An Emotion-Aware Voice Portal

Felix Burkhardt, Markus van Ballegooy*, Roman Englert**, Richard Huber****

T-Systems International GmbH*, Deutsche Telekom Laboratories**, Sympalog Voice Solutions GmbH***

Berlin / Erlangen, Germany

{felix.burkhardt|markus.van-ballegooy}@t-systems.com,

roman.englert@telekom.de, huber@sympalog.de

Abstract

This article describes the development of a voice portal prototype that integrates the recognition of anger in the caller's voice. Often users get frustrated by talking to machines and hang up without having the possibility to express their anger in some way. In order to handle this problem, we integrated an emotional recognizer in a standard-based voice portal

If anger is detected in the caller's voice, the dialog manager component runs a conciliation strategy by first mirroring the caller's mood and in a second step transferring the caller to a human agent.

The system design is described and first evaluation results reported.

1 Introduction

Business transactions that require no personal encounter get more and more handled by automated voice portals. One channel of human communication that is not regarded until today is the emotional expression. If customers get frustrated by talking to a machine that might not directly understand the user's requests, there is no way to express this anger and the caller feels not understood even more. As unsatisfied customers can become quite expensive, industry is motivated to find ways to tackle this problem with automated strategies (e.g. [1, 2]).

Within the scope of a T-Systems R & D project concerning customer care automation we developed an emotion-aware customer portal in several expansion stages. We describe a prototype of this voice portal based on a standardized architecture that integrates a prosodic and semantic anger-detection module. If anger is detected in the caller's voice, the dialog manager component runs a conciliation strategy by first mirroring the caller's mood and in a second step transferring the caller to a human agent.

The article is structured as follows: The 2. chapter is about the architecture and modularization of the voice-portal. A next section is dedicated to the prosodic anger detector, which is described in more detail in [3] and [4]. This is followed by section 4, which deals with the automated conciliation strategies. We finish with a description on preliminary evaluation tests.

2 Architecture

VoiceXML ([5]) is a markup-language developed by the W3C and used to describe voice-dialogs in telephony-applications. It is perhaps easiest to understand the role of VoiceXML by use of the HTML-analogy: VoiceXML corresponds to HTML, the voice browser to a web-browser and the application server can be used for both channels.

Our voice-portal is based on a standardized architecture. Motivated by the wish to use standard VoiceXML 2.0, we integrated the emotion detection module on the server side although it would of course be more efficient to handle the signal analysis part in combination with the speech recognizer, as both modules share many processing steps.

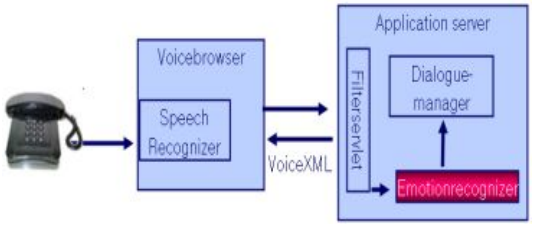

Figure 1: Architecture of Voice-portal

The integration of the emotion recognizer in a VoiceXML-based Voice portal is shown in Figure 1. Via a FilterServlet all requests from the voice browser get filtered. If an audio-file is attached it gets analyzed. The process of analysis is handled in its own thread, so the dialog manager can respond to the user without any time delay caused by the anger detection.

The emotion recognition module consists of a session-global object. This object is fed by emotional ratings from the analysis thread. The ratings get computed from two sources:

- analysis of the acoustic features and
- analysis of the semantic content.

The semantic analysis is quite crude up to now and consists simply of swear-word spotting.

The acoustic analysis is handled by a statistical classifier developed by Sympalog Voice Solutions GmbH [3], which analyses prosodic features. In a first phase of the experiment this classifier delivered a single value between 0.0 and 1.0 denoting the probability that the utterance was spoken in an angry mood. We refined that approach in a second phase by training the classifier with a three-class problem:

- not angry
- low anger
- high anger.

So the classifier delivered as output three normalized values between zero and one. Because of the normalization one of them yields always 1.0 and can be interpreted as the acoustic recognition result.

This distinction was motivated by the wish to detect two steps of anger that would be handled by the conciliation strategies. The division is not to be confused with the famous ”hot” and ”cold” anger reported in literature (e.g. [2]) as it originates from a different point of view.

3 The Acoustic Classifier

The acoustic classifier works as follows: from the user utterance a set of prosodic features is computed and this feature vector is classified into one of two or three classes, respectively, using an algorithm based on Gaussian densities. Only the speech signal is used as input in this processing step. For the classification a set of up to 31 prosodic features is used. The vector consists of features based on the pitch and the energy curve and durational features which are computed on the basis of the voiced/unvoiced decision, since no time alignment of the recognized words is yet available to this module.

After computing the voiced/unvoiced decision, the pitch and the energy curve, the following 31 prosodic features are generated:

Pitch features:

minimum, maximum, average, median, standard deviation, average of first voiced block, average of last voiced block, number of voiced blocks with positive regression coefficient, average of positive regression coefficients, number of voiced blocks with negative regression coefficients, average of negative regression coefficients, maximum of positive regression coefficients, maximum of negative regression coefficients, and regression coefficient of last voiced block.

Energy features:

minimum, maximum, average, median (all over voiced frames), ratio of energy integral between voiced and unvoiced frames, average of first voiced block, energy of last voiced block, number of voiced blocks with positive regression coefficients, average of regression coefficients of voiced blocks, number of voiced blocks with negative regression coefficients, average of negative regression coefficients in voiced blocks, maximum of positive regression coefficients in voiced blocks, maximum of negative regression coefficients in voiced blocks and regression coefficient of last voiced block.

Duration features:

average length of voiced blocks, ratio between number of voiced and unvoiced frames, number of voiced frames.

This prosodic feature vector is computed for the complete speech signal, i.e. we do not use a frame based approach. For the classification we use a Gaussian classifier with two classes (phase I) and three classes (phase II). Each class is modeled using four Gaussian densities. Strictly speaking, in both phases the acoustic classifier does not classify the speech signal, it rather calculates a score for every class.

In phase I, where the two classes *angry* and *not angry* are discriminated, each density is evaluated using the feature vector and the values are normalized. We return the maximum of those

values, if the corresponding density belongs to the class *not angry*, otherwise *value = 1 - maximum* is returned.

In phase II we calculate a score for every class, which is the minimum of all negative logarithms from the evaluation of the corresponding densities. Then, those three scores are normalized as follows:

$$min(scores) = min(S_{notangry}, S_{lowangry}, S_{angry})$$

$$S'_{notangry} = \frac{min(scores)}{S_{notangry}}$$

$$S'_{lowangry} = \frac{min(scores)}{S_{lowangry}}$$

$$S'_{angry} = \frac{min(scores)}{S_{angry}}$$

As a result, one of the three scores always has the value 1.0. Afterwards, the easiest way to classify is to decide for the class which belongs to the score with the value 1.0. Since the acoustic-prosodic processing delivers the three scores, more sophisticated decision functions, e.g. combination of thresholds, linear combinations or even neural networks could be applied.

4 Conciliation Strategies

If one asks for the purpose of uttering negative emotions in every day communication, three main functions can be identified: Uttering negative emotions may serve to

1. inform your communication partner about your own emotional status in order to give him a complete comprehension of the information you want to express (your emotional appraisal of the information given).

2. inform your communication partner about the perceived, respectively the desired quality of relation between the communication partner and yourself (e.g. denial or distrust)

3. induce a certain action or behavior of your communication partner that you want him to show (e.g. yielding or flinching).

In every day communication situations there are many possibilities for communication partners to react on the utterance of negative emotions and the affiliated intentions. Depending on the current situation, the personalities and the current goals of the communication partners there might occur different reactions with different positive or negative consequences. A fast termination of the negative emotional state can be seen as a criterion for a successful and adequate reaction. Depending on the communicators' intentions, also other types of behavior may be regarded as goal leading and successful.

In functional and professional communication contexts (e.g. call center communications), the occurrence of negative emotions (especially anger) is often not regarded as goal leading and therefore mostly unwanted. First, strong negative emotions may distract the communication from the actual communication topic and therefore lead to a loss of efficiency in communication, secondly the exposure to strong negative emotions itself poses a strong emotional strain on the communicators. For this reason so called conciliation or deescalation strategies are frequently used by professional communicators in order disarm an emotionally charged situation quickly and switch back to the actual communication topic again.

But the main intention here is not to simply skate over the negative emotions (of a customer). Today there is a prevailing insight, that it is most essential for customer satisfaction and customer retention to give him the feeling that also his negative emotions are taken seriously. The main distinctive feature of conciliation strategies is - in contrast to ”normal” everyday communication - the conscious and goal orientated use in the case of rising negative emotional states.

Examples of effective conciliation strategies for negative emotions comprise

- Distraction/ change of topic, without showing an interest for the negative emotions
- Providing information, which rebuts the possibly wrong assumptions that led to a negative emotional state.
- Feedback / Mirroring the perceived emotions in order to show awareness for the communicators emotional state.
- Emphatic utterances, that show ones' sympathy for the communicators emotions
- Further encouragement of the communicator to express his emotional state
- Pointing out alternatives for being angry or frustrated
- Humor / Joking / Teasing. Reason (appeal to the communicators senses)

While transferring conciliation strategies from the inter-human communication into dialog behavior of an emotional-aware voice portal we face two major constraints:

1. As you see in this article it is possible to detect whether a caller is angry or not by using the state of the art technology in emotion detection. Unfortunately, what can not be detected is the reason why. To be credible, only those strategies can be transferred into man-machine dialogs that are generic enough so that it is not required to make references to the content subject of the negative emotion.
2. As soon as communication between man and machine leaves the immediate task that the user actually intends to handle with the dialog system and switches over to abstract topics (e.g. to the reason why the user is angry), the result is a complexity of the dialog that is neither calculable nor manageable within a speech dialog system. Therefore, conciliation strategies in a dialog system must be designed very straightforward and narrow so that the user will not be encouraged to stray from the task .

Despite these constraints we tried to transfer two conciliation strategies into the speech dialog of our emotion aware voice portal. It was set up as a self service application where customers can look up their telephone bills and find information about their mobile phone tariffs.

1. Within dialog situations where slight anger was detected and where there were no further hints that the users request could not be handled within the voice portal, we used a strategy that is called ”conciliation by mirroring”. The goal of this strategy is to show the user that his emotions are recognized but that it is better to continue the task. Immediately after an utterance that was classified as angry by the acoustic classifier, our system interrupted the main dialog by playing a short prompt like: *”I see that your are a little bit excited. So it will be the best if I continue quickly with your query!*”

2. Within dialog situations where strong anger was detected in combination with hints that the dialog will not be completed successfully (e.g. if there where more than 2 out of grammar utterances during the history of the dialog) we used a strategy called ”conciliation by empathy and delegation”. Since we did not believe that very angry users could be handled within a speech dialog and should rather be treated by a human operator, we decided to offer him the possibility to leave the voice portal and speak to a ”real” service agent. Before the connection was put through, the system verbally showed empathy for the users situation. We used a prompt like: *”I notice that you are angry because I don’t grasp your request. I can really understand you! The best thing will be if I put you through to a service agent.*”

5 Evaluation

5.1 Subject of Evaluation

As we didn’t run a field test yet we can not say how well the emotion-aware voice portal performs with respect to usual voice portal evaluation measures like e.g. task-completion rate. From informal evaluation we can only conclude that the automatic detection of anger is a very sensible field, as the false detection of anger has a highly comical effect (just imagine a machine of accusing you to be angry although you’re not) or might even produce anger instead of lessening it. Also the handling of emotional communication results in the illusion of intelligence the machine just doesn’t possesses and might lead to excessive demands with respect to state-of-the-art natural language understanding modules. For that reason it’s probably safest to use this technology with less sensitive tasks like automatic measurement of customer satisfaction.

Here we report the evaluation of anger detection. In a first step we built the voice-portal as a demonstrator program. It was trained with approximately three hours of speech material containing about one hour low and 10 min. high anger. Note that the inter-labeler agreement was very low; while two of the three labelers agreed of about 80 minutes of angry samples, the third one judged 150 minutes of the three hours as sounding angry.

5.2 Two Phases Experiment

Up to now, we divided the project into two phases, which, with respect to the emotion-awareness, differ mainly in the features regarding the acoustical anger detector developed by Sympalog.

The following extensions were made:

- While the first module delivered one value to be interpreted as the degree of anger, the second one gave three values as output to be interpreted as probabilities for low, high anger or not angry speech.
- The set of acoustic features was extended from 13 to 31.
- While the first classifier was trained on a very limited database of acted recordings of around 3 minutes length, the second used as a basis three hours of recorded conversation.

5.3 Overall Recognition

Our first aim was to see whether the recognition of anger compared between the two phases was getting better.

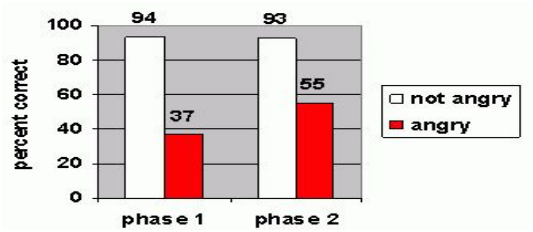

Figure 2: Comparison between the phases

As a test database we used recordings from the first demonstrator consisting of 6 minutes angry and 20 minutes not angry utterances.

In Figure 2 the overall recognition rate for angry and not angry samples is shown. Especially the recognition of anger improved from 37 % to 55 %. As we got similar results when running the phase 2-classifier with the limited feature set from phase 1, the improvement must mainly result from the enlarged database and the different classification algorithm.

5.4 Thresholds

In order to have a threshold that would permit us to tweak the recognition result in favor of correctness of non-anger-detection, we used in phase one the single probability value. Unfortunately this value wasn't distributed uniformly at all but extreme values under .1 and above .9 appeared in about 60 % of all cases, which is a consequence of the insufficient number of training utterances.

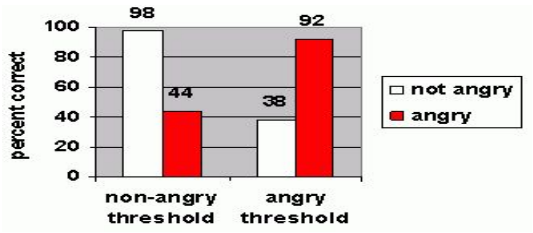

Figure 3: Results using thresholds

In phase two we used the threshold on the three recognition results. If we were interested mainly in a correct recognition of non-angry utterances, we can use a threshold on the not-angry probability irrespective of the angry-values.

Figure 3 depicts the results achieved with a not-angry (0.95) and angry threshold (0.9). As can be seen, the overall recognition rate (or equal error rate) always tends to be about 75 %, but using thresholds and a multiple class recognizer the recognition can be tweaked in favor of anger- or non-anger detection. An informal listening test indicated that the remaining angry utterances that pass the not-angry threshold do indeed sound angry to a high degree.

5.5 Distinction between low and high anger

The distinction between low and high anger was introduced without a profound definition or clear description. Neither the labelers of the training nor of the test database had an objective criterion to measure an ”anger-scale”. Therefore it is not very astonishing that the classifier does not distinguish between these classes very well. Table 1 shows a confusion matrix of the recognition results for the three classes. It’s based on the same test database as the other tests. Low anger was often confused quite often with no anger (46%) and high anger wasn’t really recognized at all (16%) but confused with low anger (54%). As stated above, the distinction isn’t clear anyway.

Table 1: *Confusion matrix for low and high anger (in %)*

Ref. \Test	no anger	low anger	high anger
no anger	89	9	1
low anger	46	49	4
high anger	28	54	16

6 Outlook

We have several plans for the further development of our emotion-aware voice-portal.

- We’ll try to enhance the anger detection by taking word boundaries into account and dismiss absolute values (like e.g. global pitch mean).

- We will go back to a binary anger-detection but will work on the confidence values.
- A personalized version where the anger probability is computed as deviation from a neutral speech sample is envisaged.
- Simplify the interface to train new data in order to feedback the system more easily.
- A first pilot operation is supposed to result in a collection of ”real” voice-data as well as evaluation data with the system.

The evaluation of the conciliation strategies' efficacy is still an open question. Unfortunately we could not run any tests with participation of real users till now. We are going to conduct focus groups and field tests in later phases of the project.

7 Conclusions

We described a voice portal system that includes the detection of anger in the caller's voice and tries to react accordingly. It is based on a standard VoiceXML framework and can easily be integrated into existing architectures.

Up to now we're still in the development phase and evaluation has been done only based on faked data, but a pilot operation is already envisaged. We're convinced that anger-detection is an important factor to enhance user satisfaction.

8 Acknowledgments

This work was partially funded by the EU FP6-IST project HUMAINE (Human-Machine Interaction Network on Emotion).

References

[1] L. Devillers, L. Lamel, and I. Vasilescu, “Emotion detection in task-oriented spoken dialogs,” in *IEEE Multimedia & Expo Proc.*, 2003.

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

[3] R. Huber, *Prosodisch-linguistische Klassifikation von Emotionen*. Logos Verlag, Berlin, 2002.

[4] R. Huber, A. Batliner, J. Buckow, E. Nöth, V. Warnke, and H. Niemann, “Recognition of Emotion in a Realistic Dialogue Scenario,” in *Proc. Int. Conf. on Spoken Language Processing*, vol. 1, Beijing, China, Oktober 2000, pp. 665–668.

[5] “Voice extensible markup language (voicexml) version 2.0.” World Wide Web Consortium (W3C), http://www.w3.org/TR/voicexml20.