The Effect of Emotional Speech on a Smart-Home Application

Theodoros Kostoulas, Iosif Mporas, Todor Ganchev, Nikos Fakotakis

Artificial Intelligence Group, Wire Communications Laboratory,
Electrical and Computer Engineering Department,
University of Patras, 26500 Rion-Patras, Greece
{tkost, imporas, tganchev, fakotaki}@wcl.ee.upatras.gr

Abstract. The present work studies the effect of emotional speech on a smart-home application. Specifically, we evaluate the recognition performance of the automatic speech recognition component of a smart-home dialogue system for various categories of emotional speech. The experimental results reveal that word recognition rate for emotional speech varies significantly across different emotion categories.

Keywords: speech recognition, emotional speech, dialogue systems.

1 Introduction

The use of spoken dialogue applications has been increased over the last decade. It is common that services such as info-kiosks, telephone centers and smart-home applications are supported by multimodal dialogue systems. The challenge in technology is to provide user-friendly human-machine interaction, which guarantees robust performance in variety of environmental conditions or behavioral styles.

To this end, a number of spoken dialogue systems have been reported. Most of them focus on designing various techniques concerning the dialogue flow between the user and the system. In [1] an adaptive mixed initiative spoken dialogue system (MIMIC) that provides movie show-time information is described. MIMIC has been implemented as a general framework for information query systems, by decoupling its initiative module from the goal selection process, while allowing the outcome of both processes to jointly determine the response strategies employed. In [2] MiPad, a wireless personal digital assistant is presented. It fully integrates continuous speech recognition and spoken language understanding. In [3] a multimodal application's architecture is described. This architecture combines finite-state multimodal language processing, a speech-act based multimodal dialogue manager, dynamic multimodal output generation and user-tailored text planning, so as to enable rapid prototyping of multimodal interfaces with flexible input and adaptive output. The application provides a mobile multimodal speech-pen interface to restaurant and subway information. In [4] a multimodal dialogue system using reinforcement learning for in-car scenarios is demonstrated. The baseline system is built around the DIPPER

dialogue manager [5]. The DIPPER dialogue manager is initially used to conduct information-seeking dialogues with a user, such as finding a particular hotel and restaurant, using hand-coded dialogue strategies. The task and user interface modules of a multimodal dialogue system development platform are presented in [6]. The system is evaluated for a travel reservation task.

The aforementioned approaches try to meet the needs for successful interaction experiences through: *(a)* Keeping track of the overall interaction of the user with the dialogue system, with a view to ensuring steady process towards task completion, and *(b)* proper management of mixed initiative interaction. The basic component that determines the performance of a dialogue system is the automatic speech recognition (ASR) component. In [7] Rotaru et al. examined dependencies between speech recognition problems and users' emotions. Moreover, previous ASR experiments prove that the user's emotional state affects significantly the recognition performance. In [8] Steeneken and Hansen performed experiments on speech data characterized by different kind of difficulties. The results show the strong effect of emotions on the performance of the speech recognizer. In [9] Polzin and Waibel demonstrated that automatic speech recognition accuracy varies significantly depending on the emotional state of the speaker. Their research focused on performing experimentations on five major emotion categories {*neutral, happiness, sadness, fear, anger*}. Further experiments, realizing emotion-specific modeling, improved the word accuracy of the speech recognition system, when faced with emotional speech. Athanaselis et al. addressed the task of recognizing emotional speech by using a language model based on increased representation of emotional utterances [10].

In the present work, we study the effect of emotional speech recognition on smart-home application. Specifically, we examine the variation in the speech recognition performance for a variety of emotion categories. The present research is important for providing estimation about the potential gain of ASR performance that can be obtained, if an emotion recognition component is incorporated into our application.

2 The spoken dialogue system of the LOGOS project

The scope of the LOGOS project is to research and develop a smart-home automation system that offers user-friendly access to information, entertainment devices and provides the means for controlling intelligent appliances installed in a smart-home environment. Fig. 1 illustrates the overall architecture of the LOGOS system, and the interface to various appliances, such as high definition television (HDTV), DVD player, mobile phone, etc. The multimodal user interface of that system allows the home devices and appliances to be controlled via the usual infrared remote control device, PC keyboard, or spoken language. In the present study, we focus on the speech interface and the spoken dialogue interaction.

Fig.1 presents in detail the architecture of the spoken dialogue interaction sub-system, which functions as follows: The speech is captured by a microphone array, which tracks the position of the user as s/he moves. Next, the preprocessed speech signal is fed to the speech and speaker recognition components that identify the command and the speaker, respectively. The dialogue manager generates feedback to

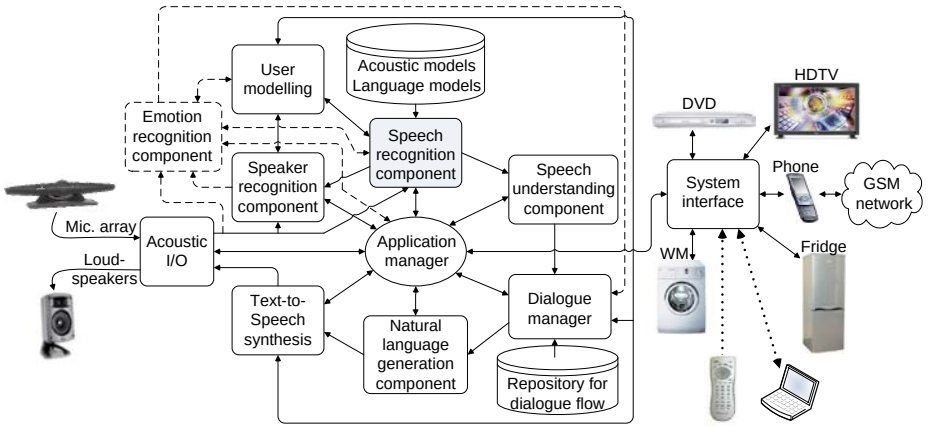

Fig. 1. The LOGOS system.

the user, according to the interpretation of commands received from the speech understanding component, the device status and other information. This feedback is delivered via the text-to-speech component, which is fed by the natural language generation component. All components of the spoken dialogue system are deployed on a personal computer, and the communication with the intelligent appliances is performed through a set-top box, which provides the interface to the intelligent appliances.

The smart-home system is designed with open architecture, which allows new components to be added for extending the system's functionality. One such addition could be an emotion recognition component (in Fig. 1 it is shown with dashed-line box), which could contribute to significant improvement of the user-friendliness of the system, as well as of its overall performance. By utilizing the feedback from this component, which ideally would capture the emotional state of the users, the system will be able to select an emotion-specific acoustic model, for improving the speech recognition performance, or to steer the dialogue flow accordingly.

3 Experiments and results

For examining the operational performance of the smart-home dialogue system under the effect of the emotional state of the user the open source Sphinx III [11] speech recognition engine was employed. A general purpose acoustic model trained over the Wall Street Journal database [12] was utilized. Speech data, consisting of recordings of 16 kHz and resolution of 16-bit, were frame blocked from a sliding window of 512 samples and further parameterized. For the parameterization, the 13 first cepstral coefficients were computed from 40 Mel filters covering the frequency range [130, 6800] Hz, as well as their first and second derivative coefficients. The acoustic model consists of 3-state left-to-right HMM phone models, with each state modeled by a mixture of 8 continuous Gaussian distributions. The number of tied states was set to 4000. No automatic gain control or variance normalization was applied.

In order to capture the variation of the recognition performance across the numerous emotional speaking styles, we used, as test data, the Emotional Prosody Speech and Transcripts database [13]. It consists of recordings of professional actors reading series of semantically neutral utterances (dates and numbers) spanning fourteen distinct emotional categories. The full set of emotional categories are *anxiety*, *boredom*, *cold anger*, *contempt*, *despair*, *disgust*, *elation*, *happy*, *hot anger*, *interest*, *panic*, *pride*, *sadness*, *shame*, and *neutral*. The recordings were down-sampled to 16 kHz, 16-bit, single-channel format, following the specifications of the acoustic model. Finally, each speech file of the database was segmented to smaller ones, resulting one single emotional state per file. The segmentation was accomplished according to the annotation files provided with the database.

During recognition process a unigram language model, consisting of the list of words that appear in the Emotional Prosody Speech and Transcripts database, was constructed. Each single-emotion speech file was pre-processed and parameterized identically to the acoustic model training data. The decoder's language weight was set to 9.5. The rest of the decoder's parameters held their default values. We used uniform setup for all the emotional classes.

The speech recognition results for the fourteenth emotional classes as well as for neutral (conversational speaking style) are shown in Table 1. The first column provides the number of words that existed in the test set for each emotion. The second column shows the number of words that were not recognized, including word substitutions, insertions and deletions. The last column provides the speech recognition performance in terms of percentage of word error rate (WER).

Table 1. Speech recognition results for different emotional categories.

Emotion	# Words	# Errors	WER (%)
Neutral	1787	111	6.21
Shame	736	69	9.38
Interest	818	86	10.51
Boredom	853	95	11.14
Pride	722	91	12.60
Sadness	766	102	13.32
Despair	875	136	15.54
Anxiety	884	141	15.95
Cold Anger	746	129	17.29
Happy	830	170	20.48
Disgust	850	176	20.71
Contempt	796	188	23.62
Panic	673	237	35.22
Elation	747	271	36.28
Hot Anger	347	153	44.09

As Table 1 presents, emotional speaking affects significantly the speech recognition performance. Neutral speaking style presents the highest performance, which was expected since the utilized acoustic model was trained purely with neutral speech. In fact, the ASR performance depends on a number of speech characteristics, such as: voice quality, speaking rate, manner of articulation, intensity, pitch range, and these differ among emotional and neutral speech. Emotional states, such as

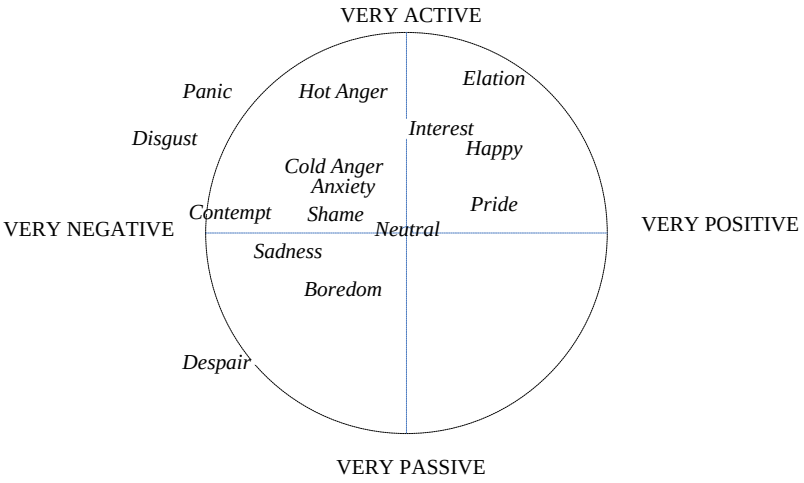

Fig. 2. Emotion categories used, mapped in the activation-evaluation space.

happiness, panic and hot anger, are characterized by: high speaking rate, breathy voice quality, tense articulation, high intensity and wide pitch range [14]. In addition, physiological changes affect the air flow propagation and impose using non-linear speech features to model the corresponding emotional speech [15]. Thus, it is logical such emotion categories to present significantly higher WER than the one reported for the neutral utterances.

On the other hand, emotion categories, such as sadness and shame, present lower increase of the WER. Previous studies conducted on emotion recognition from speech had shown that these emotions differ less from neutral speech, since they present low speaking rate, regular voice quality and slurring articulation [14]. As observed in Fig. 2 the WER results conform to the position of the aforementioned emotion categories in the activation-evaluation space [16]. As more active the emotional speech is, either positive or negative, so the WER increases.

4 Conclusions

In this paper we studied the performance of a speech recognizer for emotional speech. The experimental results show significant variation in the WER reported across different emotion categories. The results suggest the potential gain of ASR performance which can be achieved when reliable emotion recognition component is incorporated into the dialogue system. We deem that emotion-specific or emotion-adapted acoustic models for the ASR would improve the recognition performance of the verbal content of emotional speech [9, 10], thus enhance human-computer interaction.

Acknowledgments. This work was partially supported by the LOGOS project (*A general architecture for speech recognition and user friendly dialogue interaction for*

advanced commercial applications), which is funded by the General Secretariat for Research and Technology of the Greek Ministry of Development.

References

1. Chu-Carroll J.: MIMIC: An adaptive mixed initiative spoken dialogue system for information queries. In: Proc. of the 6th ACL Conference on Applied Natural Language Processing, Seattle, WA (2000)
2. Huang X., Acero A., Chelba C., Deng L., Duchene D., Goodman J., Hon H.-W., Jacoby D., Jiang L., Loynd R., Mahajan M., Mau P., Meredith S., Mughal S., Neto S., Plumpe M., Wand K., Wang Y.: MIPAD: A next generation PDA prototype. In: Proc. Int. Conf. Speech Language Processing, Beijing, China (2000)
3. Johnston M., Bangalore S., Vasireddy G., Stent A., Ehlen P., Walker M., Whittaker S., Maloor P.: MATCH: An architecture for multimodal dialogue systems. In: Proc. Annu. Meeting of the Association for Computational Linguistics (2002)
4. Lemon O., Georgila K., Stuttle M.: An ISU dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the TALK in-car system. In: EACL (demo session) (2006)
5. Bos J., Klein E., Lemon O., Oka T.: DIPPER: Description and Formalisation of an Information-State Update Dialogue System Architecture. In: 4th SIGdial Workshop on Discourse and Dialogue, Sapporo (2003)
6. Potamianos A., Fosler-Lussier E., Ammicht E., Peraklakis M.: Information seeking spoken dialogue systems Part II: Multimodal Dialogue. IEEE Transactions on Multimedia, vol. 9, no. 3 (2007)
7. Rotaru M., Litman D.J., Forbes-Riley K.: Interactions between Speech Recognition Problems and User Emotions. In: Proc. Interspeech 2005, pp. 2481--2484 (2005)
8. Steeneken, H.J.M., & Hansen, J.H.L.: Speech under stress conditions: Overview of the effect of speech production and on system performance. In: International conference on acoustics, speech, and signal processing, 4, pp. 2079--2082 (1999)
9. Polzin S.T., Waibel A.: Pronunciation variations in emotional speech. In: H. Strik, J. M. Kessens, & M. Wester, Modeling pronunciation variation for automatic speech recognition. Proceedings of the ESCA Workshop, pp. 103--108 (1998)
10. Athanaselis T., Bakamidis S., Dologlou I., Cowie R., Douglas-Cowie E., Coxb C.: ASR for emotional speech: Clarifying the issues and enhancing performance. Neural Networks: 18, pp. 437--444 (2005)
11. Lee, K.-F., Hon, H.-W., Reddy, R.: An overview of the SPHINX speech recognition system. IEEE Transactions on Acoustics, Speech and Signal processing, vol. 38, no.1, pp. 35--45 (1990)
12. Paul D., Baker J.: The design of the wall street journal-based CSR corpus. In: Proceedings of ARPA Speech and Natural Language Workshop. ARPA, pp. 357--362 (1992)
13. University of Pennsylvania, Linguistic Data Consortium, Emotional Prosody Speech, http://www.ldc.uppen.edu/Catalog/CatalogEntry.jsp?cataloId=LDC2002S28
14. Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W., Taylor J.G.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, vol. 18, no. 1, pp. 32--80 (2001)
15. Guojun Z., Hansen J.H.L., Kaiser J.F.: Nonlinear feature based classification of speech under stress. IEEE Transactions on Speech and Audio Processing, vol. 9, pp. 201--216 (2001)
16. Whissell, C.: The dictionary of Affect in Language. In: R. Plutchik & H Kellerman ed. Emotion: Theory, research and experience. vol. 4, Academic Press, New York, (1989)