A Hierarchical Approach with Feature Selection for Emotion Recognition from Speech

Panagiotis Giannoulis [1] , **Gerasimos Potamianos** [2,3]

[1] Department of Electrical and Computer Engineering, National Technical University of Athens, Greece
[2] Department of Computer and Communication Engineering, University of Thessaly, Volos, Greece
[3] Institute of Informatics and Telecommunications, NCSR "Demokritos", Athens, Greece

paniotiso@gmail.com, gpotam@ieee.org

Abstract

We examine speaker independent emotion classification from speech, reporting experiments on the Berlin database across six basic emotions. Our approach is novel in a number of ways: First, it is hierarchical, motivated by our belief that the most suitable feature set for classification is different for each pair of emotions. Further, it uses a large number of feature sets of different types, such as prosodic, spectral, glottal flow based, and AM-FM ones. Finally, it employs a two-stage feature selection strategy to achieve discriminative dimensionality reduction. The approach results to a classification rate of 85%, comparable to the state-of-the-art on this dataset.

Keywords: Emotion recognition, glottal flow, AM-FM, feature selection

1. Introduction

Automatic emotion recognition from speech has attracted significant interest in recent years, aiming at improved human-computer interaction. Several approaches have been proposed in the literature, varying in the types and number of features employed, the classifier used, and the system developed. For example, there have been works using traditional features from the speech recognition literature or higher level voice quality ones (He et al., 2010), some adopting flat and simple systems, and others that employ complex classifiers and systems of a hierarchical form (Mao and Zhan, 2010; Shaukat and Chen, 2008). Up to now, most of the highest performing speaker independent emotion recognition systems use large feature sets and rather complex classifiers (Schuller et al., 2006; Lee and Narayanan, 2005). For example, Schuller et al. (2006) extract a large set of 4k features, achieving an average recognition rate of 87% on the Berlin database of emotional speech (Burkhardt et al., 2005).

The proposed approach in this paper achieves comparable performance with the best research efforts by using only 112 features in total. We accomplish this by designing a system that is based on smaller and more specially trained sub-systems that focus on pairs of emotions. We also use a two-stage feature selection scheme, in contrast to the state-of-the-art simple sequential selection. We finally involve feature sets that are not typically employed in the emotion recognition literature, such as glottal flow features and AM-FM ones. Thus, although the total number of features remains small, there is much variety in their types.

The rest of the paper is organized as follows: First, in Section 2, we describe the different feature sets considered, together with their basic theoretical background. In Section 3, we present our two-stage feature selection approach, and, in Section 4, we explain how the final system works. Following these, in Section 5, we present our experiments and associated results. Finally, we conclude the paper with a short summary in Section 6.

2. Feature Extraction

A number of feature sets are considered in our approach. In more detail they fall within the following four categories:

2.1. Prosodic Features

Such features are strongly related to the emotional state of the speaker and are extensively used in the literature. In this paper, we extract two features relevant to prosody intonation and intensity. First, we calculate the pitch contour of the utterance using the RAPT algorithm (Talkin, 1995), and, second, we compute signal energy to obtain information about speech intensity.

2.2. Spectral Features

As in the state-of-the-art in speech recognition, we employ the Mel frequency cepstral coefficients (MFCCs), together with their first-order derivatives (Young et al., 2002). In addition, we compute the zero crossing rate (ZCR) of each frame (Young et al., 2002).

2.3. Glottal Flow Features

The volume velocity of air-flow through the glottis is the excitation source for voiced speech. The glottal flow is related with several voice quality features, such as breathiness, harshness, and creakiness, and therefore it provides useful information about the emotional state of the speaker. The estimation of the glottal flow is based on Fant's source-filter theory, according to which the voice excitation and the vocal tract are linearly separable. In this manner, speech production can be modeled by a cascade of linearly separable filters. In order to obtain the glottal flow, we perform inverse filtering by using the IAIF algorithm, which employs the discrete all pole (DAP) method to model the vocal tract and then cancels it iteratively to obtain an estimate of the glottal flow waveform (Airas et al., 2005).

Once the glottal flow is obtained, we extract time and frequency based features from its waveform, such as the:

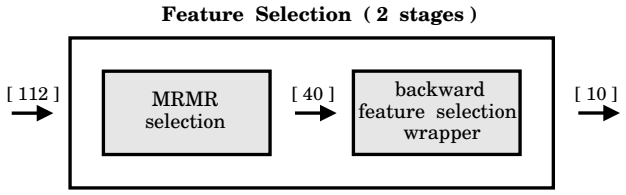

Figure 1: The two-stage feature selection scheme employed here. Numbers in brackets depict feature dimensionality.

- *Open quotient*, which is the ratio of the time in which the vocal folds are open and the whole pitch period duration.
- *Speed quotient*, defined as the ratio of rise and fall time of the glottal flow.
- *Normalized amplitude quotient* that is used to parametrize the glottal closing phase.
- *Harmonic richness factor*, which is the ratio of higher harmonics to the first harmonic.

In total, we extract 18 time based and 4 frequency based features.

2.4. AM-FM Features

The AM-FM model considers decomposing the speech signal into a series of a few instantaneous frequency and amplitude signals. These signals can be considered as time-frequency distributions, containing acoustic information that is not captured by the linear speech model (Potamianos and Maragos, 1999).

Following the approach in (Dimitriadis and Maragos, 2005), we model each speech sound by six AM-FM signals, estimating their parameters by the energy separation algorithm (ESA). At the end, twelve (2×6) parameters are obtained for each utterance frame.

We should note that, for all aforementioned feature sets, we only utilize their first and second order statistics (mean and variance).

3. Feature Selection Strategy

Feature selection is a very crucial step in pattern recognition problems to counter the curse of dimensionality. It does so, avoiding feature transformations such as PCA or LDA, obtaining instead a subset from the initial set of features that is most relevant to the classification problem at hand.

In general, this requires a search strategy to select candidate subsets and an objective function to evaluate such candidates. Depending on their objective function, feature selection algorithms can be divided into filters and wrappers. Filters evaluate feature subsets by their information content, typically statistical dependence or information-theoretic measures. Wrappers, in contrast, evaluate the subsets by their classification rate on test data. One can claim that filters have better generalization properties, as they are not related to any classifier. On the other hand, wrappers can interact with a specific classifier and find a subset that is also appropriate for the problem at hand.

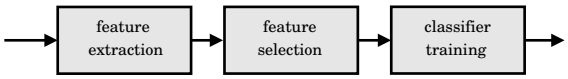

Figure 2: Training of each emotion recognition sub-system.

Two of the simplest and most used wrappers belong to sequential feature selection algorithms, and are *forward feature selection* (FFS) and *backward feature selection* (BFS). FFS starts from the empty set and sequentially adds the feature that results in the highest recognition rate, when combined with the features that have already been selected. In contrast, BFS starts from the full feature set and sequentially removes the feature that leaves a subset with the highest recognition rate. Both constitute greedy approaches and may be trapped in local minima. In our experiments, we compare these two wrappers and show that BFS performs better, especially when the initial set of features is large.

In our system we use a two-step feature selection scheme that takes advantage of both generalization properties of filter algorithms and classifier-adaptive properties of wrappers. As a filter we employ the *maximum relevance-minimum redundancy* (MRMR) algorithm (Peng et al., 2005). This method tries to select a feature set that has maximum relevance with the two emotions involved in each two-class emotion classification sub-problem (see Section 4), as well as minimum intra-redundancy in terms of mutual information between its features.

In more detail, if $\mathcal{S}$ denotes a set of selected features within the set of all possible features Ω, then a measure of its *redundancy* is given by

$$W_{\mathcal{S}} = \frac{1}{|\mathcal{S}|^2} \sum_{i,j \in \mathcal{S}} I(i,j) \ ,$$

where $I(i,j)$ represents the mutual information between features i and j, and $|\mathcal{S}|$ denotes the number of features in set $\mathcal{S}$. The *minimum redundancy criterion* seeks the set of features $\mathcal{S}$ that minimizes $W_{\mathcal{S}}$. Next, if $c \in \mathcal{C}$ denotes the class of interest (for the two-class emotion classification problem we will have $\mathcal{C} = \{c_1, c_2\}$), we can calculate the *relevance* of feature set $\mathcal{S}$ as

$$V_{\mathcal{C},\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} I(c,i) \ .$$

The *maximum relevance criterion* seeks feature set $\mathcal{S}$ that maximizes $V_{\mathcal{C},\mathcal{S}}$ for the specific classification problem $\mathcal{C}$.

The MRMR algorithm then tries to achieve both of the two previous criteria by maximizing the following quotient:

$$\max_{\mathcal{S} \subset \Omega} \frac{\sum_{i \in \mathcal{S}} I(c,i)}{\frac{1}{|\mathcal{S}|} \sum_{i,j \in \mathcal{S}} I(i,j)} \ .$$

Obtaining the optimal solution to the above through exhaustive search is clearly intractable. In practice, one proceeds with a sequential, incremental, non-optimal approach, by first selecting as the first feature the one that maximizes the relevance criterion, and subsequently adding

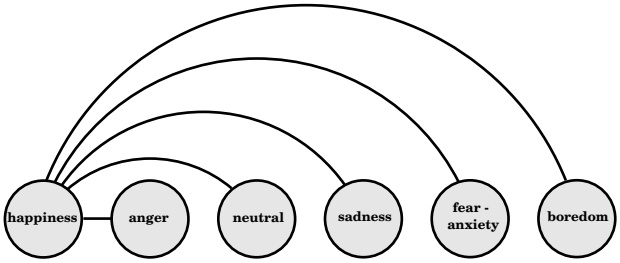

Figure 3: The five sub-systems for recognizing "happiness", denoted by the five lines connecting the circles.

one feature at a time, similarly to the FSS approach mentioned earlier. Assuming that the selected set at the current iteration is $\mathcal{S}$, then feature $i \in \Omega - \mathcal{S}$ will be selected as

$$i = \arg\max_{i \in \Omega - \mathcal{S}} \frac{I(c,i)}{\frac{1}{|\mathcal{S}|}\sum_{j \in \mathcal{S}} I(i,j)} .$$

The process terminates, once we arrive at the desired feature set size. We should also note that, in the above, all mutual information quantities are computed after appropriate discretization of all continued-valued features.
The two-stage feature selection scheme employed in our approach is depicted in Figure 1.

4. Final System

Our emotion classification system has a *hierarchical* structure, as it is composed from multiple specialized systems. Each one of the smaller systems is trained to distinguish between only one pair of emotions (e.g., fear vs. happiness, anger vs. happiness, etc.), and it is trained separately, as depicted in Figures 2 and 3. Since, in this paper, we concentrate on six basic emotions (see Section 5), the proposed approach gives rise to 15 sub-systems. Each such sub-system is implemented based on a simple linear support vector machine (SVM) classifier. Majority voting over the sub-system classification results is then used to provide the final classification. In addition, we use the a-priori knowledge of gender information, as such has been shown to play an important role in emotion recognition in the literature.

5. Experiments and Results

In our experiments we used the well-known Berlin database of emotional speech. The corpus has been recorded at the Technical University of Berlin, and it consists of 493 utterances by 10 professional actors (5 male and 5 female). It contains seven emotions (acted), namely anger, happiness, sadness, fear, boredom, neutral, and disgust (Burkhardt et al., 2005), of which we concentrate on the first six in accordance with similar work in the literature. We follow a leave-one-speaker-out experimental paradigm to provide speaker independence.
In the following tables we present results in the form of confusion matrices for several different experiments. From these, we can immediately observe the superiority of the gender dependent vs. the gender independent approach. Clearly, a-priori gender knowledge significantly improves

	Anger	Fear	Sadness	Boredom	Neutral	Happiness
Anger	**87**%	2%	0%	1%	0%	10%
Fear	7%	**78**%	4%	3%	1%	7%
Sadness	1%	7%	**84**%	6%	2%	0%
Boredom	0%	6%	7%	**61**%	23%	3%
Neutral	3%	3%	0%	12%	**82**%	0%
Happiness	14%	10%	0%	1%	1%	**74**%

Table 1: Confusion matrix for the *gender dependent* experiment with *forward selection* algorithm used for feature selection from the entire set of features. The overall accuracy is **77.08**%.

	Anger	Fear	Sadness	Boredom	Neutral	Happiness
Anger	**86**%	1%	0%	1%	0%	12%
Fear	6%	**82**%	3%	3%	1%	5%
Sadness	0%	5%	**88**%	5%	2%	0%
Boredom	0%	4%	6%	**64**%	22%	4%
Neutral	2%	3%	0%	10%	**84**%	1%
Happiness	15%	9%	0%	1%	2%	**73**%

Table 2: Confusion matrix for the *gender dependent* experiment with *backward selection* algorithm used for feature selection from the entire set of features. The overall accuracy is **79.71**%.

algorithm performance. Furthermore, we observe the gradual improvement of recognition accuracy of almost all emotion classes, first when applying the BFS scheme in place of FSS, and subsequently when employing the proposed two-stage approach that includes the MRMR algorithm. One should note that these results are in par with human emotion perception experiments reported at 84.3% by Schuller et al. (2007).
Next, in Figure 4, we show some results on the several subsystems performance with different initial features sets. As we can observe, the addition of glotal and AM-FM features lead to better recognition results. In a few sub-systems performance is somewhat better when using only MFCC and prosodic features. This demonstrates that such features are sufficient for the specific sub-systems.
Finally, concerning the types of features selected by the proposed algorithm, it should be noted that different feature combinations are selected for each subsystem, a fact that backs our hierarchical approach. For example, when one of the two classes of interest is fear, more glottal features are selected.

6. Conclusions

In this paper, we proposed a hierarchical classification system that is based on the discriminative power of its appropriately trained sub-systems. These sub-systems employ feature sets of various types appropriately selected through a two-stage feature selection algorithm. Finally we took advantage of the well-established method of gender dependent systems to achieve better results.
Our final system achieved an overall classification accuracy of 85.18% that is comparable to the state-of-the-art in the field. Both glottal flow and AM-FM features were selected by the feature selection scheme employed and indeed improved recognition results in comparison to the state-of-

	Anger	Fear	Sadness	Boredom	Neutral	Happiness
Anger	**90**%	2%	0%	1%	0%	7%
Fear	5%	**81**%	3%	6%	0%	5%
Sadness	1%	3%	**96**%	0%	0%	0%
Boredom	0%	3%	7%	**73**%	15%	2%
Neutral	0%	2%	0%	8%	**90**%	0%
Happiness	16%	2%	0%	1%	0%	**81**%

Table 3: Confusion matrix for the *gender dependent* experiment with the final proposed feature selection scheme (*backward selection* followed by *MRMR*). The overall accuracy is **85.18**%.

	Anger	Fear	Sadness	Boredom	Neutral	Happiness
Anger	**85**%	4%	0%	0%	0%	11%
Fear	5%	**74**%	5%	7%	2%	7%
Sadness	0%	3%	**86**%	6%	5%	0%
Boredom	0%	10%	6%	**71**%	11%	2%
Neutral	1%	3%	1%	9%	**84**%	2%
Happiness	17%	3%	0%	3%	0%	**77**%

Table 4: Confusion matrix for the *gender independent* experiment with the final proposed feature selection scheme (*backward selection* followed by *MRMR*). The overall accuracy is **80.09**%.

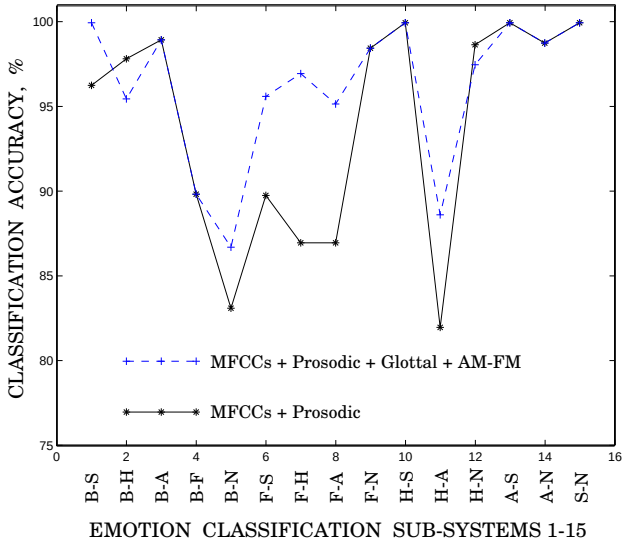

Figure 4: The recognition rates of the 15 different subsystems on female utterances with two different feature sets: One with only MFCCs and prosodic features, and one with all feature set types. Sub-systems are identified by the initials of their corresponding emotion pairs.

the-art systems that employ only MFCCs and prosodic features. Finally, the combination of the MRMR filter based algorithm with the BFS wrapper based one outperformed simple sequential feature selection.

In future work, we will investigate the use of temporal feature information (feature contours), as opposed to just their means and variances. Furthermore, we will consider larger datasets and spontaneous speech data.

7. Acknowledgments

The authors would like to thank Prof. Petros Maragos with the Electrical and Computer Engineering Department of the National Technical University of Athens, Greece, and Dr. Dimitrios Dimitriadis with the Networking and Service Laboratory of AT&T Labs Research, Florham Park, NJ, USA, for sharing the implementation of AM-FM feature extraction from speech.

G. Potamianos would also like to acknowledge partial support from the European Commission through FP7-PEOPLE-2009-RG-247948 grant AVISPIRE.

8. References

M. Airas, H. Pulakka, T. Bäckström, and P. Alku. 2005. A toolkit for voice inverse filtering and parametrisation. In *Proc. Interspeech*, pages 2145–2148, Lisbon, Portugal.

F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. 2005. A database of German emotional speech. In *Proc. Interspeech*, pages 1517–1520, Lisbon, Portugal.

D. Dimitriadis and P. Maragos. 2005. Robust AM-FM features for speech recognition. *IEEE Signal Process. Lett.*, 12(9):621–624.

L. He, M. Lech, and N. Allen. 2010. On the importance of glottal flow spectral energy for the recognition of emotions in speech. In *Proc. Interspeech*, pages 2346–2349, Makuhari, Japan.

C. M. Lee and S. S. Narayanan. 2005. Toward detecting emotions in spoken dialogues. *IEEE Trans. Speech Audio Process.*, 13(2):293–303.

Q.-R. Mao and Y.-Z. Zhan. 2010. A novel hierarchical speech emotion recognition method based on improved DDAGSVM. *Comp. Sc. Information Sys.*, 7(1):211–221.

H. Peng, F. Long, and C. Ding. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Machine Intell.*, 27(8):1226–1238.

A. Potamianos and P. Maragos. 1999. Speech analysis and synthesis using an AM-FM modulation model. *Speech Comm.*, 28:195–209.

B. Schuller, D.Arsic, F. Wallhoff, and G. Rigoll. 2006. Emotion recognition in the noise applying large acoustic feature sets. In *Proc. Int. Conf. Speech Prosody*, Dresden, Germany.

B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl. 2007. Towards more reality in the recognition of emotional speech. In *Proc. Int. Conf. Acoustics Speech Signal Process.*, volume 4, pages 941–944, Honolulu, HI, USA.

A. Shaukat and K. Chen. 2008. Towards automatic emotional state categorization from speech signals. In *Proc. Interspeech*, pages 2771–2774, Brisbane, Australia.

D. Talkin. 1995. A robust algorithm for pitch tracking (RAPT). In W. B. Klein and K. K. Paliwal, editors, *Speech Coding and Synthesis*, chapter 14, pages 495–518. Elsevier.

S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. 2002. *The HTK book*. Cambridge University, United Kingdom.