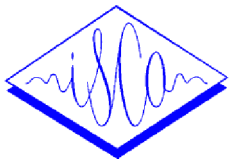
ISCA Archive
http://www.isca-speech.org/archive



Speech Prosody 2008
Campinas, Brazil
May 6-9, 2008

Using Prosody for Automatically Monitoring Human-Computer Call Dialogues

Woosung Kim

Convergys Corporation
201 E. 4th Street, Cincinnati, OH 45202, USA
woosung.kim@convergys.com

Abstract

In human-computer call dialogues, human callers often get frustrated or angry due to, e.g., the computer's mistakes. Detecting such emotions would be beneficial for many purposes; nevertheless, emotion detection so far has been studied primarily as a classification task. Taking a step forward from classifying emotions from a single utterance, this paper investigates whether emotions, detected by prosodic features, can be used *practically* at the dialogue level, i.e., for monitoring human-computer dialogues to detect BAD calls requiring human agent's assistance. We first show emotion detection can be improved by a regression model. In combining emotion detection and dialogue monitoring, we demonstrate decision level fusion is better than feature level fusion. Our experiments also confirm that NEGATIVE emotions may be a sufficient, but not a necessary condition for detecting BAD calls. Finally, we show that BAD calls due to caller's NEGATIVE emotions may be identified by other clues.

1. Introduction

Due to the advances in computer telephony integration, the last decade has seen a steep and continuous growth in the call center market. As a result, the typical number of calls that a call center receives is huge, which can easily go over a few hundred thousand calls even in a single day. When the call volume is so large, manually monitoring, i.e., listening to and analyzing calls—one of the main call center operation tasks—is practically impossible, which leads us to seek an alternative, *automatic call monitoring*. Call monitoring is in fact very important. E.g., how can we evaluate the agent's performance? Or, how do we know whether callers are satisfied or not? To answer those questions, we need to monitor calls, preferably in an automatic manner.

While human agents are the majority in a call center, many call centers have been increasingly using automated (computer) agents, in an effort to reduce operation costs by automating a call center. These automated agents, which are capable of listening to and understanding the caller's speech as well as prompting callers with speech, can replace human agents but at a much lower cost. Indeed, callers may no longer need any human agents until they finish the call to complete their task if the automated agent receives the call[1]. The downside, on the other hand, is that the automated application is not perfect, often causing miscommunication, subsequently caller frustration and dissatisfaction. Reluctant to talk to a computer, accordingly, many callers ask for *expensive* human agents, which makes call center's automation effort futile. Hence, this trade-off between call center operation costs and customer dissatisfaction is the main challenge for automating a call center.

[1]Such automated agent-based applications are often called *self-care* as callers can complete the call by themselves without human agents.

Recently, it has been proposed that automatic call monitoring can also be used for human-computer, self-care applications [1]. The basic idea is to let the caller use the self-care application first and monitor the call, automatically. Whenever the caller has problems with self-care, the monitoring system detects it and brings in a human agent to help the caller. Notice that a human agent is brought in *only if* the caller has a problem; otherwise, the caller keeps using the self-care until the end of the call. This way, we can minimize the cost for human agents and keep callers from being frustrated or dissatisfied.

While the approach—which classifies GOOD or BAD calls from caller response sequence features—shows a promising result, it is interesting to investigate whether emotions can be used additionally for call monitoring assuming emotions are reliably detected. Intuitively, troubled callers would express NEGATIVE emotions such as frustration or anger, naturally, and such emotions may be a good indicator for identifying problematic calls requiring human assistance. Furthermore, we initially hypothesize that NEGATIVE emotions are a *sufficient* condition for BAD calls but may not be a *necessary* condition. Put differently, if a caller is angry or frustrated, we should bring in a human agent; however, there are other callers, though not angry or frustrated, who need human assistance. These issues certainly deserve an investigation, which is the main motivation for this paper.

2. Prior work

For the sake of completeness, this section briefly reviews previous work related to emotion detection (ED) and dialogue monitoring (DM). Being able to detect human emotions is certainly fascinating in human-computer interaction, as it allows the computer to adaptively change its behavior and to better serve the human (caller). Along the same line, ED has been extensively studied in the context of human-computer dialogue systems [2, 3, 4, 5]. Current dialogue systems, though, are inevitably erroneous, often causing users to get angry or irritated. The intent is, therefore, to the detect caller's emotion especially when they get emotionally aggravated, which leads to poor customer service. There are other clues, as well as emotions, for detecting communication problems in human-computer dialogues, which have been studied elsewhere [6, 7]. Most studies, however, have focused on classifying emotions or miscommunication from a single utterance. One exception is to detect communication errors at the dialogue level, but the decision is made after a fixed number of turn exchanges without considering the whole dialogue content [8, 9].

Our goal is to find a solution for the trade-off between cost and customer (dis)satisfaction, and it is, therefore, crucial to detect BAD calls as accurately as possible. Hence, our strategy is to monitor the call and withhold our decision until enough clues are observed, which distinguishes this approach from others.

3. System overview

In essence, our task is to combine two technologies, ED and DM: i.e., we extract useful information in terms of emotions from human-computer dialogues and use it for DM, to detect BAD calls where the callers are having troubles with a self-care application. Both DM and ED are essentially a classification problem where the task is to decide the most probable target class out of previously defined classes, given test input data. The target class would be either a NEGATIVE emotion or a NON-NEGATIVE (¬NEGATIVE) emotion in ED (cf. §4.1) and a GOOD call or a BAD call in DM (cf. §4.2). Such hard decisions on the target class, however, may not be suitable for combining ED and DM. Rather, it would be better to generate scores on a certain decision like a soft decision from each and combine the scores to make the final decision, which leads us to approach this as a regression problem, i.e., using model trees [10]. Model trees are similar to decision trees and regression trees as they all share the same tree-like structure [11]. Unlike decision trees where each leaf node has a single class label, model trees have a linear equation to yield scores at each leaf node. They also differ from regression trees as a leaf node in regression trees has an averaged numeric value instead of a linear equation.

One issue in applying a regression model to a classification problem is how to assign numeric target values for nominal class-labeled data. If the task is binary classification, we can, e.g., assign one class (¬NEGATIVE emotion in ED and GOOD call in DM) to 0 and the other to 1[2]. In fact, model trees have been successfully applied to the classification problem showing better performance than decision trees in many cases [12].

Fig. 1 depicts the overview of our approach. On the left side of the figure is our baseline DM system which is based on call logs. More specifically, it extracts numeric features such as the number of caller turns (cf. §5.2) from call logs and detects BAD calls using a classifier. Our ultimate goal is to improve the baseline DM by incorporating ED, based on prosodic features, as shown on the right side of the figure. One issue is there are multiple ways, as listed below, for coupling ED with DM, out of which we are going to exploit choices 2 and 3.

1. We could *fuse* raw prosodic features with call log features at the feature level and proceed to use the extended features. However, this does not require or take advantage of emotion-labeled data at all, and hence we exclude this.
2. Also, we could first build an ED system using emotion-labeled data, obtain ED results (scores), and then use the ED scores as additional features, as depicted by the dotted line in the figure, to build a final classifier.
3. Finally, we could build an ED system, obtain ED scores, and combine the ED scores with the baseline DM scores at the decision level, e.g., via linear combination.

4. Corpora

4.1. Emotion detection corpus

For ED, we have collected a human-computer call dialogue corpus from a self-care application. This application is, in fact, operated by a human wizard and thus the corpus is collected by the *Wizard of Oz* method. Each call is first segmented into caller's turns and then each turn segment is emotionally labeled (cf. Table 1). Each segment is labeled only once. Since we are mostly interested in NEGATIVE emotions, we merge POSITIVE and NEUTRAL data, yielding ¬NEGATIVE data. Notice here that NEGATIVE emotions occurred much less (12% of all turns) than ¬NEGATIVE emotions, which agrees with what was observed by others [3].

[2]Even multi-class classification can be solved by constructing multiple binary classifiers and then combining the results from each classifier.

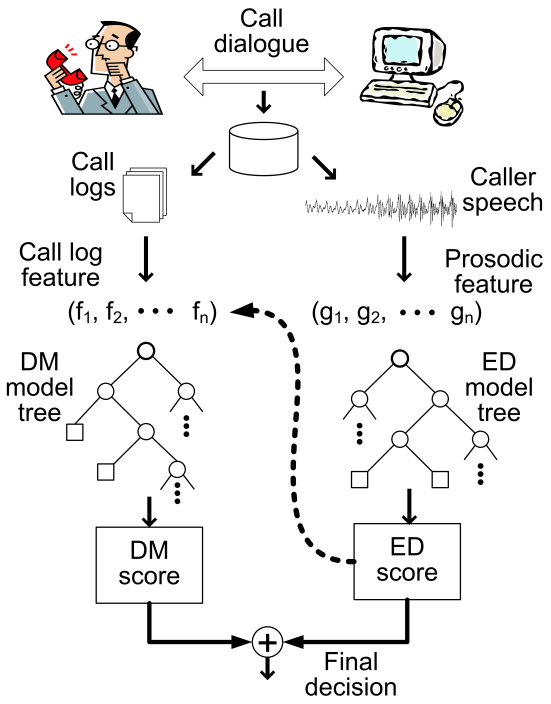


Figure 1: *System overview*

Table 1: *ED corpus summary*

No. Calls	No. Turns	Emotion Label (No. Turns)		
		NEUTRAL	POSITIVE	NEGATIVE
738	3314	2815	20	419

4.2. Dialogue monitoring corpus

Our ED corpus, however, does not have any DM labels, meaning no notion of GOOD/BAD call is available, which forces us to build another corpus for DM. We have obtained another human-computer call dialogue corpus from another self-care application. Unlike the ED corpus in which the labels are annotated at the turn level, the DM corpus is labeled at the call level, i.e., whether the call is GOOD or BAD. Our broad definition of a BAD call is either that the caller has trouble with the self-care or the caller opts out of the self-care for any reason. For BAD calls, therefore, a human agent should be brought in to help the caller. Following the automatic labeling approach [1], all calls have been automatically labeled as GOOD/BAD using a call-flow finite state machine, as summarized in Table 2.

Table 2: *DM corpus summary*

No. Calls	No. Turns	DM Label (No. Calls)	
		GOOD	BAD
1714	8554	676 (39.4%)	1038 (60.6%)

5. Experiments

5.1. ED test

We begin our experiments with testing ED: i.e., given input utterances (caller turns), the task is to classify each emotion as either NEGATIVE or ¬NEGATIVE. Similarly to [4], the following prosodic features are extracted from speech using Entropic ESPS xwaves `get_f0`[3] and used for testing ED:

- Pitch (f0): max, min, mean, standard deviation, median
- Energy (RMS) : max, min, mean, standard deviation, median
- duration, preceding pause, internal silence.

Notice that the data distribution is noticeably skewed as in Table 1, which *should*[4] be balanced so that they have equal priors. To this end, similarly to [13], we apply bagging [14] in a *class-conditioned* manner. That is, we first reserve a test set with equal priors (42 NEGATIVE & ¬NEGATIVE utterances, each). From the remaining data, we build 10 distinct training sets by randomly selecting data with equal priors (350 NEGATIVE & ¬NEGATIVE utterances, each), repeatedly. Then, we build 10 classifiers via the WEKA machine learning toolkit[5] using the 10 training sets and test the classifiers against the test set. Since there are 10 classifiers, each built from one training set, there are 10 test results which are then averaged or voted as below. To compare, two classifiers are built for ED, decision trees (classification model) and model trees (regression model), noted as bagged DT and bagged MT, respectively. As bagged DTs generate class labels as output, we simply take the majority vote of the 10 labels to yield the final classification label. Likewise, as bagged MTs generate scores (regression numbers), we take the average of the 10 scores and use a fixed threshold to determine each as NEGATIVE or ¬NEGATIVE.

Table 3: *Emotion detection performance*

ED Model	Class	Prec.	Rec.	F-score	Accuracy
Bagged DT	NEGATIVE	.667	.824	.737	.762
	¬NEGATIVE	.857	.720	.783	
Bagged MT	NEGATIVE	.767	.786	.776	.774
	¬NEGATIVE	.780	.762	.771	

Table 3 shows ED performance in terms of precision, recall, F-score, and accuracy. We first notice that model trees outperform decision trees, especially for NEGATIVE emotions as observed by a higher F-score. Also, though not fully satisfied with the overall performance ($< 80\%$ accuracies), we regard it as acceptable considering that 1) unlike other commonly used ED corpora, our corpus was labeled only once, meaning some emotion labels may be subjective, 2) emotions are often not explicit because our corpus is collected from a real-world application, and 3) most utterances are very short, as they are from system-directed, menu-style call dialogues, which makes it difficult to confidently decide the emotion. Notice that detecting NEGATIVE emotions performs comparably to detecting ¬NEGATIVE emotions—due to data balancing—which is favorable for DM.

[3]Freely available at http://www.speech.kth.se/speech/esps/esps.zip.

[4]Indeed, we initially tried building and testing models without balancing the data. This, however, resulted in classifiers with a very low recall in detecting NEGATIVE emotions—which is not desirable for DM—though they have a good accuracy. I.e., most of the time it classifies test data as ¬NEGATIVE.

[5]Freely available at http://www.cs.waikato.ac.nz/ml/weka/.

5.2. DM with call log features only

We next move on to testing our baseline DM system, i.e., DM with call log features only. The DM corpus has been first divided into training data (75%) and test data (25%). Then, the following features are extracted from call logs: the number of caller turns, NOMATCHES, NOINPUTS (timeouts), DTMF inputs, Yes/No answers, and finally speech recognizer's average confidence score. Next, two DM classification models, again, decision trees and model trees, are built using the training data and subsequently tested against the test data. Table 4 shows the baseline DM test results. As observed in the ED test (cf. Table 3), we have obtained a significant gain by using model trees (at the statistical significant test p-val < 0.05).

Table 4: *DM performance with call log features*

DM Model	Class	Prec.	Rec.	F-score	Accuracy
Decision Trees	BAD	.840	.904	.871	.841
	GOOD	.843	.749	.793	
Model Trees	BAD	.858	.904	.881	.855
	GOOD	.844	.788	.815	

5.3. DM with ED results only

As a first attempt to apply ED for DM, we use ED results directly for DM without considering anything else. As decision trees generate classification labels, decision trees are not used; only scores from ED model trees are used for DM. That is, we take ED regression scores from the DM test data, and then make decisions in terms of DM using a threshold.

Notice here that emotions are detected per each caller's turn while DM decisions are made per each call. E.g., if there are 5 turns in a call, there will be 5 ED scores while we need one DM decision for that call, meaning we need to somehow choose a single score from the 5 ED scores. We are mainly interested in caller's NEGATIVE emotions as they are a good indicator of caller dissatisfaction, and hence it stands to reason to pick the most NEGATIVE, i.e., the maximum ED score from each call as a feature for making a DM decision[6].

Unfortunately, this does not yield any comparable results to our baseline DM results. This is primarily due to the fact that some calls are BAD even though the callers are not emotionally aggravated. This experiment, however, confirms our hypothesis—NEGATIVE *emotions may be a sufficient condition for detecting* BAD *calls, but not a necessary condition*.

5.4. DM plus ED: feature level fusion

We continue our experiments with using ED results additionally for feature level fusion. As in §5.3, the maximum ED score from each call is extracted and combined with call-log features (feature level fusion) to build the classifiers. However, this again does not show any improvements over our baseline DM results. Our explanation for this is that even if a caller is emotionally aggravated, it may be equivalently detected by inspecting call logs. E.g., caller's NEGATIVE emotions may result in a low speech recognition confidence score, which corresponds to a NOMATCH. Indeed, feature analysis using principal component analysis shows that ED scores are not significant.

[6]We actually tried other features such as the average, range of the ED scores from a single call, and <max, average, range> as a vector, but all performed not better than the case with the maximum score.

Table 5: *Decision level fusion performance*

Decision Fusion	Class	Prec.	Rec.	F-score	Accuracy
Decision Trees	BAD	.864	.900	.882	.857
	GOOD	.845	.793	.818	
Model Trees	BAD	.863	.916	.888	.864
	GOOD	.866	.788	.825	

5.5. DM plus ED: decision level fusion

Finally, we try to fuse ED scores with call log-based DM scores at the decision level. That is, we first build two separate regression models, one for DM using call log features and the other for ED using prosodic features. We next generate regression scores from the DM training set—as we need another model for decision fusion which needs to be trained. The decision fusion model is then trained using those training data scores. After all, there are three models: (call log-based) DM model trees, ED model trees, and the decision fusion model for which we try both decision trees and model trees. Once the three models are built, we repeat the test procedure using the DM test data.

As Table 5 shows, we have obtained a modest improvement using decision level fusion. This is, in fact, the first result showing an improvement after using ED in addition to call log-based DM. Unfortunately, however, the improvement is not statistically significant (p-val ≈ 0.07) compared to the case without using ED. To find out the reason for such a small gain, we have randomly selected about 100 calls from the DM test set and have manually listened to them checking how often callers are emotionally aggravated. Interestingly, it turns out that only 9 calls (9%) have the emotional content which is much lower than the ED corpus (25%). This indicates that the ED corpus is much richer than the DM corpus in terms of emotions, which is not uncommon. In other words, in some application, callers are more often emotionally aggravated than in other applications. This is probably due to many factors such as the nature and/or complexity of the application. Conversely, this also implies that if we choose the ED corpus, label it in terms of DM, and test DM on that corpus, then we may be able to get a better result. Furthermore, notice that only 9% of the calls have the emotional content, which means the upper bound of the improvement using ED additionally is 9%. Despite such a small room for improvement, we have managed to obtain a modest gain, which is remarkable. Finally and more importantly, out of the 9 calls having the emotional content, 8 calls were already classified as BAD by the call log-based DM classifier. Supported by this observation, we can conclude that BAD calls having caller's NEGATIVE emotions may be identified by other call log features.

6. Conclusions

Going beyond simply detecting emotions at the utterance level, we have exploited possibilities of using ED practically for DM at the call dialogue level. Our findings and contributions, as summarized below, are fruitful. First, we have tackled ED as a regression problem using model trees, yielding gains in performance. This also enables us to easily use ED results for other purposes, e.g., DM. Second, we have confirmed that NEGATIVE emotions may be a sufficient condition for BAD calls, but not a necessary condition. Third, we have compared two alternatives to incorporate ED results for DM and demonstrated that decision level fusion is better than feature level fusion. Although our experiments have shown only a limited improvement after using ED, we attribute this to the fact that our corpus is not emotionally rich. Taking into account that the room for improvement is small, our gain is noteworthy. Finally, we have shown by our experiments that emotionally aggravated callers may be identified by other features. We are currently working on real-time online call dialogue monitoring to detect BAD calls as early as possible and incorporating ED for online monitoring as well.

7. References

[1] W. Kim, "Online call quality monitoring for automating agent-based call centers," in *Proc. of INTERSPEECH 2007–Eurospeech*, Aug. 2007, pp. 130–133.

[2] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. of ICSLP*, Sep. 2002, pp. 203–207.

[3] C. M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, March 2005.

[4] D. Litman and K. Forbes-Riley, "Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors," *Speech Communication*, vol. 48, no. 5, pp. 559–590, 2006.

[5] A. Kazemzadeh, S. Lee, and S. Narayanan, "Using model trees for evaluating dialog error conditions based on acoustic information," in *Proc. of ACM Int'l Workshop on Human-centered Multimedia*, 2006, pp. 109–114.

[6] A. Bosch, E. Krahmer, and M. Swerts, "Detecting problematic turns in human-machine interactions: Rule-induction versus memory-based learning approaches," in *Proc. of ACL*, July 2001, pp. 499–506.

[7] J. Hirschberg, D. Litman, and M. Swerts, "Prosodic and other cues to speech recognition failures," *Speech Communication*, vol. 43, no. 1-2, pp. 155–175, June 2004.

[8] I. Langkilde, M. Walker, J. Wright, A. Gorin, and D. Litman, "Automatic prediction of problematic human-computer dialogues in 'How may I help you?'," in *Proc. of ASRU*, Dec. 1999, pp. 369–372.

[9] M. Walker, I. Langkilde, H. Hastie, J. Wright, and A. Gorin, "Automatically training a problematic dialogue predictor for a spoken dialogue system," *Journal of Artificial Intelligence Research*, vol. 16, pp. 293–319, 2002.

[10] J. R. Quinlan, "Learning with Continuous Classes," in *5th Australian Joint Conf. on Artificial Intelligence*, 1992, pp. 343–348.

[11] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, CA, 1984.

[12] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. H. Witten, "Using model trees for classification," *Machine Learning*, vol. 32, no. 1, pp. 63–76, 1998.

[13] Y. Liu, N. Chawla, M. Harper, E. Shriberg, and A. Stolcke, "A study in machine learning from imbalanced data for sentence boundary detection in speech," *Computer Speech & Language*, vol. 20, no. 4, pp. 468–494, 2006.

[14] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.