Automatic Recognition of Anger in Spontaneous Speech

Daniel Neiberg, Kjell Elenius

Centre for Speech Technology, CSC, KTH (Royal Institute of technology), Stockholm, Sweden

neiberg@speech.kth.se, kjell@kth.se

Abstract

Automatic detection of real life negative emotions in speech has been evaluated using Linear Discriminant Analysis, LDA, with "classic" emotion features and a classifier based on Gaussian Mixture Models, GMMs. The latter uses Mel-Frequency Cepstral Coefficients, MFCCs, from a filter bank covering the 300-3400 Hz region to capture spectral shape and formants, and another in the 20-600 Hz region to capture prosody. Both classifiers have been tested on an extensive corpus from Swedish voice controlled telephone services. The results indicate that it is possible to detect anger with reasonable accuracy (average recall 83%) in natural speech and that the GMM method performed better than the LDA one.

Index Terms: spontaneous speech, natural emotions, anger

1. Introduction

Emotional corpora may be categorized into three classes: acted, induced and natural [1][2][3]. Acted emotion corpora are often based on recordings by actors in studio conditions. Induced speech is speech where the emotions are induced by exposing the speaker to emotionally loaded material such as videos, slides or text [3]. Natural emotional speech is spontaneous speech where the emotions are authentic, and not manipulated in any way. While the collection of acted emotional speech is straightforward and can give an even distribution for various emotional categories, the collection of natural speech generally results in a relatively low percentage of emotional data. Also, the distribution of emotions may be greatly skewed, depending on various conditions. A difficulty in the study of real-life speech is that it is hard to exactly know what emotions the speakers were actually experiencing.

An influential theory in emotion research is the discrete emotion theory, which is based on the belief that there are a small number of basic emotions, such as anger, fear, joy, sadness, surprise and disgust [5]. These emotions can to a large extent be communicated cross-culturally, although there is a cultural effect so expressions from one's own culture are slightly better recognized, see, e.g., Laukka [6]. Laukka also shows that the discrete emotions are associated with relatively distinct acoustic categories. Thus, for our purposes, we rely on the following assumption, Schrerer [1]: "*The basis of any functionally valid communication of emotion via vocal expression is that different types of emotions are actually characterized by unique patterns or configuration of acoustic cues*".

The two major measures for detecting emotions in speech are acoustic and verbal. Acoustic parameters are usually measured on a frame basis, e. g. every 10th millisecond. These measures may be used directly as features in models capable of classifying sequences of data, such as GMMs and Hidden Markov Models, HMMs [7]. Alternatively, low order statistics on a segment level may be computed to give a feature vector per segment which is fed to a classifier, using LDA, Support Vector Machines or decision trees. A segment may, e.g. be an utterance, a word or a syllable, and typical statistics used are minimum, maximum, mean, range and standard deviation of an acoustic parameter [3].

A standard approach to emotion recognition is to first extract a large number of features derived from measures of fundamental frequency, intensity and formants and then use a vector classifier to search for a smaller set of relevant features [7]. While this approach performs quite well on acted data, results on induced and natural data are less promising [9]. Since the extraction of acoustic features targets specific speech acoustics considered to be influenced by emotions, this approach may be seen as knowledge based. It is also common enough to be considered as a baseline.

A general problem with acoustic features is that they are speaker dependent, which, however, may be overcome to a large extent by speaker adaptation. Other measures, such as words and human noises are the output from speech recognition systems, while linguistic features such as parts of speech (PoS), may be produced by a natural language parser. Although it seems likely that human noises, such as sighs, laughter and frustration (swear) words [10][11] are correlated with emotion, it has also been shown that measures of repetitions, PoS and Dialog Acts [7][9] may be useful.

As it is widely known that prosody is dependent on context, it is reasonable to expect additional benefits from using word or context dependent models for acoustical features. A benefit of using acoustic features as opposed to linguistic features is that they can be used for classification before an utterance is finished. A most important aspect in human-computer interaction is the naturalness of the interface. By adopting natural human-human interaction as a gold standard in spoken dialog systems, maximum efficiency and user satisfaction is assumed to be achieved [12]. As found in the literature [9][11], emotions in spoken dialogs is not limited to human-human interaction, but is also found in human-computer interaction.

By letting a dialog system be aware of the user's emotions, a proper response may be generated. If the user shows frustration or anger, it is likely that it is due to communication problems. It may be a recognition error, a dialog system inflexibility or user inability to use the service. In this case, the dialog system may forward the user to an operator or enter a specific mode designed to resolve the problem.

In this work, we first present an extensive corpus of natural emotions from human-computer interaction. Then two methods are described and evaluated. The first is a classical knowledge based approach based on per syllable statistics and a LDA classifier. The second is a data driven approach based on broad spectral features, MFCCs, and GMMs.

2. The Voice Provider Corpus

The speech material in our study is telephone speech recorded at 8 kHz by the Swedish company Voice Provider that runs more than 50 different voice-controlled telephone services covering information regarding airline and ferry traffic, postal assistance and a lot more. The large majority of utterances are neutral (non-expressive), but some percent show frustration, often after misrecognitions by the speech recognizer. Limited parts of this corpus has been used in other studies [13][14][15].

2.1. Annotation

The utterances are labeled by an experienced, senior voice researcher into neutral, emphasized or negative (frustrated) speech. When labeling a speaker's dialog it is at times obvious that the speaker is emphasizing, or hyper-articulating, an utterance rather than expressing frustration. This is, however, not obvious without taking the dialog context into account. Other emotions than frustration are very rare and therefore not meaningful to handle. Since the utterances are recorded from a real-life telephone services the emotions expressed must be regarded as natural. As mentioned, this means that labeling them is not straightforward, since it is impossible to be sure about what emotion (if any) the speaker was expressing. However, our results indicate that the decisions of the annotator were consistent with some expressive content across all utterances. Also, a subset of the material was labeled by five different speech researchers and the pair-wise inter-labeler kappa was 0.75 - 0.80.

2.2. Description of the Corpus

The corpus consists of 20,807 dialogs with 61,078 utterances giving an average number of 2.9 turns per dialog. The median is 2 turns and the maximum is 206 turns. The proportion of neutral utterances is 96.1%, while the number of negative and emphatic utterances is 2.2% and 1.7% respectively (Table 1). Thus, emotional utterances are rare in our data, which may be seen as an indication that the recorded services work mostly well. The length of the dialogs in turns is shown in Figure 1. A dialog "emotion" is classified as follows: neutral if all turns are neutral, negative if one or more turns are negative, and otherwise emphatic. The proportion of neutral, negative and emphatic dialogs (relative to all neutral, negative and emphatic dialogs respectively) with length 1 to 20 turns is shown in Figure 1. The negative and emphatic dialogs increase to turn 3 after which they decrease. This is most probably because the callers become frustrated and decide to hang up.

Note that the curves for negative and emphatic are very similar. This indicates that when there is a problem people will either speak louder or get frustrated. Of the negative dialogs 5% are always negative and 12% start with a negative turn. Longer connected sequences of negative turns are rare; 77% of the negative dialogs have one negative turn, 51% end with negative, 11% end with 2 negative and 4% end with 3 negative followed by hang-up, 35% become negative and calm down to neutral, 30% are preceded by an emphatic turn somewhere in the dialog, and 13% have an emphatic turn immediately before a negative one. Although there is a tendency for users to hang up after getting angry their behavior spans over a wide range of patterns.

Development set	Value	Percent
Neutral	38229	95,8
Emphatic	685	1,7
Negative	977	2,5
Total	39891	
Evaluation set	**Value**	**Percent**
Neutral	20445	96,5
Emphatic	370	1,7
Negative	372	1,8
Total	21187	
Total	**61078**	

Table 1: *Utterances in the Voice Provider corpus.*

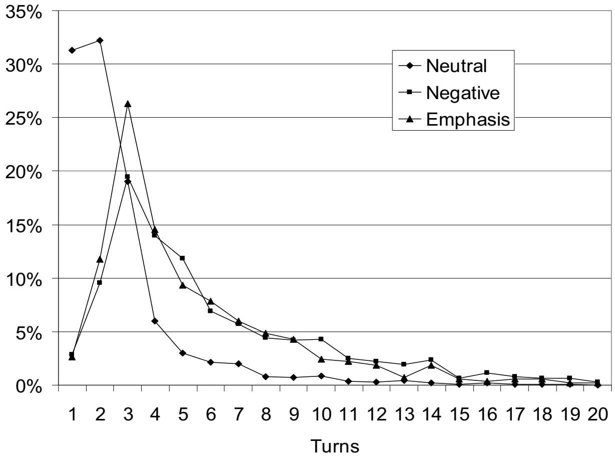

Figure 1: *Distribution of neutral, negative and emphatic dialogs. Since neutral utterances constitute 96.1 % of all the distributions are normalized as described in section 2.2.*

3. System Descriptions

The approach adopted in this work is to set up a baseline emotion classification system using knowledge based features and LDA and compare it to a GMM based system using broad spectral features, MFCCs.

3.1. System 1: A Baseline System

In our baseline system we calculate "classic" features used in emotion recognition. They are calculated on pseudo-syllabic segments, defined as a sequence of unvoiced, voiced, unvoiced segments with intensity and duration above certain thresholds. Finally we use the average value of the features calculated over all segments of an utterance.

The features can be categorized into the following broad classes: F0, intensity, formants, voice source and duration. For the first two we used minimum, maximum, mean, range, quantiles 1-5, slope, frame-wise delta, frame-wise delta-delta, standard deviation, the relative position of the minimum and the maximum, and percentage of frames with rise/static/fall. For F0 also quantiles 1-5, standard deviation and range of F0 were extracted after first subtracting the average slope of F0 over the segments. The formant features used are mean, standard deviation and median bandwidth for formants 1 to 4. The duration features are silence, mean duration, mean syllable duration and percentage voiced.

For the voice source features the following spectral approximations are used: *Open Quotient*: F0 amplitude - 2nd F0 harmonic amplitude; *Glottal opening:* F0 amplitude - F1 amplitude; *Amplitude of Voicing:* F0 amplitude; *Rate of closure:* F0 amplitude - F3 amplitude; *Skewness:* F0 amplitude - F2 amplitude; *Completeness of Closure:* F1 mean bandwidth; *log Approximate Normalized Amplitude Quotient:* F3 amplitude + 20*log (F0 mean) [16].

In addition, jitter and shimmer where added to the voice source measures. Jitter is measured as the average absolute difference between consecutive differences for consecutive periods, divided by the average period. Shimmer is measured as the average absolute difference between consecutive differences for amplitudes of consecutive periods.

All features are extracted using Praat [17], where the amplitudes used for the voice source features were measured using the Long Time Amplitude Spectrum (LTAS), without any corrections for F0 or formants. These "classic" features are modeled using LDA, but without taking the prior distribution of emotional categories into account.

3.2. System 2: A GMM and MFCC System

A standard filter bank in the 300-3400 Hz region is used. The signal is pre-emphasized and a FFT is calculated every 10 ms using a 25.6 ms Hamming window. 24 Mel-warped logarithmic filter bank coefficients are cosine transformed to 12 dimensions, followed by a RASTA-processing (position of pole 0.94), and appended with the 0'th component. Finally, delta features and delta-delta features are added, resulting in a 39 dimensional feature vector.

Another filter bank in the 20-600 Hz region is used to model prosody. It is computed similarly as above. This novel method was first reported in [13]. Optimizations on our development set showed that 48 filters performed better than 24, but a projection to 24 cepstral dimensions did not yield any significant improvement over 12.

GMMs with 512 Gaussians per emotion category and filter bank are used according to a procedure described in [13]. The log-likelihoods of the GMMs were combined using linear addition.

4. Experiments

The corpus was split into three sets with roughly equal distribution of emotion categories. Two sets are used for development and one for evaluation. The development sets are used for cross-wise training and testing. The speaker identities were not known and splits were made between dialogs. Since the LDA model is computationally inexpensive, a brute force forward selection scheme to find the most discriminative features was adopted. Essentially, each feature was tested and the one that maximized average recall was kept in an incremental way. Each part of the development set was used as training and test in a cross-wise fashion, resulting in two ranked lists of features. Finally a joint list was created, where the rank of each feature was calculated as the weighted mean rank from the two lists. The feature order in the joint list was then used to add features one by one in cross-wise tests on the development set. As the features were added incrementally, the number of necessary features was determined as the number where maximum of average recall was achieved. Finally, a LDA model was trained on the full development set using these features and tested on the evaluation set.

Our first experiments showed that with our methods it was utterly difficult to distinguish between the emphatic class we had defined and the neutral class, and therefore the emphatic class was merged with the neutral.

Since some key features, such as F0 range and mean used in System 1 may be speaker dependent, we also did speaker adaptation experiments by subtracting the mean of each feature estimated on one neutral utterance per dialog. As some dialogs only have one turn, or start with non-neutral emotions, a slightly different split was made for this particular experiment in order to make sure that at least one neutral utterance existed per dialog. Average recall is used as performance measure.

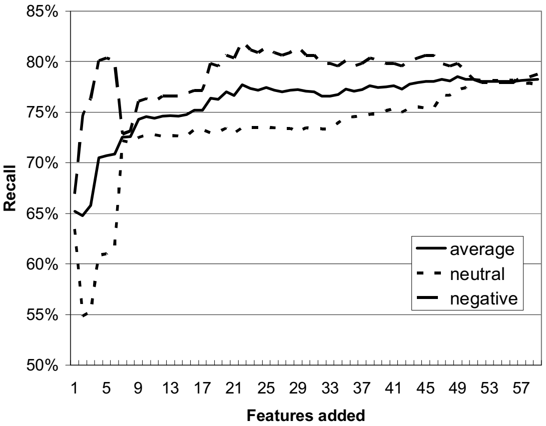

Figure 2: *Performance on the evaluation set as a function of the best added features estimated on the evaluation set.*

5. Results

For System 1, a maximum average recall rate of 71% was achieved on the development set with the 19 best ranked features achieved with the method described above, see Table 2. Recall rates with these ranked features on the evaluation set are shown in Figure 2. The curves are similar to those of the development set, but there is no peak at 19 features, which was not unexpected.

1. mean delta-delta F0
2. mean delta intensity
3. mean delta F0
4. logANAQ
5. mean delta-delta intensity
6. fraction F0 rise
7. syllable mean duration
8. relative position of intensity maximum
9. silence mean duration
10. F1 std.
11. F4 bandwidth median
12. fraction voiced
13. fraction intensity rise
14. F3 mean
15. F1 mean
16. intensity range
17. F2 mean
18. intensity quantile 5
19. F2 bandwidth median

Table 2: *Ranked list of the best features for the development set.*

The confusion matrices for System 1 and System 2 are shown in Table 3. Separate results for the feature sets of System 1 are shown in Table 4. The performance for the high and low filter banks of System 2 are shown in Table 5. Speaker adaptive results yielded only 63% average recall on the development set, and was considered a failed approach.

6. Discussion

The results for various System 1 feature sets in Table 4 show that the voice set (voice source and formants) performs best, followed in order by intensity, duration and F0. This order may be seen as an indication of how sensitive the respective feature sets are to anger; the more sensitive the better the classification. Note that the 19 best features determined on the development set, see Table 2, were better than each of the feature sets.

The low performance for speaker adaptation may be explained by the fact that only one, often short, utterance may not provide a good enough estimate for the normalization.

From Table 3 it is obvious that both System 1 and 2 have a discriminative ability, but with System 2 being better.

Regarding System 2, the two filter banks used seem to contain an almost equal amount of discriminative information, see Table 5. The somewhat better performance for the low filter bank indicates that the prosodic information it was designed for is important for our classification task.

A couple of studies have classified speech into angry/frustrated or neutral and report classification accuracy of around 70-80%, which is comparable with the accuracy of the classification of negative emotions in the present study (e.g., [10][18]. The improvement compared to the previous approach where a subset of the same corpus was used [13], may be explained by the larger numbers of utterances for each category, and by the more robust way of combining classifiers reported here.

7. Conclusions

Based on these results, it is our opinion that it is possible to detect anger with reasonable accuracy in natural speech. Specifically, it is shown that a classifier based on a filter-bank in the 20-600 Hz region has as good performance as a filter bank in the 300-3400 Hz region.

8. Acknowledgements

This work was performed within CHIL, Computers in the Human Interaction Loop, an EU 6th Framework IP (506909). We thank Voice Provider for providing the speech material and Inger Karlsson for her invaluable help in its annotation.

	System 1: LDA		System 2: GMM	
	Neutral	Negative	Neutral	Negative
Neutral	0.73	0.27	0.87	0.13
Negative	0.20	0.80	0.22	0.78

Table 3: *Confusion matrix with recall rates for the evaluation set.*

Feature set	Average recall	Neutral recall	Negative recall
Voice	0.74 ± 0.012	0.70 ± 0.003	0.78 ± 0.024
Intensity	0.72 ± 0.012	0.66 ± 0.003	0.77 ± 0.022
Duration	0.69 ± 0.012	0.73 ± 0.003	0.65 ± 0.021
F0	0.68 ± 0.012	0.70 ± 0.003	0.67 ± 0.025
19 Best	0.76 ± 0.012	0.73 ± 0.003	0.80 ± 0.021

Table 4: *System 1 feature set results on the evaluation set with one standard deviation.*

Filter bank	Average recall	Neutral recall	Negative recall
300-3400	0.80 ± 0.011	0.85 ± 0.002	0.75 ± 0.022
20-600	0.83 ± 0.010	0.85 ± 0.002	0.81 ± 0.020
Both	0.83 ± 0.011	0.87 ± 0.002	0.78 ± 0.021

Table 5: *System 2 feature set results on the evaluation set with one standard deviation.*

9. References

[1] Scherer, K. R. “Vocal communication of emotion: A review of research paradigms,” Speech Communication, vol. 40, pp. 227– 256, 2003.

[2] Campbell, N. “Databases of emotional speech,” in Proc. of the ISCA Workshop on Speech and Emotion, Northern Ireland, 2000, pp. 34–38.

[3] Ververidis, D. and Kotropoulos, C. “Emotional speech recognition: Resources, features, and methods,” Speech Communication, vol. 48, no. 9, pp. 1162–1181, September 2006.

[4] Velten, E. “A laboratory task for induction of mood states,” Behavior Research and Therapy, no. 6, pp. 473–482, 1968.

[5] Ekman, P. “An argument for basic emotion,” Cognition and Emotion, vol. 6, no. 169-200, 1992.

[6] Laukka, P. “Vocal expression of emotion: Discrete-emotions and dimensional accounts,” Ph.D. dissertation, Uppsala universitet, 2004.

[7] Nwe, T. L., Foo, S. W. and De Silva, L. C. “Speech emotion recognition using Hidden Markov Models,” Speech Communication, vol. 41, no. 4, pp. 603–623, November 2003.

[8] Schuller, B., et al. “The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals,” in Proceedings of Interspeech, Antwerp, Belgium, August 2007.

[9] Batliner, A., Fischer, K., Hubera, R., Spilkera, J. and. Nöth, E. “How to find trouble in communication,” Speech Communication, vol. 40, pp. 117–143, 2003.

[10] Ang, J., Dhillon, R., Krupski, A., Shriberg, E. and Stolcke, A. “Prosody-based automatic detection of annoyance and frustration in human-computer dialog” in ICSLP, Denver, 2002.

[11] Chul, M. L. and Narayanan, S. “Toward detecting emotions in spoken dialogs,” IEEE, Transactions on Speech and Audio Processing, vol. 13, no. 2, pp. 293–303, March 2005.

[12] Edlund, J., Heldner, M., Hjalmarsson, A., and Gustafson, J. (in press). Towards human-like spoken dialogue systems. Speech Communication: Special Issue Spoken Dialogue Technology.

[13] Neiberg, D., Elenius, K. and Laskowski, K. “Emotion recognition in spontaneous speech using GMMs,” in Proceedings ICSLP-2006, Pittsburgh, 2006, pp. 809–812.

[14] Forsell, M., Elenius, K., and Laukka, P. (2007). Acoustic correlates of frustration in spontaneous speech. Proceedings of Fonetik, TMH-QPSR, 50(1), 37-40.

[15] Laukka, P., Elenius, K., Fredriksson, M., Furumark, T., and Neiberg, D. (in press). Expression in spontaneous and experimentally induced affective speech: Acoustic correlates of anxiety, irritation and resignation. Proc. LREC Workshop on Corpora for Research on Emotion and Affect. Marrakech, Marocko, 2008.

[16] Toubol, E. “Recognition of Emotions, Automatic extraction of spectral correlates from the glottal source, Master Thesis, Speech, Music and Hearing, 1993.

[17] Boersma, P. and Weenink, D. Praat: Doing phonetics by computer. [Online]. Available at: http://www.fon.hum.uva.nl/praat/

[18] Burkhardt, F., Ajmera, J., Englert, R., Stegmann, J., & Burleson, W. Detecting anger in automated voice portal dialogs, in: Proc. 9th International Conference on Spoken Language Processing, Pittsburgh, 2006.