AE 76

EMOTIONAL SPACE IMPROVES EMOTION RECOGNITION

Raquel Tato[1], Rocío Santos[1], Ralf Kompe[1], J.M. Pardo[2]

[1]Sony International (Europe) GmbH
Advanced Technology Center Stuttgart/ MMI Lab
{tato,santos,kompe}@sony.de

[2]Universidad Politécnica de Madrid
E.T.S.I.T. / Grupo de Tecnología del Habla
{rsantos,pardo}@die.upm.es

ABSTRACT

A number of recent studies have focused on the conceptualized expression of emotions as a three-dimensional space. This paper proposes a new approach to emotion recognition, making use of two of the emotional dimensions and their relationship with different kinds of features. The main idea consists in associating prosodic features, derived from pitch, loudness, and duration, with the activation or arousal dimension, and quality features, i.e. phonation type, articulation manner, voice timbre, with the evaluation and pleasure dimension, in a way that different classification methods can be applied for each specific case. Most important results are achieved for the speaker-independent case and three classes, with a recognition rate close to 80%.

1. INTRODUCTION

The variety of human emotional states and affects can be represented in a three-dimensional space, i.e. arousal, pleasure and power, called emotional space. The arousal dimension, also named activation, refers to the degree of intensity, the pleasure or evaluation dimension refers to how positive or negative, i.e. pleasant, the emotion is perceived, and the power dimension relates to the sense of control over the emotional state. Emotions, which are placed close to each other in this space, have similar features in terms of acoustics and they are more difficult to distinguish in classification [1] [2].

Most of the prior art work [5] [6] makes only use of prosodic features, which are easier to handle but give mainly information concerning the arousal dimension of emotions. This study aims to show the need to take into account at least a second dimension of the emotional space, i.e. pleasure, and how much it is expressed by quality features, i.e. auditory features that arise from variation in the source signal and vocal tract properties [3]. For instance, happy and angry are emotions expressed with high intensity, i.e. with a high level in the arousal dimension, which makes them very difficult to classify based on prosody features. However, they are situated opposite to each other in the pleasure axis, and therefore, quality features will contribute effectively to enhance this classification.

Assuming that both dimensions of the emotional space relate to different features of the speech, with different degrees of complexity, it makes sense to divide the problem and apply hierarchical classification. The classification will be accomplished in terms of levels, from high to low, in two of the emotional dimensions, looking at a subspace of the features (prosody or quality features).

The target scenario is the Sony entertainment robot AIBO. To that end, an emotional database has been recorded, simulating different possible situations, which comprises all the desired emotions. From the application point of view it is interesting to have five different emotional states, angry, happy, sad, bored and neutral. According to the distribution of these emotions in the emotional space [4], it was decided to have a first classifier looking at prosodic features and giving as output three levels in the activation dimension (high=angry-happy, medium=neutral, low=sad-bored), and a second classifier, looking at quality features, and making the final decision, concerning an emotional state.

2. DATABASE

A clear difference in the performance of an emotion recognizer, can be achieved, depending on which kind of speech is used, i.e. actors, read speech, WOZ [5].

In order to obtain relevant results, it is desired to have a speech database, as close as possible to spontaneous emotional speech in the target scenario. With that purpose in mind, people were put in an emotional state by some context action and then asked to read the commands. About 40 commands in five emotions (angry, happy, sad, bored and neutral) were recorded for 14 German non-actor speakers, 7 male and 7 female, overall around 2800 utterances. For training and evaluation purpose, a "leaving-one-speaker-out" cross validation algorithm was implemented.

Following the same strategy, one of the speakers was recorded twice, approximately 400 utterances, to provide enough data for speaker-dependent experiments. The speaker-dependent database was distributed into 80% for training and 20% for testing.

3. FEATURE SUBSPACES

Previous research on feature extraction for emotion recognition has focused on prosodic features, based on different linguistic units as utterance vector [5], word vector [6], sliding windows [7], intervals [8]. In the present study we attempt to recognize emotions from the speech signal given a short command (approximately between 2 and 4 seconds), without getting any profit from context or linguistic information. In the long term, our goal is to have a speaker and language independent emotion classifier. Such a challenging purpose, leads us to deal only with global acoustic features, computed for a whole utterance or command, which seem to have the favor of many recent studies [9] [10].

Currently our emotional database contains around 40 commands per emotion and per speaker. In order to have a reasonable training set, according to the parameters estimation in a neural network (NN) classifier, feature selection methods had to be applied to find a reduced set of features.

As it was mentioned previously, prosody and quality features are processed in a different way as independent feature subspaces.

3.1. Prosody features

A set of 37 features has been used, as a first approach to localize a certain utterance in the arousal dimension. 26 of these features model logarithmic F0, energy and durational aspects, as proposed by Batliner [5] (see also [11] [12]).

- Logarithmic F0: maximum, minimum, maximum and minimum position, mean, standard deviation, regression coefficients, mean square error for regression coefficients, and F0 for the first and last voiced frame.
- Energy: maximum, maximum and minimum positions, mean, regression coefficients, and mean square error for regression coefficients.
- Durational aspects: number of voiced and unvoiced regions, number of voiced and unvoiced frames, longest voiced and unvoiced region, ratio of number of voiced vs. unvoiced frames, ratio of number of voiced vs. unvoiced regions, ratio of number of voiced vs. total number of frames, ratio of number of voiced vs. total number of regions.

Additional 11 features, model jitter, tremor and pitch derivative statistics, as suggested by Dellaert [9].

3.2. Quality features

Latest emotion recognition approaches also include information related to articulatory precision or vocal tract properties, e.g. formant structure, as in [10]. There is perceptual evidence, in terms of emotions expression, of the additional importance of phonatory quality parameters, i.e. auditory qualities derived from variations in the glottal excitation [13] [14].

We chose 16 quality features, describing the first three formants, their bandwidths, harmonic to noise ratio, spectral energy distribution, voice to unvoiced energy ratio, and glottal flow [15], to classify the quality aspect of the emotion. Basically, those features are used to distinguish between happy and angry, as well as sad and bored.

All the quality features described were obtained using the phonetic analysis software PRAAT.

3.3. Feature Selection

NNs are able to handle redundant and irrelevant features, assuming that enough training patterns are available to estimate the weights during the learning process.

Since our database does not satisfy those requirements, linear regression models have been applied as feature selection method, to find a subset of features for improving prediction accuracy. That means, modeling emotions by linear combination of features and selecting only those, which significantly modify the model. For some specific cases (speaker-independent), also quadratic regression models were used. Both models were implemented using R, a language and environment for statistical computing and graphics.

4. CLASSIFICATION

The research has been conducted in two steps. First, the basic approach, necessity of sequential classifiers using emotional space concept, was verified in the framework of a speaker-dependent scenario. Later on, the results were extended to the speaker independent case.

Several learning algorithms were evaluated for different NN configurations, in order to find the most suitable one for our purpose (c.f. below), using the Stuttgart Neural Network Simulator SNNS. For all of them, the classification of the patterns depends on the unit with the highest output, i.e. WTA (winner takes all). Pruning algorithms for learning have been applied on top of the best learning algorithm found. These techniques try to make NN smaller by pruning unnecessary edges and nodes. In our case "Magnitude Based Pruning" is applied, which removes after each training the link with the smallest weight. Though this method is very simple, it rarely yields worse results than the more sophisticated ones.

In all the experiments concerning prosodic features and three levels of arousal classification, the neutral database was duplicated in order to balance the amount of input patterns to the NN. As reference label, the intended emotions were used. The labels will be manually corrected in the future, since a certain percentage of the errors are due to the labeling.

4.1. Speaker-dependent experiments

As a first approach to verify the relationship between the different feature types, i.e. quality and prosody, with the two dimensions of the emotional space, i.e. arousal and pleasure, the speaker-dependent case was addressed as a reduced problem.

An experiment was carried out taking the whole set of 37 prosodic features as input to the NN, with no hidden layer, to assess the confusability between the five emotional states. The confusion matrix, obtained averaging the activation of the output nodes over the whole test database, is presented in Table 1:

OutNode / Reference	Angry	Bored	Happy	Neutral	Sad
Angry	**0,46**	0.09	0,24	0,19	0.1
Bored	0.1	**0,39**	0.09	0,2	0,34
Happy	0,25	0.09	**0,59**	0,14	0.06
Neutral	0,2	0,19	0,14	**0,3**	0,2
Sad	0.12	0,34	0.08	0,22	**0,39**

Table 1: Confusion matrix from the average activation of the output nodes.

Although the recognition rate is not particularly high, the output values clearly differentiate the position of the emotional state on the arousal axis. Table 1 shows that emotions placed close to each other in the arousal dimension, i.e. angry-happy and sad-bored, obtain similar values in the activation of the output nodes. If the reference emotion is angry, the activation of the output node for happy is also relatively high, while for sad and bored is pretty low. The observation is consistent for the rest of the reference emotions. Consequently, the new approach becomes indispensable.

From the 37 feature set, 10 features were selected via linear regression models. We took as discrimination classes three levels in the activation-arousal axis, i.e. the five emotional states are assigned to the levels as follows: high=angry/happy, medium=neutral, and low=sad/bored. After several experiments, the reduced set of features turned out to give better performance than the whole set.

Best results were obtained with two hidden layer NN, having 10 and 5 nodes in the first and second layers. The network achieves optimal performance, if first a pruning algorithm discards some features, and afterwards, R-Prop learning algorithm is applied [16]. Such algorithm overcomes the inherent disadvantages of the pure gradient-descent technique of the original Backpropagation procedure, performing an adaptation of the weight update-values according to the behaviour of the errorfunction. With that configuration it is possible to achieve 83.7% recognition rate in the three levels of arousal dimension decision (see Table 2).

From the results reflected in Table 2, we can confirm that happy and angry are far enough from sad and bored, to avoid any confusion along the arousal dimension. Being in the middle, is not the only reason for the neutral emotion to be confused with the others. Owing to the intrinsic properties of commands intended for a pet, it is very difficult to avoid any kind of emotion in the pronunciation. The commands themselves contain some predefined emotion in the meaning.

OutNode Reference	High	Neutral	Low
High	**82.1**	17.9	0
Neutral	10.3	**82.8**	6.9
Low	0	13	**87**

Table 2: Speaker-dependent arousal classification. The average recognition rate is 83.7%

Similar experiments in terms of NN configuration and learning algorithms were performed for the classification of the emotional states on the pleasure axis, through the 16 quality features. The goal was to differentiate between happy and angry and between bored and sad. Table 3 shows the results obtained with no hidden layer, pruning and R-Prop learning algorithm, for happy versus angry, and the results obtained with no hidden layer, pruning and Standard-Backpropagation learning algorithm, for bored versus sad.

This pleasure is based on the reference labels of the original database, i.e. the results of the first classifier are not taken into account. It would be interesting to apply this classification on the corresponding correct outputs of the arousal classifier, i.e. happy-angry on the 82.1% correctly classified as high arousal, and sad-bored on the 87% correctly classified as low arousal.

OutNode Reference	Happy	Angry	OutNode Reference	Bored	Sad
Happy	**75**	25	**Bored**	**76**	24
Angry	28	**72**	**Sad**	44	**56**

Table 3: Speaker-dependent happy-angry (73.5% RR), bored-sad classification (66% RR)

There is clearly a difficulty in separating bored and sad in the pleasure axis. Indeed, they are much closer than happy and angry, which are actually opposite extremes.

4.2. Speaker-independent experiments

To extend the problem to the case of speaker-independent, several experiments were conducted, following the same idea as in the speaker-dependent case. Results obtained with no hidden layer NN, Chunkwise-Backpropagation learning algorithm, and 37 prosodic features set, are presented in Table 4.

There is clear evidence that neutral emotion has a recognition rate close to chance. As we mentioned previously, the situations designed in order to provoke emotions more spontaneously, were not very suitable to evoke neutral emotion. As a result of some listening tests, we decided to record neutral commands as read speech, i.e. without any context simulation behind.

OutNode Reference	High	Neutral	Low
High	**72.4**	17.8	9.8
Neutral	34	**35.5**	35.5
Low	13.9	15.7	**70.4**

Table 4: Speaker-independent arousal classification. The average recognition rate is 59.3%

Results in Table 5, reproduce the same experiment conditions as in Table 4, but training the NN with the new neutral commands. The abrupt decrease in the accuracy of neutral commands recognition confirms, that the assumed neutral commands are either very active or passive, but rarely positioned in the middle.

OutNode Reference	High	Neutral	Low
High	**67.1**	18	14.9
Neutral	42.2	**11.5**	46.2
Low	14.2	4.6	**81.3**

Table 5: Speaker-independent arousal classification. Neutral commands from read speech for training. The average recognition rate is 61%

OutNode Reference	High	Neutral	Low
High	**68.1**	17.3	14.5
Neutral	14.3	**81.6**	4.1
Low	14.4	3.7	81.8

Table 6: Speaker independent arousal classification. Neutral commands from read speech for training and testing. The average recognition rate is 77%.

However, when the same test is performed, but substituting the new neutral commands also in the test set, the recognition rate for neutral level becomes comparable with the other levels in the arousal axis, as is shown in Table 6. There is a substantial increase in the average recognition rate from 60% to 77%.

In accordance with the idea of having two classifications, one in the arousal, and the other one in the pleasure dimension, the 16 quality features were used to classify angry versus happy, and bored versus sad, for the speaker-independent approach. The maximum recognition rate achieved for both classifications is slightly better than 60%. Since most quality features are very speaker-dependent, it would be convenient to include some speaker dependencies, e.g. age and gender, in the classification design.

5. DISCUSSION

In this paper, it is reported that experiments using only prosody features suffer from a clear weakness, trying to differentiate emotions, which are placed very close in the arousal dimension, but in the contrary being quite separated in the pleasure dimension. Applying prosody features to the specific case of arousal dimension gives pretty good results for the three level classification problem, i.e. differentiating between low, neutral and high. Specially, emotional states with high and low level of arousal are hardly ever confused. Quality features are useful to be able to give a final decision about the emotional state, assuming the level of arousal is already known. There is still a challenge, trying to classify emotions, which are very close along the pleasure axis, e.g. sad and bored.

All the above experiments indicate that an alternative way of classifying emotions can be seen as finding a place in the emotional space, and infer from such location and from additional information, i.e. context, application, if available, the intended emotion. Obviously emotions that are close in the emotional space tend to be confused more often, and furthermore, positions in the pleasure dimension can be considered very tricky to allocate. There is still further research to be done concerning the speaker dependencies of quality features, as well as the association of new possible features to a third dimension of the emotional space, i.e. power dimension.

We tried to design a database as close as possible to the real situation of our application. Along the recordings and subsequent experiments, we came across the fact, that we cannot expect pure neutral emotion in our target application. People do not speak neutral to pets, and actually ... do people speak neutral at all? That's an open question, whose answer is strongly related to the specific situation, in our case, to the final application. Emotional expressions are very contingent upon the environment. They are not the same at home, in a party, or at the office. Therefore, an emotional database reflecting the real situation is a crucial factor.

Moreover features should be selected according to the object application, paying special attention to how the target emotions are distributed in the emotional space. Our research has only focused on acoustic events. More elaborated features, such as lexical features, i.e. word-final syllable, syllable with lexical word accent, semantic features, i.e. word containing some emotional meaning, affect burst [17], can significantly benefit the recognition accuracy. Most of them are somehow language dependent. That is the main reason for excluding them from our studies.

ACKNOWLEDGMENTS

The results published in the paper are part of a Diploma Thesis performed in Sony (International) Europe, ATCS, in collaboration with the E.T.S.I. Telecommunication (Madrid) [18]. We would like to express our gratitude to Thomas Kemp and Krzysztof Marasek for their contributions in various forms to the material presented in this paper. We would also like to extend our thanks to The University of Erlangen, Chair of Pattern Recognition, for their cooperation and support with software tools for basic feature extraction.

REFERENCES

[1] Pereira, C., "Dimensions of emotional meaning in speech", *in [19].*

[2] Cowie, R., "Describing the Emotional States Expressed in Speech", *in [19].*

[3] Scherer, K., "A Cross-Cultural Investigation of Emotion Inferences from voice and Speech: Implications for Speech Technology", *ICSLP 2000*, Beijing.

[4] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schröder, M., "FEELTRACE: An Instrument for Recording Perceived Emotion in Real Time", *in [19].*

[5] Batliner, A., Fisher, K., Huber, R., Spilker, J., and Nöth, E., "Desperately Seeking Emotions: Actors, Wizards, and Human Beings" , *in [19].*

[6] Huber, R., Nöth, E., Batliner, A., Buckow, J., Warnke, V., and Niemann, H., "You BEEP Machine – Emotion in Automatic Speech Understanding Systems", *TSD'98*, Brno, Masaryc University.

[7] Amir, N., "Classifying Emotions in Speech: a Comparison of Methods", *EUROSPEECH 2001*, Scandinavia.

[8] Nogueiras, A., Moreno, A., Bonafonte, A., and Mariño, J.B., "Speech Emotion Recognition Using Hidden Markov Models", *EUROSPEECH 2001*, Scandinavia.

[9] Dellaert, F., Polzin, T., Waibel, A., "Recognizing Emotion in Speech", *ICSLP 96*, Delaware.

[10] Petrushin, V. A., "Emotion Recognition in speech Signal: Experimental Study, Development, and Application", *ICSLP 2000*, Beijing.

[11] Kießling, A., "Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung", *Berichte aus der Informatik*, Aachen 1997.

[12] Kompe, R., "Prosody in Speech Understanding Systems", *Lecture Notes for Artificial Intelligence*, Berlin 1997.

[13] Klasmeyer. G., "The Perceptual Importance of Selected Voice Quality Parameters", *ICASSP 97*, Munich.

[14] Rank, E., and Pirker, H., "Generating Emotional Speech with a Concatenative Synthesizer", *ICSLP 98*, Sydney.

[15] Henrich, N., d'Alessandro, C., and Doval, B., "Spectral Correlates of Voice Open Quotient and Glottal Flow Asymmetry: Theory, Limits and Experimental Data", *EUROSPEECH 2001*, Scandinavia.

[16] Riedmiller M., and Braun H., "RPROP-A Fast Adaptive Learning Algorithm", *to appear in ISCIS VII.*

[17] Schröder, M., "Experimental Study of Affect Bursts", *in [19].*

[18] Santos, R., "Emotional Speech Recognition", Technical University of Madrid, E.T.S.I.T, Grupo de Tecnología del Habla. *To appear.*

[19] *ISCA workshop on Speech and Emotion*, Belfast 2000.

http://www-gth.die.upm.es/partners/sony/main.html