



Automated Vocal Emotion Recognition Using Phoneme Class Specific Features

Géza Kiss, Jan van Santen

Center for Spoken Language Understanding, Oregon Health & Science University

geza.kiss@cslu.ogi.edu, vansanten@cslu.ogi.edu

Abstract

Methods for automated vocal emotion recognition often use acoustic feature vectors that are computed for each frame in an utterance, and global statistics based on these acoustic feature vectors. However, at least two considerations argue for usage of phoneme class specific features for emotion recognition. First, there are well-known effects of phoneme class on some of these features. Second, it is plausible that emotion influences the speech signal in ways that differ between phoneme classes. A new method based on the concept of phoneme class specific features is proposed in which different features are selected for regions associated with different phoneme classes and then optimally combined, using machine learning algorithms. A small but significant improvement was found when this method was compared with an otherwise identical method in which features were used uniformly over different phoneme classes.

Index Terms: emotion recognition, phoneme class specific features, biomedical application

1. Introduction

Automated vocal emotion recognition (AVER) has numerous applications, including user-sensitive avatars and other advanced user interface technologies, as well as the study of vocal emotion in special populations. An area of specific interest in our Center is the study of vocal emotion in individuals with autism spectrum disorders. We need to address questions such as: Are the acoustic features used for vocal emotion production different from those in the general population? Is vocal emotion less sensitive to conversational context? Is the frequency distribution of certain discrete emotion classes different? For these biomedical applications, AVER serves two purposes. First, AVER systems can provide discrete emotion class labels that will help to answer these questions. Second, even if AVER classification performance is imperfect, the questions can also be addressed in terms of the acoustic features that are found to be key for AVER. The present study aims at accurate classification performance, but the latter, that is finding highly emotion-relevant acoustic features, is one of its important sub-goals.

Despite the wealth of potential applications, AVER is still a substantially unsolved problem, with even the best systems performing significantly worse than human performance which, in turn, is also far from perfect (see e.g., [1]). One potential reason for the imperfect human emotion recognition is that in the overwhelming majority of natural situations, vocal emotion information is processed in conjunction with other channels, such as the contents of the spoken communication, facial expression, and gesture. It is plausible that each channel has its own limitations in terms of which emotion classes it conveys well and which ones it does not, by analogy to certain speech sound distinctions being conveyed well via visible speech (e.g., bilabial versus alveolar) and

others via the acoustic signal (e.g., stops versus nasals). An example of this is the active vs. passive distinction, which may be conveyed well via the vocal channel (e.g., distinguishing happy and angry from sad and fearful), while the positive vs. negative distinction (happy versus angry) may be conveyed more accurately via the facial channel. Thus, there may be inherent limitations of the emotion information present in the acoustic signal, which may put an upper bound on the AVER performance.

Nevertheless, even with this upper bound in mind, there is significant room for AVER performance improvement. In this paper, we explore an AVER method based on the hypothesis that extracting the same types of features not from the entire utterance, but separately from segments of the waveform belonging to different phoneme classes would furnish us with additional features that improve the recognition performance. The intuition underlying this hypothesis is based on work by, among others, Scherer ([1], Table 6), showing the relevance for vocal emotion of a wide variety of features. These features include ones that are clearly only applicable to vowel and semivowel regions (e.g., pitch, formants), while others (e.g., intensity) are relevant for all phoneme classes. In addition, there are powerful effects of manner of production and voicing on some of these features (e.g., intensity, spectral balance). It would thus seem that representing the emotion-relevant acoustic information in an utterance handling all phoneme regions in the utterance uniformly is sub-optimal.

2. Method

In this section we describe the components and choices of the emotion recognition system that we used to evaluate our hypothesis: the emotional corpus, the features used, the feature selection and normalization method, and the classifier model.

2.1. Emotional corpus

We used an American English acted emotional corpus [2] for this work, henceforth referred to as the Actor corpus. This corpus contains acted realizations of four emotions from adult, teenager and child speakers of American English, from both genders. The corpus was generated as follows. For each of 24 emotionally neutral sentences (e.g., "it is round") a brief vignette was written that ended on the neutral sentence and that was targeted to evoke a particular emotion class (anger, happiness, sadness, fear). The sentences were semantically unbiased in their affective content, that is, it was impossible to predict which affect was intended from the text alone. The sentences consist of a single phrase that is 2-5 words in length. The simulated vocal expressions obtained in this manner will yield more intense, prototypical expressions of affect [1]. A perceptual experiment showed that listeners were generally able to reliably recognize the intended affects reflecting the fact that these recordings represent normal expression patterns. Untrained actors read all 96 vignettes, blocked by emotion, enabling them to "get into" a given affect. The actors were

recruited with maximal diversity with respect to age, intrinsic pitch, voice quality, and habitual speaking rate and loudness. The actors did not produce neutral renditions of the sentences. For further details, see Table 1.

Table 1. *Some characteristics of the Actor emotional corpus.*

characteristics	Values
number of speakers	14
sentences per speaker	21-24 per emotion
total length	40 minutes
age of speakers	9-45; mostly 25-35
emotion type	acted
emotion label source	intended emotion
emotion classes	Anger, Fear, Happiness, Sadness

A prerequisite for phoneme-class specific features is that we need to know the boundaries of phonemes in the speech, or at least those of the different phoneme types. Advancements in ASR systems have made automatic segmentation possible with a relatively high performance (see for example [3][16]), so the use of such features may even be feasible in real-time emotion recognition. For the purposes of this work, we used manually checked phonemic transcriptions and aligned them with the speech using the CSLU Toolkit forced alignment system [5]. The phoneme boundaries were not hand-corrected.

2.2. Features

We used the baseline feature types used in the Interspeech 2009 Emotion Challenge (IS09 EC) [4]. These proved to perform quite well in the challenge, as no one submitted better features in the Feature sub-challenge [6]. We expect that by later adding features that work particularly well for phenomena observable only on certain phoneme types, we can further enhance the utility of the phoneme specific features.

The IS09 EC features are calculated as statistics from low-level descriptors (LLD) extracted from frames of the waveform, which are generated with a 25-ms Hamming window and 10-ms shift. The LLDs are zero-crossing rate, RMS frame energy, fundamental frequency normalized to 500 Hz, harmonics-to-noise ratio, and 12 mel-frequency cepstral coefficients. The statistics calculated from these are the mean, standard deviation, kurtosis, skewness, minimum, maximum, the position where minimum and maximum occurred, range, two linear regression coefficients and their mean square error. We used the openSMILE open-source system [7] to extract these from different portions of the utterances, after necessary modifications of the code.

We extracted the features from the whole length of the utterances (baseline features), and separately for the segments belonging to different phoneme classes (proposed features). We classified each frame to the phoneme whose center it was closest to. The phoneme classes we used were: vowels, glides, nasals, fricatives, plosive closures, plosive bursts, and non-speech. Although we extracted phoneme-class specific features from all parts of the waveform, we also considered using the baseline features as possible contributors to an overall improvement.

2.3. Feature selection

We used the mRMR (minimum Redundancy Maximum Relevance) feature selection method [8], which works as a filter method for feature selection. It creates a ranking of the

features, where the first ones in the list have the maximum relevance to the class, while having minimum dependency among each other. One can either choose a certain number of features from the beginning of the list, or use it in a wrapper framework in a forward or backward selection scheme.

We trained and evaluated the classifiers (see below) on the first 10, 20, and so on features taken from the start of the ranking. We divided the training data into design sets and corresponding development sets in a leave-one-speaker-out cross-validation framework. For each classifier, we chose the feature set size whose average performance on the development sets was the best. If two feature sets performed similarly, we chose the smaller set; we found this beneficial in an exploratory evaluation.

2.4. Classifiers

We used the static classifier scheme described in the IS09 EC paper [4], except that we trained SVMs with a polynomial kernel instead of using SMO; we used LIBSVM [9] as incorporated into the openEAR [7] framework (version 0.1.0) for this purpose.

We trained a classifier for each phoneme-class specific feature set extracted from the different phoneme classes, and fused their outputs for decision fusion. For the utterances where a phoneme class was not represented, the corresponding classifier output equal probability scores for the emotion classes. We chose this approach instead of merging all features into one classifier so as to avoid having too many highly correlated features, which may cause instability in the SVM training and other artifacts.

We normalized the features using the LIBSVM toolkit [9] by determining an offset and a scaling factor for each feature such that they fall into the $[-1,+1]$ range over the whole training set. The same scaling factors and offsets were used to normalize the test samples. We expected that performing the normalization independently for phoneme types contributes to the improvement over the baseline, since several of the LLDs have ranges that are peculiar to each phoneme; for example their energy levels. The statistics for these LLDs may have a quasi-random variance in the baseline feature set, depending on the ratios of the phoneme classes in each sentence. We may be able to achieve some improvement by a further phoneme-specific normalization within the broad phoneme class, but we assume that the within-class differences are much smaller than the inter-class ones.

2.5. Decision fusion

We used the FoCal toolkit [10] for fusing the outputs of the classifiers. FoCal uses Cllr [17], an information theoretic metric, for finding the optimal fusion parameters for combining classifier outputs, and contains possibilities for discriminative as well as generative training; furthermore, it can transform output scores of any kind into probabilities using a Gaussian model.

In our search for the optimal fusion type, we trained the parameters of different types of fusion methods on a design set, and evaluated these on separate development set. For the quadratic fusion models, we got better performance on the design set and worse performance on the development set than for the logistic linear regression fusion (LLR); in other words we had a case of overfitting. The quadratic models have many more parameters than LLR, and this result indicates that the Actor corpus is too small for training these parameters reliably. For this reason, we decided to use LLR, which

generates a scaling factor and an offset for the probability scores of each classifier.

3. Results

3.1. Evaluation method

We applied our classification scheme to a multi-class classification task for four emotions (see Table 1).

In order to achieve reliable performance estimation, we used a leave-one-speaker-out cross-validation method. For every speaker, we trained the classifiers on the data of all the other speakers, evaluated it on the remaining speaker, and finally combined the performance on these cross-validation folds. This way we tested the classifiers on speakers that were not represented in the training data, while we could still use most of the data for training, which was important for this corpus of relatively small size.

We used unweighted average recall (UAR) as our main metric for comparing results: the recalls on the different emotions are averaged, not taking into account the percentage with which the emotion is represented in the corpus. The use of this metric is in accordance with the IS09 EC guidelines, and is often used to reward a system with good performance on minority classes. Consequently the weighted—unweighted distinction is relevant especially in the context of a spontaneous emotional corpus, where there is generally a great imbalance between the emotional categories, and neutral utterances tend to be the majority, whereas an acted emotional corpus is balanced by design. For the balanced Actor emotional corpus, the unweighted and weighted performances were about the same.

3.2. Optimal classifier fusion

After experimenting with all possible combinations of the classifiers, including the one trained on the IS09 EC baseline features, we found that the following combination gave the overall best performance: the fusion of four classifiers trained on the features extracted from vowels, glides, nasals, and non-speech parts. These were not the ones with the best individual performances, as you can see in Figure 1; the amount of complementary information must be a more important factor.

We were surprised to find that the classifier trained on the non-speech part contributed to the overall best result. This may be explained by two factors. One is the presence of non-speech vocalizations, for example the sound of breathing, which can be different for these emotions. The other is that the non-speech frames contain some speech: partly because the sliding window sometimes overlaps two phonemes at the boundary; and partly the phoneme boundaries were not hand-corrected, therefore systematic errors of the forced alignment system have placed some phrase beginning and ending segments into the non-speech part. An abrupt start, or a low-intensity utterance ending speech segment can indeed be characteristic of anger or sadness, respectively.

3.3. Performance

The fused classifier gave an improvement on most, but not all speakers (i.e. cross-validation folds), as shown in Figure 2; but it resulted in a significant improvement on the average. The performance of the baseline and the fused classifier were 48.7% and 52.1% UAR respectively, which is 3.4% absolute and 7% relative improvement. We evaluated the significance of the improvement with a one-tailed one-sample t-test on the

14 per-speaker performance changes. This is justified by the facts that there was no chance of getting a lower performance, since we evaluated all possible combinations of one or more classifiers including the baseline classifier; and that the speech samples are not independent from each other, but the speakers are. We found that the improvement was significant, with $t(13)=1.803$, $p=0.047$.

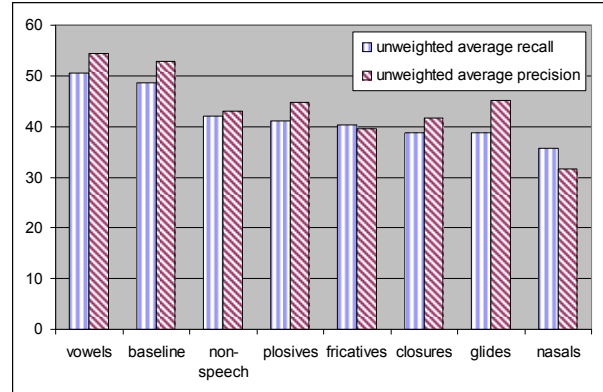


Figure 1: Comparison of the overall performances of classifiers trained on phoneme-class specific features, in decreasing order of unweighted average recall.

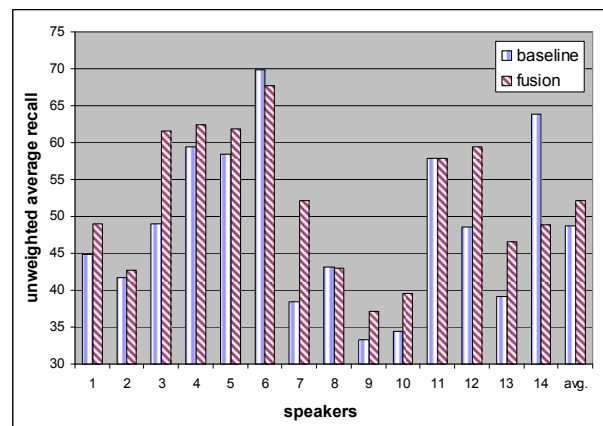


Figure 2: Comparison of the performances of the baseline and the fused classifier on the leave-one-speaker-out cross-validation folds.

3.4. Comparison to previous work

Two papers submitted to IS09 EC used features that were extracted from only certain parts of the waveform: Luengo et al. [11] identified the vowel and consonant parts of the speech using an HMM model with 80 percent accuracy, and extracted prosodic features for the vowels. Kockmann et al. [12] used a Hungarian speech recognizer for voice activity detection, and discarded the non-speech part from the feature extraction. Although these studies segmented the speech with regard to the type of speech content, our approach is different in the sense that we extracted features from all available phoneme types, and evaluated their individual contribution to the overall performance. Lee et al. [15] also examined the effect of phoneme-class dependent emotion recognition on four emotions. They extracted 13 MFCC coefficients, together with delta and acceleration, and used GMM modeling in five phoneme-specific HMM models for each emotion. They

reached a significant improvement compared to a general HMM model trained on the whole of the speech. Both their study and ours examined the same problem, but with a different approach: while their work examined the use of a dynamic framework, our study concentrated on the use of static features for the task.

3.5. Discussion

We have shown that extracting static LLDs separately from different phoneme classes does significantly improve emotion recognition performance. As can be seen in Figure 1, the vowel classifier by itself performs better than the baseline classifier, although the difference is not significant in this case.

The performance we reached on this data set is much higher than that the performance on spontaneous corpora (e.g., [4]), but lower than some similar results for four emotions on acted corpora (e.g., [15], [18]). The difference can be explained by the fact that the database has a relatively small size, and at the same time great diversity in age, and furthermore we used all the utterances of the corpus, it was not filtered to retain utterances with a high labeler agreement.

Evaluating the confusion matrix of this four emotion classification task, shown in Table 2, we found in accordance with other results that anger and happiness, as well as fear and sadness, are more often confused with each other than with other emotions. These are not differentiated by their activation (or arousal) level, and several authors (see e.g., [14]) have shown that even humans tend to mistake them for each other.

Table 2. Confusion matrix of the fused classifier: true vs. predicted classes (true class on left).

	Anger	Fear	Happiness	Sadness
Anger	191	51	67	23
Fear	33	137	59	100
Happiness	91	54	136	22
Sadness	6	82	16	222

4. Conclusions

This work examined the usefulness of using phoneme-class specific features for emotion recognition. We used the Interspeech 2009 Emotion Challenge baseline feature types, extracted them from speech segments for six phoneme types and non-speech segments for an acted American English speech database. We found that the fusion of four of the classifiers trained on these resulted in a moderate but significant improvement in the classification performance, which justifies the viability of this approach. Possible directions for further research include finding features that are especially useful for certain phoneme types, including spectral balance and duration features, which are phoneme-specific by nature. We also plan to examine other classifier schemes, including dynamic frameworks. Investigating the conditions for using such features in a real-time emotion recognition task is also an area of interest.

5. Acknowledgements

We thank Esther Klabbers for her help in working with the Actor corpus and for her remarks on an earlier version of this paper; and Paul Hosom for his advice in the use of the CSLU forced alignment system. Géza Kiss would like to thank the Fulbright Fellowship for providing a scholarship enabling him to partake in this research. This research was also supported by grants from the National Institute on Deafness and Other

Communication Disorders, 1R21DC010239 (Lois Black, PI) and the National Science Foundation, 0905095 (Jan van Santen, PI). The views herein are those of the authors and do not necessarily reflect the views of either the above individuals or the funding agencies.

6. References

- [1] Scherer, K. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40, 227–256. 2003.
- [2] Klabbers, E., Mishra, T., van Santen, J., “Analysis of affective speech recordings using the superpositional intonation model”, *Proc. 6th ISCA Workshop on Speech Synthesis (SSW6)*, Bonn, Germany, 339-344, 2007
- [3] Ali, A.M.A., Van der Spiegel, J., Mueller, P. and Haentjens, G. and Berman, J., “An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech”, *ICASSP 1999 Proc.*, Vol. 3, 118-121, 1999.
- [4] Schuller, B., Steidl, S., Batliner, A., “The Interspeech 2009 Emotion Challenge”, *Interspeech 2009 Proc.*, Brighton, UK, 312-315, ISCA, 2009.
- [5] Hosom, J.P., “Speaker-independent phoneme alignment using transition-dependent states”, *Speech Communication*, 51(4), 352-368, Elsevier, 2009.
- [6] Schuller, B., Steidl, S., Batliner, A. and Jurcicek F., “The INTERSPEECH 2009 Emotion Challenge: Results and Lessons Learnt”, *SLTC Newsletter*, October 2009. Online: <http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2009-10/interspeech-emotion-challenge/>, accessed on 29 April 2010.
- [7] Eyben, F., Wöllmer, M., Schuller, B.: “openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit”, *ACII 2009 Proc.*, IEEE, Amsterdam, The Netherlands, 2009.
- [8] Peng, H., Long, G., Ding C., “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (8), 1226-1238, 2005.
- [9] Chang, C.C., Lin, C.J., “LIBSVM: a library for support vector machines”, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [10] Brümmer, N., “FoCal Multi-class”, Software available at <http://niko.brummer.googlepages.com/focalmulticlass>.
- [11] Luengo, I., Navas, E., Hernaez, I., “Combining spectral and prosodic information for emotion recognition in the Interspeech 2009 Emotion Challenge”, *Interspeech 2009 Proc.*, Brighton, 332-335, 2009.
- [12] Kockmann, M., Burget, L., Černocký, J., “Brno University of Technology System for Interspeech 2009 Emotion Challenge”, *Interspeech 2009 Proc.*, Brighton, 348-351, 2009.
- [13] Yacoub, S., Simske, S., Lin, X. and Burns, J., “Recognition of emotions in interactive voice response systems”, *Eurospeech 2003 Proc.*, 729-732, 2003.
- [14] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, “Analysis of emotion recognition using facial expressions, speech and multimodal information”, *Sixth International Conference on Multimodal Interfaces, ICMI 2004*. State College, PA: ACM Press, 205-211, 2004
- [15] Lee, C., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S., “Emotion recognition based on phoneme classes”, *Interspeech 2004 Proc.*, 205–211, 2004
- [16] Hosom, J.P., “Automatic Time Alignment of Phonemes using Acoustic-Phonetic Information”, *PhD Thesis*, Oregon Graduate Institute, Beaverton, OR, 2000.
- [17] Brümmer, N., du Preez, J., “Application-Independent Evaluation of Speaker Detection”, *Computer Speech and Language*, 20 (2-3), 230-275, 2006.
- [18] Pao, T.L., Chen, Y.T., Yeh, J.H., Li, P.J., “Mandarin Emotional Speech Recognition Based on SVM and NN,” *ICPR 2006*, Vol. 1, 1096-1099, 2006.