Automatic customer feedback processing: alarm detection in open question spoken messages

Nathalie Camelin[1], *Geraldine Damnati*[2], *Frederic Bechet*[1], *Renato De Mori*[1]

[1] LIA - University of Avignon, BP1228 84911 Avignon cedex 09, France, Fax: 33 4 90 84 35 01
[2] France Télécom R&D - 2 av. Pierre Marzin 22307 Lannion Cedex 07, France, Fax: 33 2 96 05 35 30

{frederic.bechet,nathalie.camelin,renato.demori}@univ-avignon.fr
geraldine.damnati@orange-ftgroup.com

Abstract

This paper describes an alarm detection system dedicated to process automatically customer feedbacks in call-centers. Previous studies presented a strategy that consists in the robust detection of subjective opinions about a particular topic in a spoken message. In the present study, we focus on the *alarm detection* problem in a customer spoken feedback application. We want to characterize each customer's survey with a degree of emergency. All the messages considered as *urgent* need a quick answer from the call-center service in order to satisfy the customer. The strategy proposed is based on a classification scheme that takes into account all the features that can characterize a survey: answers to the closed questions, topics and opinions detected in the open question spoken message, confidence scores from the Automatic Speech Recognition (ASR) and Spoken Language Understanding (SLU) modules. A field trial realized among real customers has shown that despite the ASR robustness issues, our system efficiently ranks the most urgent messages and brings a finer analysis on the surveys than the one provided by processing the closed questions alone.

Index Terms: Automatic Speech Recognition, Speech Understanding, Confidence Measures, Speech Mining.

1. Introduction

The constant growth of call-center services for dealing with help-desk, customer-care or telephone sale applications has pointed out the need for robust automatic speech processing. The applications of such an automatic processing are the partial or full automation of some services [4, 3] as well as speech mining in the recorded conversations for quality control and service improvement.

Along the speech mining task, important information that has to be collected is the customers' feedback after using a call-center service. This feedback is usually obtained through a survey performed on a voluntary basis among the customers that have recently called the service. The cost of surveys performed by human operators and the need to collect as many feedbacks as possible have lead to the development of automated survey applications where a customer replies to a list of closed questions. The answers to these questions are then analyzed manually or by means of Automatic Speech Recognition (ASR) techniques.

This approach has the advantage to be quite simple to process however there are two major drawbacks: firstly the list of closed questions has to be small to prevent customers dropping off the survey before the end, therefore these questions can't cover all the aspects of a call-center service; secondly the callers are often prone to add comments to their answers to closed questions, being frustrated to have only a limited set of answers not necessarily matching their opinion.

For these reasons an open question like *"Please add any further comments you would like to make on the service"* is often added to the survey. The spoken messages collected thanks to this open question are similar to the *verbatim* transcribed by operators, or collected through WEB-based surveys. However the automatic processing of such messages is particularly challenging as they contain most of the current issues in ASR research: spontaneous non constraint speech, disfluences, telephone speech with channel and background noises, ...

We presented in previous studies [2] a strategy that consists in the robust detection of subjective opinions about a particular topic in a spoken message. Distributions of positive and negative opinions on several topics were automatically extracted and compared to those obtained with a manual approach. In the present study we are going to focus on the *alarm detection* problem in a customer feedback application. Instead of estimating global distributions of opinions, we want to characterize each customer's survey with a degree of emergency. All the messages considered as *urgent* need a quick answer from the call-center service in order to satisfy the customer (and of course prevent him/her dropping off the service!).

The strategy proposed is based on a classification scheme that takes into account all the features that can characterize a survey: answers to the closed questions, topics and opinions detected in the open question spoken message, confidence scores from the ASR and Spoken Language Understanding (SLU) modules. The basic assumption is that the alarm classification process will give a higher score to the surveys containing a lot of redundancy in the expression of the dissatisfaction. A field trial realized among real customers by the telephone survey service of France Telecom has shown that despite the ASR robustness issues, our system efficiently ranks the most urgent messages and brings a finer analysis on the surveys than the one provided by processing the closed questions alone.

2. Telephone survey corpora

In this study we use several survey corpora collected from France Telecom customers.

A first corpus has been collected during a 3 month period.

Users were invited through a short message to call a toll-free number where they can express their satisfaction about the customer service they recently called. About 1 800 messages were collected, with a duration limitation of 2 minutes. Theses messages have been transcribed and annotated by human operators according to the following topics : *courtesy* representing the courtesy of the customer service operators, *efficiency* related to the efficiency of the customer service and *rapidity* concerning the amount of time they had to wait on the phone before reaching an operator. Each topic is associate with a positive (+) or negative polarity (−), leading to a set of 6 opinion labels. An example (translated from French to English) of a message with its manual annotation is given below, as we can see the human annotators have indicated in the transcribed message the beginning and the ending of an opinion expression (with the topic and the polarity). These segments are called the *support* of a given opinion.

"yes uh uh here is XX XX on the phone well I've called the customer service yep <courtesy+> *the people were very nice* </courtesy+> <efficiency+> *I've been given valuable information* </efficiency+> *but* <efficiency-> *it still doesn't work* </efficiency-> *so I still don't know if I did something wrong or [. . .]"*

This corpus contains 1779 messages and is called in this study *Train1*.

A second corpus has been collected within a more general experimental framework of automated telephone surveys mixing closed and open questions. In the first part of the survey, four closed questions were asked to the users to explicitly know his/her feeling about: the global satisfaction, the courtesy, the efficiency and the rapidity of the customer service.

Questions are provided using speech synthesis and the possible answers (namely: *completely satisfied*, *partially satisfied*, *dissatisfied*, *do not know*) are processed on-line by an ASR system. After these four closed questions, the customer is invited to leave a message through the following prompt :

"[...] Thank you for calling back and participating to the survey. If you have any other comment, please leave a message [...]"

About 700 messages were recorded on this experiment, and have been also manually transcribed and annotated by operators according to the same annotation scheme presented for the first corpus. This corpus is called *Train2*.

Finally, a last set of messages has been collected in a field trial that aims to evaluate the alarm detection system presented in this paper. In this last experiment we use the same protocol as the one used to collect the corpus *Train2*. The main difference is in the annotation scheme: the operators who processed the surveys were asked to rank each customer feedback according to its emergency, an *urgent* message being one left by a customer who needs immediate attention.

Four degrees of emergency have been defined for characterizing customers' feedback:

- *empty*: no understandable speech in the messages left after the open question;
- *none*: messages that don't require any specific attention (happy customers!);
- *moderate*: messages expressing the need for a human intervention, but with no urgency;
- *urgent*: for messages that require an urgent call-back from the service.

These degrees of emergency are called *ref emergency* labels in this paper.

A corpus was collected between January and February 2008: 352 user feedbacks were collected and constitutes the test corpus of this study. The proportions of *ref emergency* labels within this corpus are indicated in table 1. As we can see, 50% of the messages requires a call-center operator to call-back urgently the customer.

empty	*none*	*moderate*	*urgent*
7%	20%	23%	50%

Table 1: Proportions of manual emergency annotations.

3. From Opinion mining to Alarm Detection

The alarm detection system described in this paper is made of a two-step process: firstly the spoken messages collected through the open question prompt are analyzed for extracting all the opinions expressed by the customers thanks to our opinion-mining system presented in [2]; secondly the different opinions extracted, with ASR and SLU confidence measures, as well as the automatic transcripted answers given by the customers to the four closed questions are processed by another classification module in order to estimate the emergency level of each message. This two-step process is described in the next two sub-sections.

3.1. An opinion detection system

Subjectivity and opinion detection has been quite popular in recent years [1, 7]. The opinion detection system we use in this study is described in [2]. The purpose of this system is to detect opinion labels from audio messages in two steps :

- *Segmentation process*: The segmentation process is directly integrated into the ASR system thanks to specific Opinion Language Models. The output of this process is a string of segments, each of them likely to be the support of a given opinion. Each segment is associated with an ASR confidence measure. A threshold α is applied to this confidence measure for keeping only reliable opinion supports.
- *Opinion classification process*: Each segment output by the previous process is processed by a classifier based on a boosting algorithm [6]: a probability is associated to the segment for each of the six opinion labels, this corresponds to the probability for the segment to be the support of the opinion label. A threshold β is applied to this probability in order to select only the reliable opinion labels for a given segment.

3.2. Alarm Detection System

The goal of the alarm detection system is to select among all the customers feedback those which require the most urgent human intervention as presented in section 2. We want to reproduce the *ref emergency* labels on our test corpus. However we didn't have any training corpus containing such annotations. Therefore we made the following assumption: there is a correlation between the negative opinion redundancy in a spoken message and its emergency. In other words, the more a customer is dissatisfied, the more he will express his feelings with negative opinion expressions and the easier it will be for the opinion detection system to detect these expressions. Indeed we used the confidence measures of the opinion detection system as a direct indication of the emergency of a given spoken message.

In order to obtain an emergency label for each customer feedback we propose the strategy presented in figure 1, which contains three steps: processing the open question spoken message; processing the closed question answers; taking a decision of the priority label.

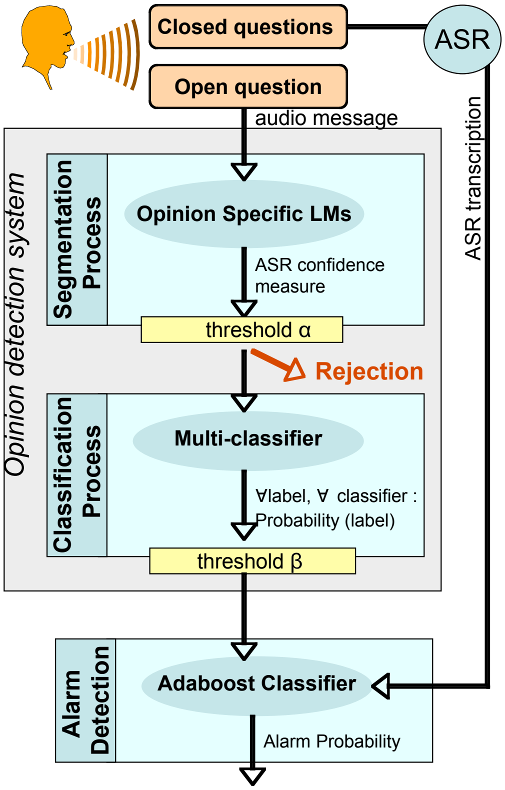

Figure 1: Alarm detection system.

3.2.1. *Processing open question spoken messages*

In [2], we have shown that varying thresholds α and β leads to different performance of the system in term of F-measure, a combination of precision and recall measures. As the goal of the alarm system is to have a high precision in opinion detection in order to evaluate the degree of emergency and also not to reject *urgent* messages, we chose values of α and β leading to the best F-measure on the corpora *Train2*, with the models trained on *Train1*.

To label each segment output by the segmentation process we used a multiple-classifier scheme. We have shown in [5], that the agreement measure among several classifiers implementing different classification algorithms is a powerful confidence indication. We used three classifiers: one based on a boosting algorithm, a decision-tree one and an SVM-based one.

Each classifier gives a score to each opinion label to be associated to a message. These scores are converted into confidence measures by means of a regression process. Inside the same message, positive and negative opinions about the same topic can be recognized. This is important as a significant portion of the messages in our corpora contain such mixed-feeling messages.

3.2.2. *Processing the closed question answers*

For each customer feedback, the answers to the four closed question are processed by an ASR module. Its goal is to identify one of the four possible answers to these questions: *completely satisfied*, *partially satisfied*, *dissatisfied*, *do not know*. If no answer is recognized, or if the customer gives a long message to justify his answer a fifth label is introduced: *unknown*.

3.2.3. *Taking a decision of the priority label*

The last process is based on a boosting based classifier which is trained on the corpus *Train2* to evaluate the dissatisfaction of the customers. During the training process, the features given to the classifier were the following:

- the result of ASR system for the four answers to the closed questions;
- the count of negative segments contained in the message according to the different classifiers with a confidence score highest than β;
- for each opinion label, the agreement measure on this label among the different classifiers.

And the label given to each set of features was either *negative* if the spoken message contained the expression of a negative opinion about the service; or *positive* if no such opinion was expressed.

During the classification process, the score output by this last classifier is converted into a confidence score with a logistic function. This score is then normalized to obtain natural values between 0 and 10. We consider that greater is this score, the more urgent is the message.

4. Evaluation

In order to validate the confidence scores representing the emergency degree of our alarm detection system, we define three levels of emergency automatically given as follows:

- *level1*: scores from 0 to 3 are the first level which represents non-urgent messages
- *level2*: the middle level is made of the scores from 4 to 6
- *level3*: scores from 7 to 10 are the third level which represents the urgent messages

Table 2 shows the correlation between the emergency labels given by the human annotators and the levels of emergency estimated by our alarm detection system.

(%)	*level1*	*level2*	*level3*
empty	21.4	0.9	0.8
none	33.0	23.4	5.4
moderate	33.1	28.0	7.9
urgent	**12.5**	**47.7**	**85.9**

Table 2: Evaluation of messages proportions according to the *ref emergency* labels and the three levels of emergency given by the alarm detection system.

There is an obvious correlation between the *ref emergency* labels and the levels obtained automatically by our alarm detection system. Indeed, 85.9% of the messages automatically classified *level3* are considered as *urgent* as opposed to only 12.5% in *level1*. Furthermore, *level3* contains only 5.8% of messages that do not require a specific attention (*none*). This confirms our assumption that the more users express their dissatisfaction, the easier is the classification task and the higher is the confidence scores obtained.

For the sake of comparison, we can evaluate the repartition of the messages of our test corpus as a function of the number of *dissatisfied* answers to the closed questions.

In table 3, proportion of *urgent* messages is evaluated according to the number of *dissatisfied* answers to the closed questions for a given survey.

#*dissatisfied* answers	corpus coverage (%)	
to closed questions	*urgent*	total
≥ 0	50.3	100
≥ 1	69.2	66.1
≥ 2	76.2	36.2
≥ 3	78.3	10.2

Table 3: Evaluation of total proportions and *urgent* messages proportions according to the number of *dissatisfied* answers to the closed questions.

Table 3 shows that only 10% of the customer answers contains at least 3 *dissatisfied* recognised answers to the closed questions. And still more than 20% of them are not considered *urgent* by the manual annotators. If we take all the feedbacks containing at least one *dissatisfied* answer to the closed question the total corpus coverage is much higher (66.1%) however it contains over 30% of non *urgent* messages.

Another interesting result is that 30% of the *urgent* feedbacks do not have any *dissatisfied* answers to the closed questions.

These results show that the answers to the closed questions cannot be used alone and this justify the need for a robust processing of the spoken messages left by the customers. Our automatic alarm detection system based on several measures including these answers to the closed question in addition to the opinions detected in the open question spoken messages with confidence scores leads to a better urgent message detection system.

The results shown in table 2 are very significant from a service point of view as too many messages are likely to be collected every day and the operators can't process manually all of them. Having a system selecting the highest urgent ones with a confidence of more than 85% is a valuable result.

The repartitions of messages according to the *ref emergency* labels and the confidence scores given by the alarm detection system are shown in figure 2.

Particularly, it is interesting to analyse the different proportions obtained when selecting all the messages that reach a score upper or equal to 7. At this step, about 36% of the messages in the test corpus are selected and more than 85% of these messages are considered as *urgent* by the human operators. Furthermore, at this step, 62% of the *urgent* messages are selected. By comparison, considering only answers to closed questions and the same corpus coverage, only 76% of *urgent* messages are selected, representing only 55% of the whole *urgent* messages.

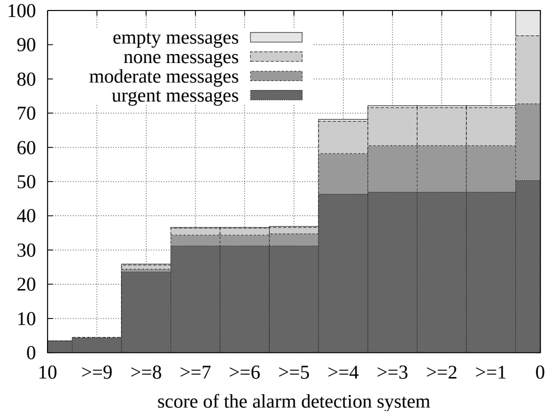

Figure 2: Proportion of selected messages according to *ref emergency* labels and the emergency score given by the alarm detection system.

5. Conclusion

We presented in this study an automatic alarm detection system that ranks efficiently customer spoken feedbacks according to a level of emergency. Several features that can characterize a survey are taken into account : answers to closed questions, topic and opinion detection from an open answer, ASR and SLU confidence measures.

The field experiment carried on at France Telecom has validated our assumption that dissatisfied customers express more clearly their negative opinions, leading the opinion detection system to produce better confidence scores that enables the alarm detection system to finely analyse the emergency degree of a message.

6. References

[1] Eric Breck, Yejin Choi, and Claire Cardie. Identifying expressions of opinion in context. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 2007.

[2] Nathalie Camelin, Fredric Bechet, Geraldine Damnati, and Renato De Mori. Speech mining in noisy audio message corpus. In *Proceedings of InterSpeech*, Antwerp, Belgium, September 2007.

[3] Geraldine Damnati, Frederic Bechet, and Renato De Mori. Spoken Language Understanding Strategies on the France Telecom 3000 Voice Agency Corpus. In *Proceedings of ICASSP'07*, volume 4, 2007.

[4] Allen L. Gorin, Giuseppe Riccardi, and Jerry H. Wright. How may I help you? *Speech Communication*, 23(1-2):113–127, 1997.

[5] Christian Raymond, Frederic Bechet, Nathalie Camelin, Renato De Mori, and Geraldine Damnati. Sequential decision strategies for machine interpretation of speech. *IEEE*, 15(1):162–171, january 2007.

[6] Robert E. Schapire and Yoram Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39:135–168, 2000.

[7] Jayce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *CICLing*, 2005.