Multimodal interaction with emotional feedback

Francesco Cutugno[1], Antonio Origlia[1], and Roberto Rinaldi[1]

LUSI-lab, Department of Physics, University of Naples "Federico II", Naples, Italy
cutugno@unina.it antonio.origlia@unina.it rober.rinaldi@studenti.unina.it

Abstract. In this paper we extend a multimodal framework based on speech and gestures to include emotional information by means of anger detection. In recent years multimodal interaction has become of great interest thanks to the increasing availability of mobile devices allowing a number of different interaction modalities. Taking *intelligent* decisions is a complex task for automated systems as multimodality requires procedures to integrate different events to be interpreted as a single intention of the user and it must take into account that different kinds of information could come from a single channel as in the case of speech, which conveys a user's intentions using syntax and prosody both.

1 Introduction

Multimodal interaction involving speech aims at providing a more natural interactive experience to the users of automated systems. While this is indeed an important and ambitious goal, it introduces a number of error sources caused by the potential ambiguities that can be found in natural language and the performance of Automatic Speech Recognition (ASR). This has introduced the need for an automated system to perform some kind of self-monitoring to evaluate its own performance, detect erroneous task selection and avoid committing the same error two times in the same interactive session. This field has been deeply explored by researchers dealing with Interactive Voice Response (IVR) platforms, where speech is the only source of information the systems can use in order to select which one of the services it offers the user is looking to obtain and in order to collect the information needed to complete the requested task. While semantic content extraction is obviously the main cue to perform task selection and data collection, paralinguistic information has been used in IVR systems to perform self-evaluation and support interaction with the user [1,15]. These early systems were trained on acted emotions while recent research is now concentrating on spontaneoud emotions: in [4] a real life anger detector trained on data collected from a voice portal was presented while in [14] the problem of multilingual anger detection was explored using recordings from an IVR system.

2 State of the art

Multimodal interface systems were introduced for the first time in the system presented in [3], where graphical objects were created and moved on a screen

using voice recognition and finger pointing. In [5] a set of theoretical guidelines were defined that were named CARE Proprieties (Complementary, Assignment, Redundancy, Equivalence). These properties establish which modes of interaction between users and systems can be implemented and, at the same time, help to formalize relationships among different modalities. The increasing amount of research and practical applications of multimodal interaction systems recently led to the definition of the Synchronized Multimodal User Interaction Modeling Language (SMUIML) [7]: a formal way of representing multimodal interactions. While the possibilities of implementing multimodal information access systems has been explored since when mobile phones started to offer internet based services [16], with the widespread adoption of touch screens on mobile devices, mobile broad band and fast speech recognition, interfaces supporting truly multimodal commands are now available to everyday users. An example is the *Speak4it* local search application [8], where users can use multimodal commands combining speech and gestures to issue mobile search queries. The great interest risen from the possibilities offered by this kind of systems, not only in a mobile environment, soon highlighted the need of formalizing the requirements an automated interactive systems needs to fulfill to be considered *multimodal*. This problem was addressed by the W3C, which has established a set of requirements, concerning both interaction design [11] and system architecture [2], formalized as proprieties and theoretical standards multimodal architectures

Concerning the use of anger detectors in IVRs, in previous studies [13,15] systems have been usually trained on acted emotions corpora before being deployed on IVR platforms. An exception to this trend is represented by [10], in which a corpus of telephone calls collected from a troubleshooting call-center database was used. In that study, the impact of emotions was shown to be minimal with respect to the use of log-files as the authors observed a uniform distribution of negative emotions over successful and unsuccessful calls. This, however, may be a characteristic of the employed corpus, in which people having problems with a High Speed Internet Provider were calling, and is therefore significantly different from the situation our system deals with, as our target consists of users of a bus stops information service.

3 System architecture

In this paper we extend a pre-existing multimodal framework, running on Android OS, based on speech and gesture to include emotional information by means of a user emotional attitude detector. We merge these concepts in a case study previously presented in [6], in which a querying system for bus stops in the city of Naples was implemented. Users can query the system by speaking and drawing on the touch screen producing requests for bus stops in a given area on the map. In a typical use case the user asks: *"Please show me the bus stops of C6 line in this area"* drawing a circle on a map on the screen while speaking.

The user can draw lines and circles on a map aiming at selecting a precise geographic area of interest concerning public transportation. In addition the user

can hold her finger for some second on a precise point on the map in order to select a small rectangular default area on the map with the same purposes. At the same time, speech integrates the touch gesture to complete the command. This way, users can ask for a particular bus line or timetable (using speech) in a given geographic area (using touch), as shown in Figure 1.

For details concerning the general architecture, we refer the reader to [6]. In the present system, the audio signal is considered as twin input: the first one connected to the linguistic content itself obtained by means of an ASR process and a subsequent string parsing process that generates a *Command* table structurally incomplete as more data are needed in correspondence with the missing geographical data completing the user request; the latter goes to an emotional attitude classifier (details will be presented in the next section) returning the anger level characterizing the utterance produced by the user.

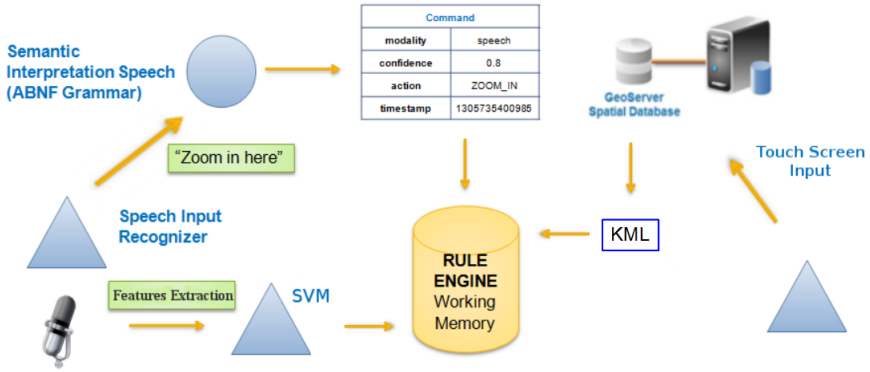
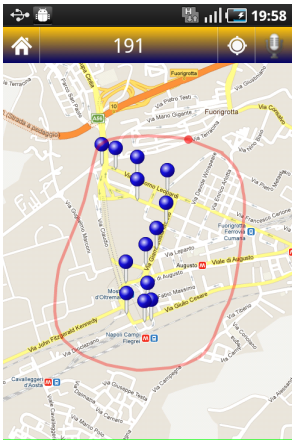


Fig. 1. System architecture: interaction example

The semantic interpreter collects the inputs from parsed ASR and from touch/geographical modules and attempts an answer using the freely available Drools (http://www.jboss.org/drools) rule engine while anger detection is used to launch backup strategies if the transaction does not succeed and the user is unsatisfied by the service as shown in Figure 2.



(a) Places of interest found by combining speech and gestures

(b) Backup strategy for unrecognized commands with angry users

Fig. 2. Screenshots of a multimodal interactive system on a mobile device. In 2a an example of a combined speech and gesture based interaction. Given the utterance "Tell me about the stops of the 191 line *# id_speech_xyz - here - #"* and the co-occurring gesture command *# id_gesture_xyz - drawCircle #*, points corresponding to bus stops of the 191 line are drawn in the area of interest. In 2b a popup menu is used as backup strategy when the transaction fails and the user is getting angry.

4 Emotion recognition module

Automatic emotion recognition is a research topic that has been gaining attention in the last years because of the additional information it brings into automatic systems about the users' state of mind. While there are a number of applications and representations of emotions in the literature, one that has found application in IVR systems is anger detection. Capturing a negative state of the speaker during the interaction is an information that has been exploited in the past, for example, in automated call centers to forward the call to a human agent. Anger detection is usually based on the response given by an automatic classifier on the basis of acoustic features extracted from a received utterance. Features extraction and classification methods for emotions are active research areas: in this work, we use a syllable-based features extraction method and a Support Vector Machine (SVM) to perform the automatic classification of an utterance into two classes: Neutral and Angry. The anger detection module is trained on a subpart of the €motion corpus [9] containing 400 angry and neutral speech recordings in Italian, German, French and English.

First, the recorded utterance is segmented into syllables. This is done by applying the automatic segmentation algorithm presented in [12]. Next, data are extracted from syllable nuclei, estimated by the -3db band of the energy peak associated with each automatically detected syllable. Syllable nuclei, being

stable spectral areas containing vowel sounds, contain more reliable information regarding the distribution of the energy among the frequencies as it was intended by the speaker. Specific spectral measurements like the spectral centroid, moreover, do make sense inside syllable nuclei only. To improve the reliability of the extracted measures, only syllable nuclei at least 80ms long were analyzed. An example of automatic syllable nuclei detection is shown in Figure 3.

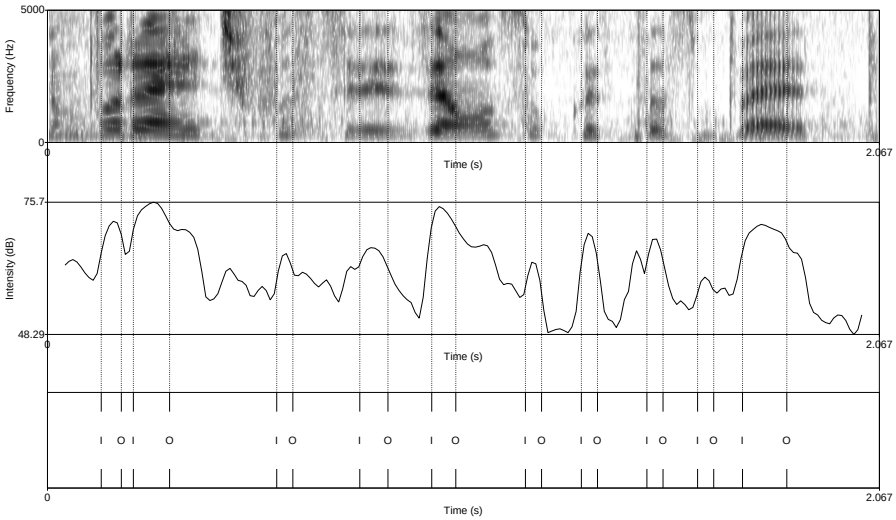


Fig. 3. Automatic detection of syllable nuclei. On the first level, the spectrogram of a speech utterance is shown. On the second one, the energy profile is reported while on the third one automatically detected syllable nuclei incipits (I) and offsets (O) are shown. Voiced areas of spectral stability are used to extract cleaner features.

From each nucleus we extract the following features: *mean pitch* (perceived fundamental frequency), *spectral centroid* (mean of the frequencies in the spectrum weighted by their magnitude) and *energy*.

To produce the final features set, global statistics were computed over the feature vectors extracted from each syllable. Mean and standard deviation were included for each feature while the maximum value was introduced for energy only. An SVM was trained and tested on the features extracted from the €motion corpus. The F-measure obtained in a 10-fold cross validation test was 90.5%.

5 Discussion

The proposed system is presently still under development so its usability has not yet been completely assessed. The multimodal interaction front-end presented in [6], here integrated with the anger detection module, will be tested in the next

future in order to validate both the accuracy of the approach in real conditions of use and the user acceptability and satisfaction. This will be done by means of both an objective and a subjective analysis. The former evaluation will be based on a background software module able to producing log-files containing all the details of the interaction session (time of interaction, number of touches on the pad, length of the speech utterance, etc.), in an evaluation release of the application the user will be requested a-posteriori to verify:

- if the ASR worked properly;
- if the request was correctly recognized and executed.

The analysis of the data collected in this way will be put in relation with those coming from a subjective investigation based on a questionnaire proposed to a further set of users (different from those involved in the former analysis) in order to estimate the subjective acceptability and the degree of satisfaction for the proposed application.

For what it concerns the data on which the Support Vector Machine classifier is trained, while we are currently using a corpus of acted emotions, we plan to use the recordings coming from the tests the system will undergo. We expect this will improve performance as the system will be retrained to work in final deployment conditions. The classifier will therefore be adapted to real-life conditions both in terms on spontaneous emotional display and in terms of recording environment as new recordings will include telephonic microphones quality and background noise.

Differently from what stated in [10], where the telephonic domain and the nature of the interaction did not encourage the introduction of an anger detection system in order to reduce the amount of hang-ups during dialogues, we believe that the mobile device domain will take advantage by the addition of an emotional state recognizer. In the case of apps for mobile devices requirements are different from those observed during telephonic dialogues and, provided that the Human-Computer Interface is well designed and correctly engineered, it is not really expected that the user closes the app before obtaining the required service. In this view, anger detection must be seen as a further effort made by the designer to convince users not to give up and close the app before reaching their goals.

6 Conclusions

We have presented a framework to design and implement multimodal interfaces with relatively little effort. As far as we know, anger detection and, in general, emotional feedback has not been taken into account in mobile applications before. The case study we presented shows a mobile application integrating speech recognition, anger detection and gesture analysis to implement a bus stops querying system. A basic release of the presented system, without speech and multimodal system is presently available on the Google Market (https://play.google.com/store/apps/details?id=it.unina.lab.citybusnapoli) and

received excellent user reviews and more than 2600 downloads (April 2012), we consider this as a very effective usability test. Multimodal without emotive feedback is also being tested for usability by means of a subjective procedure, we are now undergoing formal testing of the complete system in order to verify its usability and its stability.

Acknowledgements

We would like to thank Vincenzo Galatà for providing the speech recordings from the yet unpublished multilingual emotional speech corpus €motion we used in our experiments. We would also like to thank Antonio Caso for assisting during the extension of the original framework to include the emotional module.

References

1. Ang, J., Dhillon, R., Krupski, A., Shriberg, E.and Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proc. of of ICSLP. pp. 2037–2040 (2002)
2. Bodell, M., Dahl, D., Kliche, I., Larson, J., Tumuluri, R., Yudkowsky, M., Selvaraj, M., Porter, B., Raggett, D., Raman, T., Wahbe, A.: Multimodal architectures and interfaces (2011), http://www.w3.org/TR/mmi-arch/
3. Bolt, R.A.: "Put-that-there": Voice and gesture at the graphics interface. SIGGRAPH Comput. Graph. 14(3), 262–270 (1980)
4. Burkhardt, F., Polzehl, T., Stegmann, J., Metze, F., Huber, R.: Detecting real life anger. In: Proc. of ICASSP. pp. 4761–4764 (2009)
5. Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J., Young, R.M.: Four easy pieces for assessing the usability of multimodal interaction: the care properties. In: Proc. of INTERACT. pp. 115–120 (1995)
6. Cutugno, F., Leano, V.A., Mignini, G., Rinaldi, R.: Multimodal framework for mobile interaction. In: Proc. of AVI. pp. 197–203 (2012)
7. Dumas, B., Lalanne, D., Ingold, R.: Description languages for multimodal interaction: A set of guidelines and its illustration with SMUIML. Journal on Multimodal User Interfaces 3(3), 237–247 (2010)
8. Ehlen, P., Johnston, M.: Multimodal local search in speak4it. In: Proc. of IUI. pp. 435–436 (2011)
9. Galatà, V.: Production and perception of vocal emotions: a cross-linguistic and cross-cultural study, PhD Thesis - University of Calabria, Italy
10. Herm, O., Schmitt, A., Liscombe, J.: When calls go wrong: How to detect problematic calls based on log-files and emotions. In: Proc. of Interspeech. pp. 463–466 (2008)
11. Larson, J.A., Raman, T.V., Raggett, D., Bodell, M., Johnston, M., Kumar, S., Potter, S., Waters, K.: W3C multimodal interaction framework (2003), http://www.w3.org/TR/mmi-framework/
12. Petrillo, M., Cutugno, F.: A syllable segmentation algorithm for english and italian. In: Proc. of Eurospeech. pp. 2913–2916 (2003)
13. Petrushin, V.: Emotion in speech: Recognition and application to call centers. In: Proc. of ANNE [Online] (1999)

14. Polzehl, T., Schmitt, A., Metze, F.: Approaching multilingual emotion recognition from speech - on language dependency of acoustic/prosodic features for anger detection. In: Proc. of Speech Prosody (2010)
15. Yacoub, S., Simske, S., Lin, X., Burns, J.: Recognition of emotions in interactive voice response systems. In: Proc. of Eurospeech. pp. 729–732 (2003)
16. Zaykovskiy, D., Schmitt, A., Lutz, M.: New use of mobile phones: towards multimodal information access systems. In: Proc. of IE. pp. 255–259 (2007)