Automatic Detection of Anger in Human-Human Call Center Dialogs

Mustafa Erden[1,2], *Levent M. Arslan*[1,2]

[1]Electrical and Electronics Engineering Department, BOĞAZİÇİ University, İstanbul, Turkey
[2]Sestek Inc., İstanbul, Turkey

(mustafa.erden|arslanle)@boun.edu.tr

Abstract

Automatic emotion recognition can enhance evaluation of customer satisfaction and detection of customer problems in call centers. For this purpose emotion recognition is defined as binary classification for angry and non-angry on Turkish human-human call center conversations. We investigated both acoustic and language models for this task. Support Vector Machines (SVM) resulted in 82.9% accuracy whereas Gaussian Mixture Models (GMM) gave a slightly worse performance with 77.9%. In terms of the language modeling we compared word based, stem-only and stem+ending structures. Stem+ending based system resulted in higher accuracy with 72% using manual transcriptions. This can be mainly attributed to the agglutinative nature of Turkish language. When we fused the acoustic and LM classifiers using a Multi Layer Perceptron (MLP) we could achieve a 89% correct detection of both angry and non-angry classes.

Index Terms: emotion recognition, human-human dialogs

1. Introduction

In call centers all dialogues between agents and clients are recorded for quality measurement and agent performance monitoring. However, because of additional costs and large amounts of accumulated data only a small fraction of those conversations are reviewed. If automatic emotion recognition is applied to this data problematic calls can be identified to some extent and retention operations can be performed for disgruntled customers.

Emotion recognition on real life call center data is a challenging task. First of all, the environmental conditions and microphone qualities are from a broad range. Also it is impossible to know the exact emotion of the speaker. Therefore subjective tests are necessary for labeling the data. Additionally real life data usually contains multiple emotions at the same time even the conflictual ones [1].

Previous studies on emotion recognition are done using different data sets with different classifiers and for different number of emotion categories. Some studies are performed on acted data [2, 3]. In [4], real emotions are simulated with a Wizard of Oz setup. In recent years many studies have focused on interactive voice response (IVR) system data in order to enhance human machine interaction [5, 6]. Also human-human dialogs are investigated [1, 7]. In [1], 82% correct classification between negative and positive emotions is achieved on French medical call center data using paralinguistic features. In [7], 56% detection is achieved for five emotion classes using acoustic features as well as information extracted from orthographic transcriptions of utterances.

Prosodic parameters are proven to be good correlates for emotions as well as being applicable to all kinds of data. However, linguistic parameters are only applicable to spontaneous data since content of acted data is predetermined. In order to capture lexical information, language models are trained [8], emotional salience of words are calculated [5, 9, 10] and classification with a bag of words approach [10] is implemented.

Since Turkish is an agglutinative language, words are generated by appending suffixes to roots. This morphological structure makes it difficult to build robust word-based language models. Also reasonable size dictionaries suffer from high out of vocabulary (OOV) words. The solution is using sub-word units instead of words. This type of language models are applied for automatic speech recognition applications [11]. Additionally stemming (removing suffixes from roots) is found to be successful for document retrieval [12]. To investigate the effect of sub-word models and stemming in emotion recognition task three different models are built for calculating language model scores; i) word based ii) stem-only iii) stem+ending.

In the second part of the work database used is presented. In the third part the classifiers and features are explained. Finally the results and discussion are given.

2. Database

The database used in this study consists of utterances of real agent-client dialogs recorded in 3 Turkish commercial call centers from finance, insurance and telecommunications sectors. The dataset to be labeled is filtered out of 300 hours of conversations. Agents have indicated possibly problematic calls. This remaining set comprised of 6 hours 13 minutes of 8 kHz, 8 bit mu-law encoded audio. Using a voice activity detection (VAD) module, these 385 dialogs are split into 8512 utterances. Then the utterances are separated into 4 non overlapping subgroups. Each subgroup is labeled by a different person. The labelers marked each utterance as angry, non-angry or garbage. Garbage label is assigned for turns such as silences, DTMF tones and overlapping speech. These data are not included in experiments. It was observed that the issues that caused anger were undelivered postage, mishandling of cancelation request, unauthorized transactions, problems with services and billing etc.

Data is separated into train and test sets as in Table 1. There is no overlap in terms of speakers between train and test sets.

Table 1: *Number of utterances for train and test sets.*

	angry	non-angry
Train set	1052	6531
Test set	146	783

Apart from 4 labelers 2 other labelers labeled the test set for measuring labeler-labeler agreement. Kappa statistic is used for this purpose as in [5, 6].

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \quad (1)$$

where $P(A)$ is the probability of agreement between labelers and $P(E)$ is the probability of agreement by chance. The kappa value being less or equal to 0 means no agreement and 1 means total agreement between the labelers. Kappa value between the two labelers is found to be 0.79 which indicates a high degree of conformity.

3. Classification

In this study three different classifiers are trained with three different sources of information.

3.1. Support Vector Machines and Acoustic Features

Support Vector Machines are discriminative classifiers which aim to find separating hyperplane with maximum margin between two classes. Libsvm library [13] is used for SVM modeling end testing. Radial Basis Function kernel is applied. For SVM experiments 143 features are extracted from each utterance. These are:

- MFCC parameters : 13 coefficients with deltas and delta deltas combines to 39 features. We extract three 39 dimensional vectors from each utterance. First vector is the average MFCC vector across all frames. Second vector comprises of minimum of each element in MFCC vector across the utterance. Third vector corresponds to maximum values. When these three vectors are concatenated we obtain a single 117 dimensional vector per utterance.
- Pitch parameters : min, median, max, first quartile, third quartile, mean and max of first derivative, inter voiced maximum difference, intra voiced maximum difference. Pitch contours are extracted by robust algorithm for pitch tracking (RAPT) [14]. Then z-normalized by shifting the mean to 0 and scaling the variance to 1, for eliminating speaker differences.
- Energy parameters : min, max, mean, standard deviation, mean and max of first derivative.
- Microprosody parameters : min, max, standard deviation of jitter and shimmer. Jitter and shimmer are calculated using a linear filter as used in [3].

During pitch contour calculations it is observed that for noisy recordings background speech contours were also included. To eliminate these errors a bias towards unvoiced decision is introduced.

3.2. Gaussian Mixture Models

Gaussian Mixture Models are representation of probability distributions as weighted average of several Gaussians. The likelihood contributed by i'th Gaussian component for a d dimensional vector $\vec{x}$ is

$$p(\vec{x}|\vec{\mu}_i, \mathbf{\Sigma}_i) = \frac{1}{(2\pi)^{d/2}\mathbf{\Sigma}^{1/2}} * \exp\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)'\mathbf{\Sigma}_i{}^{-1}(\vec{x} - \vec{\mu}_i)\} \quad (2)$$

where $\vec{\mu}_i$ is mean and $\mathbf{\Sigma}_i$ is covariance matrix of i'th Gaussian. The total likelihood for the given model is calculated as

$$p(\vec{x}) = \sum_{i=1}^{N} w_i p(\vec{x}|\vec{\mu}_i, \mathbf{\Sigma}_i) \quad (3)$$

where N is the number of Gaussians and w_i is the weight of Gaussian i. GMM is trained by Expectation-Maximization algorithm which iteratively increases the likelihoods until a predetermined convergence criterion is reached.

Gaussian Mixture Models for angry and non-angry model classes are trained with spectral features. Spectral features are extracted using HTK [15] toolkit. 13 Mel Frequency Cepstrum Coefficients are calculated for 25 ms Hamming windowed frames with 10 ms skip rate. With the addition of delta features a 26-dimensional vector for each frame is formed. We used 64 mixtures for each GMM.

Becars software [16] is used for GMM training end testing.

3.3. Language Model

For language modeling all utterances are manually transcribed. Only words and human noises are included in transcriptions. In order to split words into sub-word parts Morfessor [17] is used. Morfessor is an unsupervised morphological analyzer that segments words into *morphs* which are morpheme-like units. First morph of a word is labeled as a stem while remaining morphs are concatenated and labeled as endings. Also endings are marked by a special character for avoiding possible confusions. After creating stemmed and stem+ending forms the language model processing applied is the same as word model processing.

Table 2: *Language model train set lexical analysis.*

	Angry	Non-angry
# of words	10259	38390
# of distinct words	2985	6953
# of distinct stems	1922	3913
# of distinct morphs	2387	4788

For language modeling two different unigrams are trained for angry model and non-angry model. The classification decision is made by comparing the difference between the likelihoods with a threshold. Language model train set lexical analysis is given in Table 2. The training data is imbalanced with non-angry model containing at least two times that of distinct units in the angry model. This imbalance creates a disadvantage during likelihood calculations for the angry model. In order to compensate for this "add delta smoothing" between angry and non-angry language models is applied. In add delta smoothing unobserved words in the dictionary are assigned a very small probability. Delta value of 0.25 is found to be optimum in terms of classification accuracies in the train set.

Table 3: *Percentage of OOV for different language models.*

	Angry Model	Non-angry Model
word	29.85%	7.85%
stem-only	24.35%	6.39%
stem+ending	23.80%	6.24%

The difference in out of vocabulary percentages for different language models implemented are presented in Table 3. Call

center conversations use limited vocabulary and sentence forms. On the contrary for a general purpose Language model trained from a large Turkish corpus with stem+ending model OOV rate is 2.5% [11]. Therefore it will be interesting to investigate those three models in the context of emotion detection for call center conversations.

3.4. System Combination

System combination at the decision level is implemented in order to avoid dimensionality problems. A multi layer perceptron (MLP) is trained for this purpose. An MLP creates a mapping from given input vector to the desired output. It consists of nodes with nonlinear transfer functions. Outputs of nodes in a layer are weighted and added while being delivered to the next layer.

Training set utterance scores for SVM, GMM and LM classifiers are normalized and mapped to [-1 1] range where higher values imply higher probabilities of being angry. These scores are three inputs of MLP and the expected labels are the desired output. Using all 6531 utterances in the non-angry train set created a bias towards this class. To remove this bias only 1052 non-angry utterances are included for training which is the same number as angry utterances. A two layer MLP having three neurons in first layer and a single neuron in second layer applying gradient descent backpropagation training algorithm is found to be optimum for decision merging.

4. Experiments

Results for LM classifiers with different modelling units are given in Figure 1 as anger recall vs non-anger recall curves. When we compare the equal recall rates for anger and non-anger classes, word, stem-only, and stem+ending models resulted in 70.5%, 69.3% and 72.0% respectively. Therefore we can conclude that stem+ending based model performs better than the two other methods.

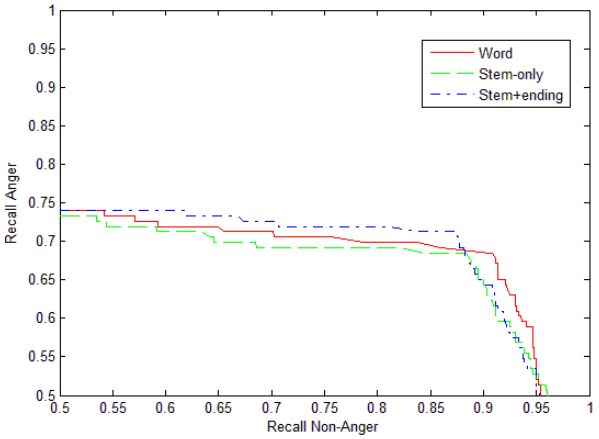

Figure 1: *Results for different LM classifiers.*

In Turkish, polite speaking style differs from impolite speaking style mostly by usage of different suffixes. An example for this difference is given in Table 4 for two words with two suffixes. The English meanings of words are given in parentheses. Since angry people tend to use less polite speaking style this is reflected in the choice of endings. Therefore removing endings results in degradation of the performance. Also frequencies of semantically similar endings are increased for stem+ending model which explains the performance superiority.

Table 4: *Turkish stemming example.*

Impolite Style	Polite Style
bırak+ın (leave it)	bırak+ınız (please leave it)
çıkart+ın (remove it)	çıkart+ınız (please remove it)
kapat+ın (close it)	kapat+ınız (please close it)

Results of different classifiers are given in Figure 2. Error rates for various operating points are calculated by applying a threshold to the differences between angry model scores and non-angry model scores.

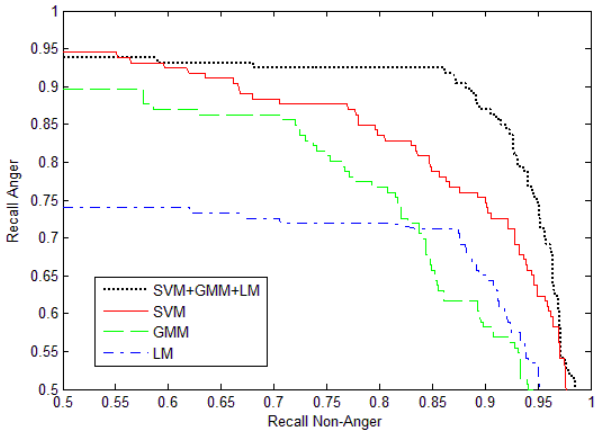

Figure 2: *Results for different classifiers.*

To measure the similarity between classifiers Q statistic is calculated as in [5].

$$Q = \frac{N_{11}N_{00} - N_{01}N_{10}}{N_{11}N_{00} + N_{01}N_{10}} \quad (4)$$

where N_{11} is the number of both classifiers being correct, N_{10} is the number of first classifier being correct and second classifier wrong, N_{01} is the number of first classifier being wrong and second classifier correct, N_{00} is the number of both classifiers being wrong. Absolute value of Q being closer to 0 means the classifiers are independent whereas higher values indicate higher similarities between the two classifiers.

Table 5: *Q statistic for classifiers.*

Q(SVM, GMM)	0.63
Q(SVM, LM)	0.05
Q(GMM, LM)	-0.16

The pairwise Q values for SVM, GMM, and LM classifiers are given in Table 5. SVM and GMM classifiers have a relatively high correlation as they both model acoustic characteristics. The Q values between LM classifier and others are lower which means linguistic channel contains complementary information to acoustic parameters. Additionally negative Q value

with GMM method indicate that correct classification is done for different samples of the test set [18].

5. Discussion

In order to extract linguistic information from utterances Unigram LM classifiers are implemented based on words, stem-only and stem+ending units. Stem+ending based model achieved accuracies higher than word based model because of the level of politeness conveyed through suffixes in Turkish. Additionally we observed that LM classifier could categorize utterances correctly which are misclassified by acoustic classifiers. As a result, after merging the scores of three classifiers by an MLP recognition accuracies of anger and non-anger increased considerably.

Language model anger recall values did not increase above 0.74 while non-anger values were above chance level. We can conclude that some of the utterances in the test set labeled as angry, does not contain any linguistic information which can favor angry model over non-angry model. In other words speakers may sometimes speak completely with neutral content even though they are emotionally disturbed. Therefore acoustic information is necessary for correct classification of these utterances.

6. Conclusion

In this paper different sources of information are utilized to recognize emotions in real-life human-human call center data. SVM classifier with prosodic and spectral features outperformed GMM with spectral features. The data set used contains noisy recordings which are recorded through various telephones. Features extracted for SVM classification are more robust to these conditions resulting in a better performance.

In [8], it is observed that stemming enhances lexical model performance on French human-human call center data. This is contrary to our findings which can be attributed to the different morphological structures of Turkish and French. While Turkish is mostly agglutinative, French show fusional aspects. In fusional languages for some words it is very difficult to segment to morphemes as they are fused together.

Human-human dialog data is rarely investigated in literature in terms of emotion recognition. A similar study to this one is presented in [1], which is on human-human call center data classifying between negative and positive emotions. We have achieved comparable accuracies with SVM classifier using acoustic features.

Future work includes considering the whole dialog while making a decision on utterance level as well as dialog level. Also the effect of using automatic speech recognizer (ASR) hypothesis instead of manual transcriptions for language modeling will be investigated.

7. Acknowledgements

This work is partially funded by TÜBİTAK within a TEYDEB project number 3070164.

8. References

[1] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers," in Proceedings of Interspeech, 2005.

[2] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, "Automatic Emotion Recognition using Prosodic Parameters", in Proc. of Interspeech, pp. 493-496, 2005.

[3] S. Yacoub, S. Simske, X. Lin and J. Burns, "Recognition of emotions in interactive voice response systems", In Eurospeech-2003, 729-732.

[4] T. Vogt, and E. Andre,"Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition", Proc. ICME 2005, Amsterdam,Holland.

[5] C. M. Lee, and S. Narayanan, "Toward detecting emotions in spoken dialogs", IEEE Trans. on Speech & Audio Processing, 13(2), 293-303, 2005.

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

[7] L. Vidrascu, and L. Devillers, "Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features", workshop Paraling07, 2007.

[8] L. Devillers, and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs", Interspeech 2006.

[9] T. Polzehl, S. Sundaram, H. Ketabdar, M. Wagner, and F. Metze, "Emotion Classification in Childrens Speech Using Fusion Acoustic and Linguistic Features", in Proc. Interspeech, Brighton, UK, 2009.

[10] F. Metze, A. Batliner, F. Eyben, T. Polzehl, B. Schuller, and S. Steidl, "Emotion Recognition using Imperfect Speech Recognition", Proceedings of Interspeech, pages 478–481, Makuhari, Japan, 2010.

[11] E. Arisoy, M. Saraclar, T. Hirsimäki, J. Pylkkönen, T. Alumäe, H. Sak, Fr. Mihelic, J. Zibert, (eds.), "Statistical Language Modeling for Automatic Speech Recognition of Agglutinative Languages" Speech Recognition : Technologies and Applications, Ch. 10, pp. 194–204, 2008

[12] F. Ç. Ekmekçioğlu and P. Willet. "Effectiveness of stemming for Turkish text retrieval", Program, 34(2):195200, April 2000.

[13] Chang, Chih-Chung and Lin, Chih-Jen., "LIBSVM: a library for support vector machines", 2001.

[14] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in Speech coding and synthesis (Elsevier, ed.), pp. 495518, 1995.

[15] S. Young, D. Ollason, V. Valtchev and P. Woodland, "The HTK Book" (for HTK Version 3.2), Entropic Cambridge Research Laboratory, 2002.

[16] R. Blouet, C. Mokbel, H. Mokbel, E. Sánchez Soto, G. Chollet, and H. Greige, "Becars: a free software for speaker verification", ODYSSEY '04, Spain, 2004.

[17] M. Creutz and K. Lagus, "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor", Publications in Computer and Information Science, Report A81, Helsinki University of Technology, March, 2005.

[18] L. Kuncheva, and C. Whitaker, "Measure of diversity in classifier ensembles," Mach. Learn., vol. 51, pp. 181207, 2003.