and miss recognition errors. In fact, it specifies the percentage of negatively recognized frames required for making a decision that the emotional state of the input utterance is negative.

4. Speech corpora

During the development of the DNES component we relied on English language acted speech database [18]. Specifically, this well-known database was utilized in investigation and comparative evaluation of various speech parameterization techniques, and afterwards for the choice of speech features and classifier.

Since our smart-home dialogue system operates in near real-world conditions and targets Greek language users, Greek emotional speech recordings were utilized to train the speaker-dependent acoustic models. These recordings were collected during the interaction of naïve users with the smart-home system.

We assume that our DNES component, which operates on acoustic level and does not exploit linguistic information, behaves consistently between these two languages, and that the difference in performance between the two datasets (reported in Section 5) is due mainly to their different nature: acted speech recorded by professional actors vs. speech recorded from naïve users in near real-world scenario. In the following we provide a brief outline of the both data sets.

4.1. Acted speech data

The Linguistic Data Consortium 2002S28 (Emotional Prosody Speech and Transcripts) database [18] consists of English language acted speech recordings. Eight actors (five females and three males), were provided with descriptions of each emotional context. Flashcards were used in order to display four-syllable dates and numbers in 15 different emotional categories. The following emotion categories are covered: {*neutral*, *hot anger*, *cold anger*, *happy*, *sadness*, *disgust*, *panic*, *anxiety*, *despair*, *elation*, *interest*, *shame*, *boredom*, *pride* and *contempt*}.

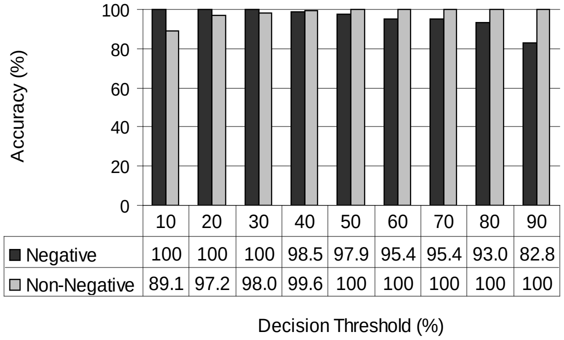
The database [18] consists of 30 data files in sphere format and their transcripts. These data were split, so as to extract the separate speech utterances, belonging to different emotional states, which were named according to the annotation information. For the purpose of our experimentations, all recordings were down-sampled to sampling frequency of 8 kHz. The acted speech data were split in a way of examining the speaker-dependent behavior.

4.2. Real-world speech data

The real-world data consist of recordings collected during interactions of naïve users with the smart-home dialogue system. Specifically, 29 people participated, 15 males and 14 females. The age of the participants ranged from 14 to 37 years old while their mean age was 21 years. None of the participants was an actor, and none of them had previous experience with a smart-home dialogue system.

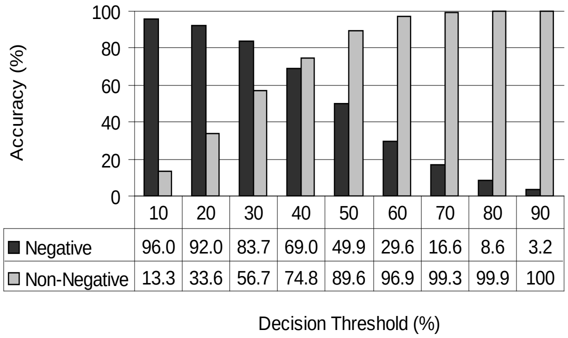
Each participant was recorded in a single session. Before entering the room, where the smart-home system is installed, all participants were asked to read a general instruction concerning the functionality of the smart-home system. The information provided to the participants prior the experiment didn't reveal that the purpose of data collection was to retrieve emotional reactions. Next, the participants filled in questionnaires in which they were asked to provide information for their background profile. Specifically, they were asked to report their experience with voice interaction systems in general, and their expectations towards the dialogue system they were about to interact with.

Afterwards, each participant entered the studio decorated in smart-home living room environment. Once they entered the room, a set of 10 task cards constructing a real world scenario, was provided. These scenarios guided the users which devices they should use but no specific commands or templates were provided. Each participant remained in the room for about 25 minutes trying to complete the tasks as described in the aforementioned task cards. The user had the ability to observe the system's responses to his commands through a simulator installed on a personal computer that animates the operation of the intelligent appliances. No help was provided to the user unless she/he failed completely to complete a given task, after trying more than five times. In addition, no intentional manipulation of the system's response (for provoking emotional reactions) took place during the experiments. Finally, each participant was asked to fill in



	10	20	30	40	50	60	70	80	90
■ Negative	100	100	100	98.5	97.9	95.4	95.4	93.0	82.8
□ Non-Negative	89.1	97.2	98.0	99.6	100	100	100	100	100

Figure 3. Recognition results for acted speech and different decision thresholds.



	10	20	30	40	50	60	70	80	90
■ Negative	96.0	92.0	83.7	69.0	49.9	29.6	16.6	8.6	3.2
□ Non-Negative	13.3	33.6	56.7	74.8	89.6	96.9	99.3	99.9	100

Figure 4. Recognition results for real-world speech and different decision thresholds.

post-questionnaires where they judged different quality aspects and noticed specific difficulties when using the smart-home system. After the end of the whole procedure each participant was asked if she/he agree that their voice can be utilized in research on emotion recognition.

All recordings were annotated and labeled manually in three subsequent stages. Initially, during pre-annotation procedure, all recordings segmented by the speech recognizer (single channel audio, sampling rate 8 kHz, resolution 16 bit) were annotated by one annotator, so as to tag the user's command, emotion and discard silences and noise-corrupted files. In average, there were about 110 utterances (interactions) per user. During the labeling for emotion, six emotion tags were used: {*delighted*, *pleased*, *neutral*, *confused*, *angry* and *hot angry*}.

In the second stage, the pre-processed recordings were annotated by six additional annotators, which were instructed to tag the utterances according to their human intuition. Eventually, each utterance was labeled with specific emotional tag only when at least five of all seven annotators converged to categorization in the same class, i.e. when there was agreement of at least 5 humans.

At the third stage of annotation, the remaining utterances (1179 out of 3269) that did not gain agreement of at least 5 annotators, i.e. utterances which were not labeled at this second stage, were post-processed by a committee of 3 experts. Thus, these 1179 utterances were force-categorized into one of the six classes.

5. Experimentations and results

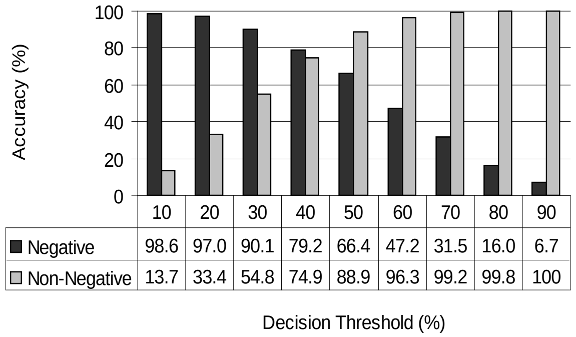
In this section we present results from the evaluation of the DNES component on both the acted speech and the real-world data. Specifically, distinguishing between two emotional categories was of interest: *negative* and *non-negative* speech. For that purpose, in the experiments with the acted speech database, the class labels {*neutral*, *hot anger*}, were utilized. In the experiments with the real-world data, the negative class consisted of utterances labeled as *anger* or *hot anger*, and the non-negative one of the utterances labeled as {*neutral*, *pleased*, *delighted*}. Only 22 out of 29 participants had sufficient number of occurrences of negative speech, which we consider here.

In all experiments the speech data were processed in a uniform manner as described in Section 3. Utilizing the speech feature vectors a GMM consisting of 16 mixture components was built for each class, through the EM algorithm. The training of each GMM was ended when the error between two subsequent iterations decreases with less that 0.1, or when the maximum number of iterations (in our case 20) was exceeded. In all experiments, we used the leave-one-out technique for better utilization of the available data.

Figure 3 illustrates the averaged for all speakers recognition rate of negative against non-negative emotional states for the acted speech data. Each bar in the figure presents the averaged recognition results obtained with respect to different threshold, i.e. for threshold that corresponds to various percentages of frames recognized as negative. Specifically, here the decision threshold varies from 10% to 90%. These different thresholds offer us a raff impression what will be the performance of the DNES component when the threshold is manipulated ether to adapt to the specific needs and preferences of different users, or to adapt to different application scenarios.

As expected, as the decision threshold for characterizing an utterance as negative increases, so the recognition accuracy of detecting negative emotional states decreases. However, the overall performance achieved on the acted speech remains very high even for decision threshold of 80%. This outcome is probably due to the fact that in acted speech the emotions might be overacted [22].

The recognition results on the real-world speech



	10	20	30	40	50	60	70	80	90
■ Negative	98.6	97.0	90.1	79.2	66.4	47.2	31.5	16.0	6.7
□ Non-Negative	13.7	33.4	54.8	74.9	88.9	96.3	99.2	99.8	100

Figure 5. Recognition results for utterances with human confidence at least 5/7.

data are shown in Figure 4. As the figure presents, the accuracy of the DNES component for real-data is much lower, when compared to the one achieved for the acted speech (Figure 3). This drop of performance can be partially explained with the variance in the human behavior and the difficulty to model it [22], but also with the fact that, often, in real speech one utterance might include more than one emotional state.

As mentioned in Section 4, for about 30% of the real-world recordings the seven annotators disagreed on the labeling. To investigate this phenomenon, we performed further experimentations, where only these data for which five or more annotators out of seven agreed. The utterances that were force categorized at the third stage of annotation are not considered in this experiment. The experimental results are illustrated in Figure 5. As it can be seen, the recognition performance is notably higher, which can be explained with the greater confidence of the annotation tags – the emotional state of the speaker was distinguished by 5 or more humans, who represent a qualified majority of the annotators.

In Figure 6, the DET (Detection Error Trade-off) plot curves that correspond to the aforementioned three experiments are presented. Each curve represents one experiment, and is obtained by combining the results obtained for all speakers. Here, the error rates of the DNES component is reported in terms of false alarm probability and miss probability. These DET plots provide threshold-independent (i.e. application independent) evaluation of the models build for the target classes. Under the assumption of speaker-independent threshold, which is common for all speakers, the DET plot allows selection of the working point for the detector of negative emotional states according to the needs of the specific application. (However, we should note here that using a combined DET plot doesn't take into consideration the possibility of having user-specific thresholds.)

As seen in Figure 6, a notable difference in the accuracy of the DNES component can be observed, when comparing the results for the acted speech database and the real-world data. Besides the different nature of the two databases (acted vs. non-acted), we deem that the dissimilarity in the recording setup, SNR conditions, and mobility of speakers are important factors, which also contributed for the significant difference in the recognition performance. Specifically, the real-world recordings consist of spontaneous speech, recorded in living room environment, from moving speakers, speaking from different locations and variable distance, and by using distant microphone array. On the other hand, the acted speech data consist of prompted speech, recorded in studio environment, using high-quality close-talking microphone.

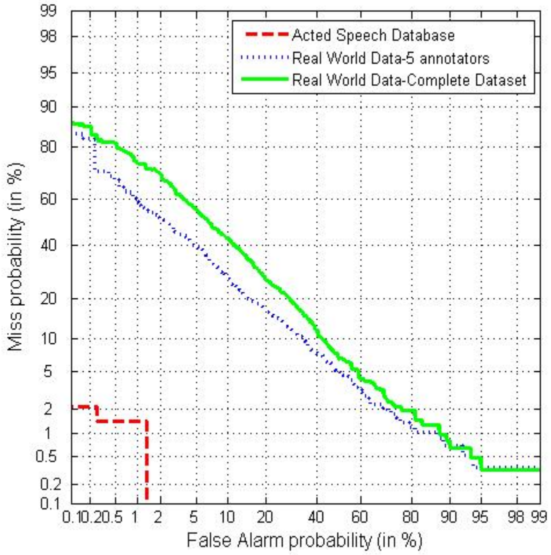


Figure 6. Plot of DET Curves for the evaluation of the detector of negative emotional states

6. Discussion

Due to the specifics of our application, it is important the emotion recognition component has low false alarm error rate. This way, the dialogue flow will not be interrupted unnecessarily, since this would be distractive for the user. On the other hand, the system should reliably detect negative emotional states and provide support to the user when difficulties in communication arise. In order to address these contradicting requirements, we plan to implemented adjustable user-specific decision threshold, which depends on the user's expertise level, needs and preferences. For example, as illustrated in Figure 4, for novice users we can set the decision threshold to 50%. In this way only one out of ten times that the user is not angry, help would be offered by the system, and half of the times that he or she is angry help would be available. On the other hand, for expert users, the decision threshold can be set to 60% so as to minimize potential distraction without need.

The aforementioned concept can be summarized in choosing the appropriate working point, corresponding to the real-world data, on the DET-curve built for each speaker (unlike the DET-plot illustrated in Figure 6, which shows the combined scores for all speakers). Such a speaker-dependent working point can be chosen so as to achieve low miss probability for novice users or low false alarm for experts. A balance between these two target groups can be kept by setting the initial decision threshold to a value around the equal error rate (EER), and letting the users to adjust it on their wish. Another option is the system to perform automatic adjustment of the threshold depending on the expertise level of the user, which usually changes with time.

Adjustment for the speaker-dependent threshold is possible even for unknown to the system speakers, whose expertise level can be judged on the basis of expertise level detected from the percentage of successful speech recognition attempts, the confidence of recognition of commands, speaking style, etc.

7. Summary and conclusion

The present work deals with recognition of negative emotional states for enhancing the spoken dialogue interaction in a smart-home scenario. In order to develop our detector of negative emotional states we initially used a well-understood corpus of acted speech recordings. Later on, taking advantage of the smart home facility deployed at our research laboratory, we collected speech from naïve users. Afterwards, these speech recordings were annotated and utilized in validation experiments that assess the actual performance of our detector of negative emotional states when trained and tested on real-world data. The experimental results reveal significant reduction of the recognition performance for real-world speech when compared to the acted speech. This is explained with the fluent transition from one emotional state to another and the lack of strict borders between the different emotional states in real-world human-machine interaction scenario. An interesting phenomenon observed during the real-world database annotation and validation phases was that humans frequently disagree when labeling specific emotional utterances.

Finally, we would like to emphasize that, although there is room for improvement in recognition performance, our detector of negative emotional states demonstrated encouraging capability to generalize on real-world data, and thus it has the potential to provide the means for improving the user-friendliness of our smart-home dialogue system.

8. Acknowledgement

This work was partially supported by the LOGOS project, which is funded by the General Secretariat for Research and Technology of the Greek Ministry of Development.

9. References

[1] O. Lemon, K. Georgila and M. Stuttle , “An ISU dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the TALK in-car system, EACL (demo session),” 2006.

[2] A. Potamianos, E. Fosler-Lussier, E. Ammicht and M. Peraklakis, “Information seeking spoken dialogue systems Part II: Multimodal Dialogue,” IEEE transactions on multimedia, vol. 9, no. 3, April 2007.

[3] K. Kvale and N. Warakagoda, “A Speech Centric Mobile Multimodal Service useful for Dyslectics and Aphasics,” Proc. Intespeech 2005, pp. 461-464, 2005.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J.G. Taylor, “Emotion recognition in human-computer interaction,” IEEE Signal Processing magazine, vol. 18, no. 1, pp. 32-80, 2001.

[5] Y. Wang and L. Guan, “Recognizing human emotion from audiovisual information,” Proc. ICASSP 2005, pp. 1125-1128, 2005.

[6] P. Pal, A.N. Iyer and R.E. Yantorno, “Emotion detection from infant facial expressions and cries,” Proc. ICASSP 2006, vol. 2, pp. 721-724, 2006.

[7] B. Schuller, R. Müller, M. Lang and G. Rigoll, “Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles,” Proc. Interspeech 2005, pp. 805-808, 2005.

[8] S. Yacoub, S. Simske, X. Lin, and J. Burns, “Recognition of emotions in interactive voice response systems,” Proc. Eurospeech 2003, pp. 729-732, 2003

[9] J. Liscombe, G. Riccardi and D. Hakkai-Tür, “Using context to improve emotion detection in spoken dialog systems,” Proc. Interspeech 2005, pp. 1845-1848, 2005.

[10] L. Devillers and L. Vidrascu, “Real life emotions detection with lexical and paralinguistic cues on human-human call center dialogs,” Proc. Interspeech, pp. 801-804, 2006.

[11] F. Burkhardt, J. Ajmera, R. Englert, J. Stegmann and W. Burleson, “Detecting anger in automated voice portal dialogs,” Proc. Interspeech 2006, pp. 1053-1056, 2006.

[12] H. Ai, D. J. Litman, K. Forbes-Riley, M. Rotaru, J. Tetreault and A. Purandre, “Using system and user performance features to improve emotion detection in spoken tutoring dialogs,” Proc. Interspeech 2006, pp. 797-800, 2006.

[13] R. Nisimura, S. Omae, H. Kawahara and T. Irino, “Analyzing dialogue data for real-world emotional speech classification,” Proc. Interspeech 2006, pp. 1822-1825, 2006.

[14] M. Rotaru and D.J. Litman, “Using word-level features to better predict students emotions during spoken tutoring dialogues,” Proc. Interspeech 2005, pp. 881-884, 2005.

[15] A. Batliner, S. Biersack, and S. Steidl, “The prosody of pet robot directed speech: evidence from children,” Proc. Speech Prosody, pp. 1-4, 2006.

[16] C. M. Lee and S.S. Narayanan, “Towards detecting emotions in spoken dialogs,” IEEE Transactions on Speech and Audio Processing, Vol 13, No. 2, pp. 293-303, 2005.

[17] A. Vovos, B. Kladis and N. Fakotakis, “Speech operated smart-home control system for users with special needs,” Proc. Interspeech 2005, pp. 193-196, 2005.

[18] Linguistic Data Consortium, “Emotional Prosody Speech,” Available: www.ldc.uppen.edu/Catalog/CatalogEntry.jsp?cataloId=LDC2002S28, University of Pennsylvania.

[19] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” Proc. 17 IFA 1993, pp.97-110, 1993.

[20] S.B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on ASSP 28, pp. 357-366, 1980.

[21] I.T. Nabney, Netlab Algorithms for Pattern Recognition, Springer, 2002.

[22] J. Wilting, E. Kramber and M. Swerts, “Real vs. acted emotional speech,” Proc. Interspeech 2006, pp. 805-808, 2006.