Emotion Recognition in Spontaneous Speech Using GMMs

Daniel Neiberg[1], *Kjell Elenius*[1] *and Kornel Laskowski*[2]

[1] Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
[2] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[neiberg, kjell]@speech.kth.se, kornel@cs.cmu.edu

Abstract

Automatic detection of emotions has been evaluated using standard Mel-frequency Cepstral Coefficients, MFCCs, and a variant, MFCC-low, calculated between 20 and 300 Hz, in order to model pitch. Also plain pitch features have been used. These acoustic features have all been modeled by Gaussian mixture models, GMMs, on the frame level. The method has been tested on two different corpora and languages; Swedish voice controlled telephone services and English meetings. The results indicate that using GMMs on the frame level is a feasible technique for emotion classification. The two MFCC methods have similar performance, and MFCC-low outperforms the pitch features. Combining the three classifiers significantly improves performance.

1. Introduction

Recognition of emotions in speech is a complex task that is furthermore complicated by the fact that there is no unambiguous answer to what the “correct” emotion is for a given speech sample. The vocal emotions explored may have been induced or acted or they may be have been elicited from more “real”, life-like contexts [1], [2]. Spontaneous speech from actual telephone services could be counted as such a material. The line of emotion research can roughly be viewed as going from the analysis of acted speech [3] to more “real” [2], [4], [5]. The motivation of the latter is often to try to enhance the performance of human-machine interaction systems, such as voice controlled telephone services.

A difficulty with spontaneous emotions is in their labeling, since the actual emotion of the speaker is almost impossible to know with certainty. Also, emotions occurring in spontaneous speech seem to be more difficult to recognize compared to acted speech [2]. In [6], a set of 6 features selected from 200 is claimed to achieve good accuracy in a 2-person corpus of acted speech. This approach is adopted by several authors. They experiment with large numbers of features, usually at the utterance level, and then rank each feature in order to find a small golden set, optimal for the task at hand [7].

Classification results reported on spontaneous data are sparse in the literature. In [5], the corpus consists of recordings of interactions between users and an automatic voice service. It is reported that performance flattens out when 10 out of 60 features are used in a linear discriminant analysis (LDA) cross-validation test. The performance is improved by introducing corpus reduction procedures. In [4], data is recorded from various air travel booking systems. The authors report good performance for discrimination between *frustration* versus *else* and *annoyed* + *frustration* versus *else*. However, they rely heavily on language model features, repetitions and manual annotation of speaking style. In [8], data from a commercial call center was used. As is frequently the case, the results for various acoustic features were only slightly better than a system classifying all exemplars as neutral. There is an indication that the golden set of acoustic features has yet to be found, or simply does not exist (without conditioning on other sources of information). As mentioned, some authors use several hundred features per utterance, which means that almost the entire spectrum is covered. The natural continuation is simply to use spectral features, such as Mel-frequency (MFCC) or linear prediction (LPCC) cepstral coefficients, possibly with additional pitch measures. Delta MFCC measures on the utterance level has been used earlier, e.g. in [6], [9]. However, we have chosen to model the distribution of the MFCC parameters on the frame level in order to obtain a more detailed description of the speech signal.

In spontaneous speech the occurrence of canonical emotions such as happiness and anger is typically low. The distribution of classes is highly unbalanced, making it difficult to measure and compare performance reported by different authors. The difference between knowing and not knowing the class distribution will significantly affect the results. Therefore we will include results from both types of classifiers in our results.

2. Material

We have used two different spontaneous speech corpora for our experiments as described below.

2.1. The Voice Provider material

The first material used in our study was recorded at 8 kHz at the Swedish company Voice Provider, called VP in the following, which runs more than 50 different voice-controlled telephone services. The services cover information regarding airlines, ferry traffic, postal assistance and much more. Most utterances are neutral (non-expressive), but some percent are frustrated, most often due to misrecognitions by the speech recognizer. The utterances are labeled by an experienced, senior voice researcher into neutral, emphasized or negative (frustrated) speech. When labeling a speaker’s

Table 1: *The Voice Provider Material*

Development set		
Neutral	3865	94 %
Emphatic	94	2 %
Negative	171	4 %
Total	4130	
Evaluation set		
Neutral	3259	93 %
Emphatic	66	2 %
Negative	164	5 %
Total	3489	

dialogue 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 dialogue context into account. This is our reason for introducing the emphatic label, since these utterances should not be counted as frustrated, if possible. A subset of the material was labeled by 5 different speech researchers and the pair-wise inter-labeler kappa was 0.75 – 0.80.

The VP material was split into a development and evaluation set according to the proportions in Table 1. In order to ensure that enough data for negative and emphatic classes are available in both sets, utterances were sampled from the entire corpus, but no sessions were split between the development and evaluation set.

2.2. The ISL Meeting Corpus

In addition to the VP data, we apply our approach to meeting recordings. The ISL Meeting Corpus consists of 18 meetings, with an average number of 5.1 participants per meeting and an average duration of 35 minutes. The audio is of 16 bit, 16 kHz quality, recorded with lapel microphones. It is accompanied by orthographic transcription, and, more recently, annotation of a three-way emotional valence (*negative*, *neutral*, *positive*) at the speaker contribution level [10]. For our purposes, this corpus was split into a development set and an evaluation set, as shown in Table 2. The emotion labels were constructed by majority voting (2 of 3) for each segment. Split decisions (one vote for each class) were removed. Finally, the development set was split into two subsets that were used for cross-wise training and testing.

Table 2: *The ISL Meeting Corpus*

Development set		
Neutral	6312	80 %
Negative	273	3 %
Positive	1229	16 %
Total	7813	
Evaluation set		
Neutral	3259	70 %
Negative	151	3 %
Positive	844	19 %
Total	4666	

3. Features

Three main sets of features are described; 1) standard MFCCs 2) MFCC-low using filters from 20 Hz to 300 Hz 3) pitch.

3.1. Mel-frequency Cepstral Coefficients

Mel-frequency Cepstral Coefficients, MFCCs, are extracted using pre-emphasized audio, using a 25.6 ms Hamming window every 10 ms. For each frame, 24 FFT-based Mel-warped logarithmic filter bank coefficients from 300 to 3400 Hz are extracted and then cosine transformed to 12 dimensions, followed by RASTA-processing (position of pole 0.94) [11], and appending of the 0'th component, which corresponds to energy. Finally, the delta (computed from 2 frames backward and forward) and delta-delta (computed from 2 frames backward and forward from delta) features are added, resulting in a 39 dimensional feature vector. For the 16 kHz ISL data, we use 26 filters from 300 to 8000 Hz; otherwise the processing is identical.

3.2. MFCC-low

These features are computed in the same way as MFCCs but the filter banks are placed in the 20 - 300 Hz region instead. We expect these low frequency MFCCs to model F0 variations. Three different frame size and frame rate combinations were tested: (1) 25.6 ms frames every 10 ms; (2) 64 ms frames every 25 ms; and (3) 128 ms frames every 50 ms.

3.3. Pitch and derivative

The algorithm for pitch tracking uses the Average Magnitude Difference Function (AMDF) due to [12]. The variant used here was introduced in [13]. Pitch is extracted on a logarithmic scale and the utterance mean is subtracted. Finally, delta features are added. We also tried out delta-delta features, but several numerical problems occurred in the modeling stage and we ultimately abandoned delta-delta features. In initial tests, we also tried not to subtract the utterance mean, but the result was worse compared to performing the subtraction.

4. Classifiers

All acoustic features are modeled using Gaussian mixture models (GMMs) with diagonal covariance matrices measured over all frames of an utterance. First, using all the training data, a root GMM is trained with the Expectation Maximization (EM) algorithm with a maximum likelihood criterion, and then one GMM per class is adapted from the root model using the maximum a posteriori (MAP) criterion [14]. MAP adaptation protects against overtraining, and removes the need to optimize the number of Gaussians per class which may be necessary due to differences in the amount of available training data per class. We use 512 Gaussians for MFCCs and 64 Gaussians for pitch features. These numbers were optimized empirically in initial tests. This way of using GMMs has proved successful for speaker verification [15].

In addition to acoustic features, we also used average log-likelihoods of n-grams using manual orthographic transcriptions for the training and test data. Only human noises and words from the transcriptions were used. N-grams without human noises were also explored but the result was no better than random. The SRILM toolkit was used for n-gram modeling [16].

4.1. Classifier combination

The output from multiple classifiers was combined using multiple linear regression, with the final class selected as the argmax over the per-class least square estimators. The transform matrix was estimated from the training data...

5. Experiments

We ran our experiments with the features and classifiers described above. We also used combinations of them. The acoustic combination was composed by the GMM modeling MFCC, the best MFCC-low GMM for the particular corpus, and pitch GMM. The combination matrix was estimated by first testing the respective GMM with its training data. For the ISL corpus, a joint result was constructed from the two development subsets by concatenating the outputs of each estimated GMM on its training data. The result was used for optimizing the combination matrix, which was used for the cross-wise test of the devel-

opment set and also for the evaluation set. This should provide a more robust estimate for the matrix since it relies on the output of two classifiers. For the evaluation set runs, a new GMM was trained on the entire development set.
The performance for the n-gram experiments flattened out for *n=3* and therefore we do not report lower order n-grams. Various minimum count pruning methods where also tried out, and even though the performance improved a little, these features were not used because they performed poorly during initial tests on the training data.

6. Results

Performance is measured as absolute accuracy, average recall (for all classes) and f1, computed from the average precision and recall for each classifier. The results are compared to two naïve classifiers: a random classifier that classifies everything with equal class priors, *random with equal priors*, and a random classifier knowing the true prior distribution over classes in the training data, *random using priors*. The combination matrix, see section 4.1., accounts for the prior distribution in the training data, making the neutral class heavily favored. Therefore a weight vector which forces the matrix to normalize to equal prior distribution was also used. Regarding this we report two more results: *acoustic combination* with equal priors, that is optimized for the accuracy measure and *acoustic combination using priors*, which optimizes the average recall rate. Thus, all classifiers in our tables under the *random equal priors* heading do not know the a priori distribution of the classes and should only be compared to each other. The same holds for the classifiers under the heading *random using priors*. Note that the performance difference in percentages is higher for a classifier not knowing the prior distribution compared to the corresponding random classifier, than for the same classifier knowing the prior distribution compared to its corresponding random classifier. This is due to the skewed prior distributions.

Table 3: *Voice Provider results for Neutral vs. Emphasis vs. Negative. Accuracy, Average Recall and f1.*

Classifier	Acc.	A. Rec.	f1
Random with equal priors	*0.33*	*0.33*	*0.33*
MFCC	0.80	0.43	0.40
MFCC-low 10 ms	0.78	0.39	0.37
Pitch	0.56	0.40	0.38
Acoustic combination	0.90	0.37	0.39
Random using priors	*0.88*	*0.33*	*0.33*
Acoustic comb. using priors	0.93	0.34	0.38

Table 4: *Voice Provider results for Neutral and Emphatic vs. Negative. Accuracy, Average Recall and f1.*

Classifier	Acc.	A. Rec.	f1
Random with equal priors	*0.50*	*0.50*	*0.50*
Acoustic combination	0.95	0.52	0.56
Random using priors	0.92	0.50	0.50
Acoustic comb. using priors	0.95	0.51	0.60

Table 5: *Voice Provider results for Neutral vs. Emphatic and Negative. Accuracy, Average Recall and f1.*

Classifier	Acc.	A. Rec.	f1
Random with equal priors	*0.50*	*0.50*	*0.50*
Acoustic combination	0.92	0.54	0.56
Random using priors	0.88	0.50	0.50
Acoustic comb. using priors	0.93	0.52	0.58

6.1. Voice Provider results

From Table 3 we note that all classifiers with equal priors perform substantially better than random. The results for the different MFCC-low tests were not very different. Only the 10 ms case is reported, which gave best results for the training set. The MFCC-low classifier is almost as good as the standard MFCC one and it performs considerably better than the pitch based GMM.
In Tables 4 and 5 the emphatic class has been combined either with the neutral or the negative class resulting in binary classification. The small differences between these tables suggest that ignoring the dialogue context makes it hard to differentiate between emphatic and negative speech, since their acoustic manifestations seem to be similar, compare section 2.1.
One can also note that combining all the three acoustic classifiers gives the best overall results, although the effect is not as pronounced when using priors, as discussed above.

6.2. ISL meeting corpus results

Table 6 shows our results for the ISL meeting corpus. Among the MFCC configurations in the low frequency region, MFCC-low 25 ms performed best for the development set. Accordingly this was used for the evaluation set. Again, the table shows that the pitch feature does not perform on the same level as the MFCC features. When the distribution among errors for the individual classes was examined, it revealed that most classifiers were good at recognizing the neutral and positive class, but not the negative one. The reason for this is most probably its low frequency, which results in poor training statistics. The 3-gram classifier gives an improvement of the performance when combined with the acoustic features.

Table 6: *ISL Meeting Corpus Evaluation set. Accuracy, Average Recall and f1.*

Classifier	Acc.	A. Rec.	f1
Random with equal priors	*0.33*	*0.33*	*0.33*
MFCC	0.66	0.49	0.47
MFCC-low 25 ms	0.66	0.46	0.44
Pitch	0.41	0.38	0.37
3-gram	0.47	0.57	0.52
Acoustic combination	0.79	0.50	0.47
Acoustic + 3-gram combination	0.80	0.53	0.50
Random using priors	*0.67*	*0.33*	*0.33*
Acoustic comb. using priors	0.82	0.42	0.48
Acoustic+3-gram comb. using priors	0.85	0.48	0.52

7. Discussion

The diverse results achieved on our two corpora, VP and ISL, are not surprising considering the differences between them. The VP corpus is labeled by a single expert, and the ISL data is annotated by three naïve labelers. Having three labelers makes majority voting possible. For the ISL data, some speakers in the development set occur in the evaluation set (a complete disjoint set is impossible to achieve for this corpus), and although this may also be the case for the VP data, the effect of it is probably negligible. The two corpora have different acoustic quality and contain different languages. The utterances in the VP data are noisy and recorded from various telephones, while the ISL data are recorded by quality microphones in silent office environment. The VP utterances generally consist of one or a few words directed to a machine. The ISL meeting utterances are longer and directed to humans. The ISL emotions were mostly positive and were easier to detect, while the VP material only contained negative emotions, due to users frustration with the system.

8. Conclusion

Automatic detection of emotions has been evaluated using spectral and pitch features, all modeled by GMMs on the frame level. The method has been tested on two corpora; voice controlled telephone services and meetings. The corpora were in two different languages, Swedish and English. Results show that frame level GMMs are useful for emotion classification.

Combining the three main acoustic classifiers significantly improved performance. Including 3-grams for the ISL corpus gave a further improvement.

The two MFCC methods used show similar performance, and MFCC-low outperforms pitch features. A reason may be that MFCC-low gives a more stable measure of the pitch. Also, it may be due to its ability to capture voice source characteristics, see [17], where the level difference between the first and the second harmonic is used as a measure of different phonations, which in turn may vary across emotions. Our reason for introducing MFCC-low was to measure pitch. However, considering the potential to measure phonation from the first two harmonics, its upper frequency should be at least doubled in order to capture these effects for all female voices.

The lower overall performance for the VP data may be due to the telephone quality of the speech and possibly also due to a larger variation in the way negative emotions are expressed. The language differences may also influence the results.

A possible way to improve performance for the VP corpus would be to perform emotion detection on the dialogue, rather than the utterance level, and also take the lexical content into account. This would mimic the behavior of the human labeler.

Above we have indicated the difficulty to compare emotion recognition results. However, it seems that our results at least compare to those in [5] that also reports results on emotion recognition for a real-life interactive voice response system.

9. Acknowledgements

Computers in the Human Interaction Loop, is an Integrated Project under the European Commission's 6th Framework Program (IP-506909). We want to thank Voice Provider for providing us with their speech material and Inger Karlsson for performing the emotion labeling of their data.

10. References

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

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

[3] Dellaert F., Polzin T.S. and Waibel A. "Recognizing emotion in speech", ICSLP, Pittsburgh, 1996.

[4] Ang J., Dhillon R., Krupski A., et al, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog", ICSLP, Denver, 2002.

[5] Blouin, C., and Maffiolo, V., "A study on the automatic detection and characterization of emotion in a voice service context", Interspeech, Lisbon, 469-472, 2005.

[6] Oudeyer, P., "Novel Useful Features and Algorithms for the Recognition of Emotions in. Human Speech". Proceedings of the 1st Int. Conf. on Speech Prosody, 2002.

[7] Batliner, A., Buckow, J., Huber, R., Warnke, V., Noth, E., and Niemann, H., "Prosodic Feature Evaluation: Brute Force or Well Designed?" In Proc. 14th Int. Congress of Phonetic Sciences, 3:2315-2318, San Francisco, 1999.

[8] Chul M. L., Narayanan, S., "Toward Detecting Emotions in Spoken Dialogs". IEEE, Transactions on Speech and Audio Processing, vol 13, no 2, pp. 293-303, March, 2005.

[9] Slaney, M. and McRoberts, G..,"Baby Ears: A Recognition System for Affective Vocalizations", ICASSP '98, 1998.

[10] Laskowski, K. and Burger, S., "Annotation and Analysis of Emotionally Relevant Behavior in the ISL Meeting Corpus", LREC, Genoa, Italy, 2006.

[11] Hermansky, H. and Morgan, N., "RASTA processing of speech", IEEE Trans. Speech and Audio Processing, vol.2, no.4, pp.578-589, 1994.

[12] Ross, M., Shafer, H., Cohen, A. Freudberg, R., Manley, H., "Average magnitude difference function pitch extraction", IEEE Trans. ASSP-22, pp. 353-362, 1974.

[13] Langlais, P., "Traitement de la prosodie en reconnaissance automatique de la parole", PhD-thesis, University of Avignon, France, September, 1995.

[14] Gauvin, J-L and Lee, C.-H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech Audio Processing, vol. 2, pp. 291-298, Apr. 1994.

[15] Reynolds, D., Quatieri, T., and Dunn, R., "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol. 10, p. 19-41, January /April/July 2000.

[16] Stolcke, A., "SRILM - An Extensible Language Modeling Toolkit", in ICSLP, Denver, Colorado, USA, 2002.

[17] Syrdal, A. K., "Acoustic variability in spontaneous conversational speech of American English talkers", In ICSLP-1996, 438-441, 1996.