A Real-World Emotional Speech Corpus for Modern Greek

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

The present paper deals with the design and the annotation of a Greek real-world emotional speech corpus. The speech data consist of recordings collected during the interaction of naïve users with a smart-home dialogue system. Annotation of the speech data with respect to the uttered command and emotional state was performed. Initial experimentations towards recognizing negative emotional states were performed and the experimental results indicate the range of difficulties when dealing with real-world data.

1. Introduction

The progress of technology and the increasing use of spoken dialogue applications raise the need of user-friendly human-machine interaction (Pantic & Rothkrantz, 2003). Awareness about the emotional state of a user can provide the means for improving the performance of a dialogue system, thus lead to more successful interaction experiences. When dealing with the design of an emotion recognition system, a crucial factor affecting the performance of the emotion recognizer is the corpus used for creating the models representing the user's emotions.

To this end, much work had been reported towards creating emotional corpora from elicited, prompted, acted speech data. For instance, the Emotional Prosody Speech and Transcripts database (LDC2002S28, 2002) 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 Berlin Database of Emotional Speech (Burkhardt et al., 2005) contains about 500 utterances spoken by 10 actors (5 males, 5 females) in the German language. The emotions covered are: happy, angry, anxious, fearful, bored and disgusted as well as neutral. The Danish Emotional speech corpus (Engberg & Hansen, 1996) consists of two male and two female actors reading texts in five emotional states: angry, happy, sad, surprise, and neutral.

On the other hand, the acted speech databases, although supporting the advance of emotion recognition technology, are not applicable for research and development on real-life emotion recognition applications, since spontaneous speech and genuine formulations differ considerably from acted speech (Batliner et al., 2003). Morisson et. al. (2007) compared two emotional speech corpora, natural and acted emotional speech, demonstrating the advantages and disadvantages of both acquisition methods and how these methods affect the end application of vocal emotion recognition. Vidarascu & Devillers (2005) faced the challenge when studying real-life non-basic emotions on speech data collected from a call centre. Earlier research (Batliner et al., 2004), described design and implementation issues, when creating a real-world emotional speech corpus: in a Wizard-of-Oz scenario, German and English children had to instruct Sony's AIBO robot to fulfil specific tasks.

Domain-specific real-world emotional speech corpus is a necessity for the development of gentle real-world applications, such as a smart-home dialogue system, call centres, voice portals, etc. In the present work, we deal with the design and the annotation of speech data collected during the interaction of naïve users with a smart-home dialogue system. The purpose of this database is to support the advance of emotion recognition technology and to assist for adaptation of this technology for real-world environments. Specifically, we plan to use the database in research and development project that aims at speaker-independent emotion recognition for user-friendly interaction.

This remaining of this paper is structured as follows: in Section 2 the primary objective of the present work and the prospective goals are analysed. Section 3 describes the design issues taken into consideration and the methodology followed. In section 4 detailed description of the recordings collected is given. Section 5 refers to the annotation procedure followed. Finally, in Section 6 preliminary evaluation of the emotional speech corpus for speaker dependent and speaker independent emotion recognition is conducted.

2. Objective

Our primary objective in this on-going project is the creation of Greek speech corpus dedicated to non-acted real-world emotional speech. This corpus is intended to serve in studies on speaker-dependent and speaker-independent emotion recognition in the smart-home domain. Our target is collecting speech from more than 100 speakers, which are either naïve or expert users of the smart-home dialogue system. We seek for a balance over ages and genders.

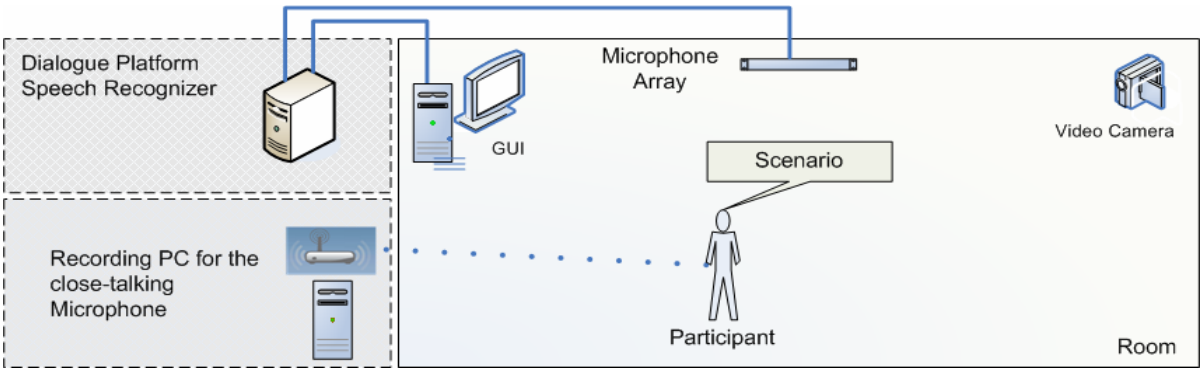

Figure 1: Test site of the smart-home dialogue system.

There are two distinctive phases in this project, each one serving a different goal: Our foremost goal in the first phase, which ended in December 2007, was to collect natural spontaneous speech from at least 40 naive users, who do not have previous experience with smart-home spoken dialogue interaction systems. Such recordings would offer the opportunity of identifying and studying genuine verbal and non-verbal phenomena, and hopefully would provide the basis for creation of an emotion recognition component that would permit the enhancement of the user-friendliness of future (smart-home) dialogue systems.

Our main goal in the on-going second phase of the project, which is planned to last by December 2008, is collecting spontaneous speech from at least 60 additional speakers. In this number we consider at least 30 users, who have previous experience with spoken dialogue systems, and at least 30 naive users. The design of the experimental setup during the second phase targets at evaluating various interaction strategies for a spoken dialogue system that possess an emotion recognition component. Furthermore, we anticipate that the second phase of the project will uncover aspects, which were not captured during the first phase.

3. Design and Methodology

The experimental setup and methodology that we followed correspond to a typical operational scenario of the smart-home dialogue system deployed at our test-site (Vovos et al., 2005). Specifically, the smart-home system provides, via speech interaction, access to information, entertainment devices, and control of intelligent appliances. The intelligent appliances are simulated via graphical interface, which provides animated feedback to the users, and pre-recorded audio. The user interface and the I/O components of the system are placed inside the room where the experiment takes place. All the other components, such as speech recognizer, dialogue platform, etc, are distributed on PCs over the local area network (please refer to Figure 1).

The recruitment of participants is done from students, university staff and external volunteers. Before the beginning of the recording session, each participant was asked to read general instructions concerning the functionalities of the smart-home system. Both before and after the recording session all participants are asked to fill in questionnaires. These questionnaires provide the means for measuring the successfulness of the interaction experience. The pre-questionnaires acquired information about for (i) the personal background, (ii) previous experience with dialogue systems and (iii) the expectations toward voice interaction systems.

Each participant was asked to interact with the dialogue system in one session for approximately 25 minutes. During the session she/he was recorded by a camera[1], a microphone array[2] and a close-talking microphone[3] (each one independently from the others). The speech recognizer was fed from the output of the microphone array. The speech signal, as it was segmented by the speech recognizer, was stored to waveform files with sampling rate of 8 kHz, single channel, resolution 16 bits. The audio signal captured from the close-talking microphone was sampled at 16 kHz and stored into the same format. The camera recordings consist of both audio and video captures, time-synchronized, and saved in the Digital Video format with frame resolution of 640x480 pixels and audio quality of 2079 kbps, 16-bit stereo, and sampling frequency of 44.1 kHz.

A set of 10 task cards, constructing a real-world scenario, was provided to each participant (please refer to Table 1). The experimental setup elicited the participant to use several appliances or access specific information services.

Task Card	Elicited Action
1	turn on the lights, turn on the radio
2	watch a movie, watch an episode
3	open the door
4	open the blinds
5	turn on TV, change channel, select channel
6	turn off TV, turn off the light
7	close the door
8	turn on the light
9	close the blinds
10	turn on the DVD, turn off the DVD

Table 1: Summary of the contents of the 10 task cards and elicited actions

[1] SONY DCR-VX1000E
[2] ACOUSTIC MAGIC Voice TrackerTM Array Microphone
[3] AKG UHFPT40 (863.100 MHz)

Age	Males	Females	Total	Current (%)	Target (%)
<18	1	2	3	6.98	>10
18-30	20	15	35	81.39	>30
31-45	1	3	4	9.30	>20
>45	1	0	1	2.33	>10
Total	23	19	43	100	

Table 2: Distribution of participants over age groups and genders.

However, in order to provoke the participants to use spontaneous speech, we did not provide any specific commands, examples or scenario-specific templates how the home appliances should be addressed. The participants were instructed to use natural language.

The participant was allowed to move freely inside the room. No help was provided to the participant, unless she/he unsuccessfully tried, more than five times, to complete a given task. No manipulation of the dialogue flow took place, for provoking emotional reactions.

Finally, the participant was asked to fill in post-questionnaires, where she/he was asked once more about her/his expectations towards voice interaction systems. In addition she/he had to judge different quality aspects and notice specific difficulties faced when using the smart-home system. No part of the aforementioned documents revealed that the purpose of the experiment was to capture emotional reactions. For minors, i.e. those participants whose age was lower than 18 years, there was trustee adult relative attending both the recording session and filling of the questionnaires. After the end of the recording procedure and the questionnaires each participant was asked if she/he agrees that her/his voice can be utilized anonymously in research on emotion recognition.

4. Recordings

The collected data consists of recordings from 43 people, among who are 20 females and 23 males. The age and gender distributions are shown in Table 2. The last column indicates our target on the recruitment of the participants. The age of the participants varies from 12 to 56 years old and their mean age is 22 years.

The number of the participants with respect to their region of origin and childhood is shown in Table 3. Twenty-five out of 43 participants (58.14%) had no previous experience with spoken dialogue applications, while the remaining 18 participants (41.86%) declared experience on using voice interaction systems for hospital telephone centres, bus ticket booking and phone banking applications.

Region	# of participants	Percentage (%)
Athens	13	30.23
Patras	10	23.26
Cyprus	8	18.60
Other	12	27.91

Table 3: Number of the participants with respect to their region of origin and childhood.

For each participant, one session equals to approximately 100 utterances (not counting the ones containing only silence). The total duration of the recordings is approximately 18 hours, which corresponds to 3.5 hours of speech.

5. Annotation

The annotation procedure consists of three subsequent steps:

Step 1: In the first step, one annotator labels each utterance segmented by the speech recognizer in one of the following emotional states: {*delighted*, *pleased*, *neutral*, *confused*, *angry* and *hot angry*}, based on his human intuition. These states were obtained by inspection of the data. This annotator also tags the issued command. Two special tags were used for: (i) labeling utterances where the participant does not issue a command and (ii) labelling files with corrupted speech or silence only.

Step 2: In the second step, six annotators, independently from each other, are asked to tag the perceived emotional state of each utterance in one of the aforementioned classes.

The annotation of the 43 participants resulted to 4413 utterances. Similarly to (Batliner et al., 2004), having the the tags from the seven labelers, a decision can be made by applying a threshold on the number of annotators that agree. Table 4 shows the number of utterances for which a decision cannot be made when setting this threshold to different number.

Threshold	Number of utterances
4	446
5	1497
6	2529
7	3673

Table 4: Number of utterances for which a decision cannot be made with respect to the threshold on the number of annotators that agree.

A widely-used measure of the inter-annotation agreement is the Kappa value (Fleiss, 1971). For our specific case of six emotional categories, we calculated Kappa = 0.31. This value corresponds to a fair agreement among the annotators.

Step 3: For the specific case of threshold 5, i.e. when we are looking at agreement among 5 annotators, there were 1497 utterances for which decision was not made. To resolve the disagreement, these utterances were

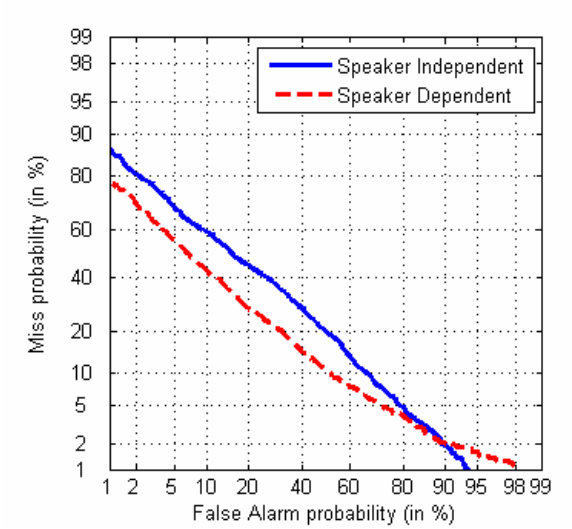

Figure 2: DET plots for the detector of negative emotional states for speaker dependent (dashed line) and speaker independent emotion recognition

post-processed by a committee of three expert-labellers, who force-categorized them into one of the six classes. The resulting numbers of utterances per category are: {*delighted* (2), *pleased* (82), *neutral* (3240), *confused* (259), *angry* (697), *hot angry* (133)}.

6. Experimental Results

In order to get impression about the collected database, we performed two experiments: one on speaker-dependent and one on speaker-independent emotion recognition. For the purpose of these experimentations we utilized the subset of data obtained for agreement among the 5 annotators. The 1497 sentences that were force categorized at Step 3 were included in the dataset.

Specifically, in the present experimental setup, we considered distinguishing two emotional categories: negative vs. non-negative speech. The negative class consists of utterances that indicate problems in the communication, i.e. utterances labelled as {*confused*, *anger* and *hot anger*}. The non-negative category indicates the "no-problem" condition and contained these utterances labelled as {*delighted*, *pleased* or *neutral*}.

A set of commonly used speech features (12 MFCCs, i.e. excluding the 0th coefficient, energy, harmonicity, fundamental frequency) and diagonal covariance Gaussian mixture models (GMM) (Reynolds & Rose, 1995) were employed.

In the speaker-dependent experiment Gaussian mixture models with 16 mixture components were built. This number of mixtures was observed to achieve the highest recognition accuracy. In order to utilize better the available data, we employed the leave-one-utterance-out technique, which resulted in 4413 experiments. The average equal error rate (EER) obtained in these experiments is 24.6%.

In the speaker-independent experiment, on the same dataset, we employed the leave-one-speaker-out technique. The same feature vector and classifier were used but the GMM had 256 mixture components to account for the larger amount of data per category. This resulted in 43 experiments, in which we obtained an average EER of 33.4%.

DET (Detection Error Trade-off) plots for these experimentations are presented in Figure 2. The aforementioned results indicate the difficulties when dealing with real-world data, and corroborate previous results reported in the literature (Batliner et al., 2006).

We consider the presented experimental results as the baseline performance for this dataset. We deem that this performance can be improved by employing more sophisticated modelling techniques, or by elaborating more advanced speech parameterization.

7. Conclusion

We reported recent efforts towards creation of real-world emotional speech database for Modern Greek language. This database was collected in support of research and development of speaker-independent emotion recognition technology in real-world environments.

Furthermore, we presented experimental results, obtained from the collected data, that can be considered as baseline performance, and that can be improved in future research activities.

We presume that large-scale emotional speech corpora would assist for proper modelling of emotional states in the speaker-independent emotion recognition tasks.

Eventually, the speech corpus will be made available for research purposes (AIG, 2008).

8. Acknowledgements

This work is partially supported by the PlayMancer project (FP7-ICT-215839-2007), which is funded by the European Commission.

9. References

AIG, (2008). Artificial Intelligence Group, Wire Communication Laboratory, University of Patras, Available at: http://www.wcl.ee.upatras.gr/ai/Research /SEmo.htm

Batliner, A., Fisher, K., Huber, R., Spilker, J., Nöth, E. (2003). How to find trouble in communication. In *Speech Communication*, vol. 40, pp. 117--143.

Batliner, A., Burkhardt, F., van Ballegooy, M., Nöth, E. (2006). A taxonomy of applications that utilize emotional awareness. In *Erjavec, T. and Gros, J.* (Ed.), Language Technologies, IS-LTC 2006, pp. 246--250.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russel, M., Wong, M. (2004) "You stupid tin box" - children interacting with the Aibo robot: a cross-linguistic emotional speech corpus. In Proc. of the LREC 2004, pp. 171--174.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B. (2005). A Database of German Emotional Speech. In *Proc. of the Interspeech 2005*, Lisbon, Portugal, pp. 1517--1520.

Engberg, I.S., Hansen, A.V. (1996). Documentation of the

Danish Emotional Speech Database DES, Aalborg

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. In *Psychological Bulletin* 76 (5): 378--382.

LDC2002S28, (2002). Linguistic Data Consortium, "Emotional Prosody Speech," University of Pennsylvania. Available: www.ldc.uppen.edu/Catalog/CatalogEntry.jsp?cataloId=LDC2002S28.

Morrison, D., Wang, R., De Silva, L.C. (2007). Ensemble methods for spoken emotion recognition in call-centres. In *Speech Communication*, vol. 49, pp. 98--112.

Pantic, M., Rothkrantz, L. (2003). Toward an affect-sensitive multimodal human-computer interaction. In *Proc. of the IEEE*, vol. 91, pp. 1370--1390.

Reynolds, D., Rose, R. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. In *IEEE Trans. on Speech and Audio Processing*, vol.3, no.1, pp. 72--83.

Vidrascu, L., Devillers, L. (2005). Real-Life Emotion Representation and Detection in Call Centers Data. In Proc. of Affective Computing and Intelligent Interaction, pp. 739--746.

Vovos, A., Kladis, B., Fakotakis, N. (2005). Speech operated smart-home control system for users with special needs. In *Proc. of the Interspeech 2005*, pp. 193--196.