Requirements for an Emotional Markup Language with Voice Portal Applications in Mind - a Position Paper

Felix Burkhardt

Deutsche Telekom Laboratories
Voice & Multimodal Solutions
Germany
felix.burkhardt@telekom.de

Abstract

We describe an emotion aware voice portal as an application that integrates emotional processing and identify the necessary steps to build such an application. Based on this, requirements for a standardized emotion description language are drawn.

1. Introduction

Anger detection is a topic that gains more and more attention with voice portal carriers as it can be useful for quality measurement and empathic dialog strategies [1, 2]. In the context of customer care voice portals it can be helpful to detect potential problems that arise from a unsatisfactory course of interaction in order to help the customer by either offering the assistance of human operators or try to react with appropriate dialog strategies.

Our group started work on emotional processing around 2003 in several sequential projects, the progress in this work was reported in [3, 4, 5, 6, 7, 8]. This was also our main motivation to participate in the HUMAINE Network on Emotional Human Machine Interaction.

As one of our main topics involve the implementation of automated voice portals, i.e. voice-only human machine dialogs over the telephone, the main application of emotional processing in our work concerns the detection of discontent (in short: anger detection) users during these interactions.

The following sections discuss our approach and what we might expect from EmotionML to support this approach.

2. Emotional processing in Voice Portals

Actually there are three applications with respect to anger detection with voice portals.

- Dispatching callers to trained agents or at least balance the load of angry customers over agents..
- Adapt the dialog design with some soothing dialog strategy that can be done automatically.
- Collect statistics in order to gather information about the contentedness of your customers over time.

Figure 1 shows the integration of the emotion module in the voice portal architecture. Because until now there is no standardized interface for emotional processing in dialog languages like VoiceXML, the module must be integrated on the server side. Here already appears a first requirement with respect to standardization: if EmotionML would be a part of dialog describing standards like VoiceXML or EMMA and manufacturers would integrate emotion recognition into their products just like speech recognition. the dialog manager would not need its own emotion module.

At least EmotionML will ease the integration and exchange of emotion recognition products into the overall system, once their is a market.

3. Stages of Application Building

The development of an emotion aware application requires the following steps.

1. The application functionality must be specified.
2. Based on the specification, emotional situations and how they should be handled must be identified.
3. Based on this, emotion related states that may appear in this situations must be described.
4. To train the detection algorithms, a train and test data set must be collected, either from a prototype, a Wizard of Oz study or a comparable application.
5. The data set must be labeled by human listeners in order to gain a ground truth.
6. The recognizer can then be integrated into the application but should be tuned / re-trained as often as possible in order to adapt to the real world situation.

With respect to step 3), standardized vocabularies of emotion related states would help to make experiences while designing applications comparable and to exchange training and test data between applications. This might be difficult in very application specific requirements, e.g. there might be the goal to differentiate between anger directed at the dialog versus anger directed against a specific product. Of course there also might appear emotion related states that are less related to emotions, like ”interest”, or ”tired”.

Step 4) includes the possibility to use already (perhaps partially) labeled data. Of course EmotionML helps to re-use data.

In step 5), the labelers should be able to express the intensity of the emotion and they confidence they feel with their decision. In a multi-emotion application, it must be possible to label several emotions for the same sample. With respect to longer samples, they might be split into time periods. In the end, each speech sample might (and should) be labeled by several humans and these labels must be unified into a single label somehow. This process is described in the next section.

With respect to step 6), EmotionML will help if the emotional module gets exchanged but the training data should still be used.

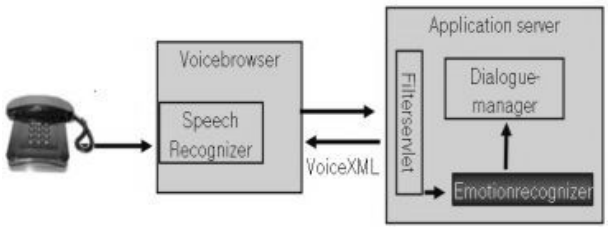

Figure 1: Integration of emotional processing in a Voice Portal architecture

Figure 2: Labeltool User Interface

4. Data Labeling

The idea of using several labelers instead of just one per sample, is to make the labels more stable and objective against the moods of the labelers and the context of the labeling situation. In order to achieve a consistent rating behavior, the labelers get written label instructions and take part in a common session where some examples are discussed. Ideally, the labeling process takes part via headphones in the same room so they can discuss difficult or prototypical samples.

To enable an efficient and easy label process, we developed a labeling tool whose interface is displayed in 2. All samples are displayed in a scrollable table, the labeler gets information on the contents being said (if already transcribed) and the decisions from previous labelers. For each turn, the labelers had the choice to assign an anger value by typing the keyboard between 1 and 5 (1: not angry, 2: not sure, 3: slightly angry, 4: clear anger, 5: clear rage), or mark the turn as ”non applicable” (garbage), i.e. he/she can express the perceived intensity of anger in the voice. After labeling, the next sample will be played automatically, so that the labeling process can be very fast (almost real time) and the concentration must not be diverted by using the mouse. Garbage samples include a multitude of files that can not be classified for some reason, e.g. DTMF tunes, coughing, baby crying or lorries passing by.

In the end one can compute the inter-labeler agreement and identify labelers that disagree conspicuously often from the others. It might be a good idea, to assign their labels with a lower confidence level in general.

We unify the ratings and map them to four classes ”not angry”, ”unsure”, ”angry” and ”garbage” for further processing conducting the following algorithm; in order to calculate a mean value for the three judgments, we assigned the value 0 to the ”garbage” labels. All turns reaching a mean value below 0.5 were than assigned as ”garbage”, below 1.5 as ”not angry”, below 2.5 as ”unsure” and all turns above that as ”angry”. One special rule was used beforehand: if at least half of the labelers voted for 0, we decided for ”garbage”. Otherwise a turn with the labels 0, 0 and 4, mean value 1.3, would have been counted as ”not angry” which is surely not correct.

Of course it would be good for the unification if the labelers would have stated their confidence, but we think it’s better to have a fast and spontaneous procedure. Like this, the labelers will express lack of confidence by assigning lower intensity.

An alternative to unification, at least for the training data, would be to present the classifier with all existent labels and let it sort out the discrepancies for itself. But at least for the test set, in order to evaluate the classifier, a single judgment per sample is needed.

Collecting data and labeling is a very costly process and it will help a lot to have a standardized way to annotate emotions with respect to data exchange. A very important aspect here is the possibility to describe as exact as possible what is meant by a certain emotion category, i.e. in our case saving the labeler instructions with the data.

5. Conclusions

We described an emotion aware voice portal as an application that integrates emotional processing and identified the necessary steps to build such an application. Based on this, requirements for a standardized emotion description language were drawn.

- A standardized interface for emotion description in dialog markup languages would enable the seamless integration into a multichannel dialog that comprises text, semantics and emotions.
- A standardized vocabulary set helps to exchange data but it must be extensible to meet exotic emotion descriptions that might arise from dialog design perspective.
- Together with the vocabulary set it must be possible to state detailed information on what the vocabulary means semantically, e.g. by saving the instructions to the labelers in the "info" element.
- Human labelers must be able to label several emotions at the same time, intensity values for each emotion and an overall confidence for their decision.

6. References

[1] S. Yacoub, S. Simske, X. Lin, and J. Burns, "Recognition of emotions in interactive voice response systems," in *Eurospeech 2003 Proc*, 2003.

[2] I. Shafran and M. R. und M. Mohri, "Voice signatures," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2003.

[3] F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber, "An emotion-aware voice portal," in *Proc. Electronic Speech Signal Processing ESSP, Prague*, 2005.

[4] F. Burkhardt, J. Ajmera, R. Englert, J. Stegmann, and W. Burleson, "Detecting anger in automated voice portal dialogs," *Proc. ICSLP, Pittsburgh*, 2006.

[5] F. Metze, R. Englert, U. Bub, F. Burkhardt, B. Kaspar, and J. Stegmann, "Getting closer: Tailored human-computer speech dialog," *UAIS journal, special issue on Vocal Interaction: Beyond Traditional Automatic Speech Recognition*, vol. 8, no. 2, 2008.

[6] F. Burkhardt, R. Huber, and J. Stegmann, "Advances in anger detection with real life data," in *Proceedings of Elektronische Sprachsignal Verarbeitung (ESSV) 2008*, september 2008.

[7] F. Burkhardt, T. Polzehl, J. Stegmann, F. Metze, and R. Huber, "Detecting real life anger," in *Proceedings ICASSP, Taipei; Taiwan*, 4 2009.

[8] F. Burkhardt, K. Engelbrecht, M. van Ballegooy, T. Polzehl, and J. Stegmann:, "Emotion detection in dialog systems - usecases, strategies and challenges," in *Proceedings Affective Computing and Intelligent Interaction (ACII), Amsterdam, The Netherlands*, 9 2009.