Since nouns are by far the most elaborated category in Wordnet, we considered as correct UWs the set of unique UWs, and as incorrect the set of duplicate UWs. As can be seen from table 4, the rate of duplicate UWs for nouns is less than 2%, a good result for the most polysemous syntactic category. Surprisingly, the results for verbs is rather good (less that 5% of error rate), although we assume that semantic arguments of verbs require human revision. On the other hand, both adjectives and adverbs yield an error rate quite high (around 14%). The possible reason for such an error rate may lie in the fact that the main lexical relations present in Wordnet are synonymy and hypernym, natural relations for nouns but not for predicates like adjectives or adverbs.

Bibliography

[Bhattacharyya et al, 2004]. N. Verma and P. Bhattacharyya, *Automatic Lexicon Generation through WordNet,* Global WordNet Conference (GWC-2004), Czech Republic. Jan, 2004

[Boguslavsky et al, 2005]. Boguslavsky, I., Cardeñosa J., Gallardo, C., and Iraola, L. The UNL Initiative: An Overview. Lecture Notes in Computer Science. Volume 3406/2005, pp 377-387. Springer Berlin / Heidelberg: 2005. ISBN 978-3-540-24523-0

[Fellbaum, 1998]. Fellbaum, C., (ed): WordNet: An Electronic Lexical Database. Language, Speech, and Communication Series, MIT Press (1998)

[Uchida et al, 2005] Universal Networking Language (UNL). Specifications Version 2005. Edition 2006. 30 August 2006. http://www.undl.org/unlsys/unl/unl2005-e2006/

Authors' Information

Igor Boguslavsky *– Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: igor@opera.dia.fi.upm.es http://www.vai.dia.fi.upm.es*

Juan Bekios *– Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: juan.bekios@opera.dia.fi.upm.es. http://www.vai.dia.fi.upm.es*

Jesús Cardeñosa *– Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: carde@opera.dia.fi.upm.es. http://www.vai.dia.fi.upm.es*

Carolina Gallardo *– Group of Validation and Industrial Applications. Escuela Universitaria de Informática. Universidad Politécnica de Madrid. Carretera de Valencia Km.7. 28041 Madrid; email: cgallardo@eui.upm.es. http://www.vai.dia.fi.upm.es*

A COGNITIVE SCIENCE REASONING IN RECOGNITION OF EMOTIONS IN AUDIO-VISUAL SPEECH

Velina Slavova, Werner Verhelst, Hichem Sahli

Abstract: *In this report we summarize the state-of-the-art of speech emotion recognition from the signal processing point of view. On the bases of multi-corporal experiments with machine-learning classifiers, the observation is made that existing approaches for supervised machine learning lead to database dependent classifiers which can not be applied for multi-language speech emotion recognition without additional training because they discriminate the emotion classes following the used training language. As there are experimental results showing that Humans can perform language independent categorisation, we made a parallel between machine recognition and the cognitive process and tried to discover the sources of these divergent results. The analysis suggests that the main difference is that the speech perception allows extraction of language independent features although language dependent features are incorporated in all levels of the speech signal and play as a strong discriminative function in human perception. Based on several results in related domains, we have suggested that in addition, the cognitive process of emotion-recognition is based on categorisation, assisted by some hierarchical structure of the emotional categories, existing in the cognitive space of all humans. We propose a strategy for developing language independent machine emotion recognition, related to the identification of language independent speech features and the use of additional information from visual (expression) features.*

ACM Classification Keywords: *I.2 Artificial Intelligence, 1.2.0.Cognitive simulation, 1.2.7. Natural language processing - Speech recognition and synthesis*

Introduction

Traditional human machine interaction is normally based on passive instruments such as keyboard, mouse, etc. Emotion is one of the most important features of humans. Without the ability of emotion processing, computers and robots cannot communicate with humans in a natural way. It is therefore expected that computers and robots should process emotion and interact with human users in a natural way. Affective Computing and Intelligent Interaction is a key technology to enable computers to observe, understand and synthesize emotions, and to behave vividly. Affective computing aims at the automatic recognition and synthesis of emotions in speech, facial expressions, or any other biological communication channel (Picard, 1997). In fact, existing automatic speech recognition systems can benefit from the extra information that emotion recognition can provide (Ten Bosch, 2003; Dusan and Rabiner, 2005). In (Shriberg, 2005), the authors emphasize the importance of modelling non-linguistic information embedded in speech to better understand the properties of natural speech. Such understanding of natural speech is beneficial for the development of human-machine dialog systems. Several applications call for recognition only of the emotions in the speech, without processing the linguistic content. Such systems should be language independent.

During the last years the research concentrated in all these problems. As example, we can site the HUMAINE (Human-Machine Interaction Network on Emotion) a Network of Excellence in the EU's Sixth Framework Programme IST (Information Society Technologies). The thematic priority of HUMAINE aims to lay the foundations for European development of systems that can register, model and/or influence human emotional and emotion-related states and processes - 'emotion-oriented systems'. For the proposed reasoning herein, we used several analyses and results of this research network, available on [http://emotion-research.net/].

Automatic emotion recognition

Automatic recognition of emotions in speech aims at building classifiers (or models) for classifying emotions in unseen emotional speech. The data-driven approaches to the classification of emotions in speech use supervised machine learning algorithms (neural networks, support vector machines, etc.) that are trained on patterns of speech prosody. The training is performed with utterances or other speech instances, labelled with a previously chosen set of emotions. Such labelled speech instances are taken from databases of emotional speech. Machine learned classifiers (ML-classifiers) can categorize other speech instances from the same database, according to the labels, used in the training procedure.

In general, the systems for speech analysis (speech recognition, speaker verification, emotion recognition) use techniques for *extraction* of *relevant* characteristics from the raw signal. Concerning emotions, the relevant information is the *Prosody* (broadly determined as: *Intonation* – the way in which pitch changes over time, *Intensity* – the changes in intensity over time and *Rhythm* – segment's durations vs. time) and in the *Voice quality (*measured in spectral characteristics).

Table 1. Feature set used in the AIBO approach (Oudeyer, 2003).

Acoustic features	Derived series of:	Statistics on the der. series,
-intensity -lowpass intensity -highpass intensity -pitch -norm of absolute Vector derivative of the first 10 MFCC Components *(MFCC - Mel-frequency cepstral coefficients)*	-minima, -maxima, -durations between local extrema - the feature series itself	-Mean, -maximum, -minimum, -range, -variance, -median, -first quartile, -third quartile, -inter-quartile range, -Mean absolute value of the local derivative

In Table 1 lists the features used in one of the contemporary feature extraction approaches developed for the Sony's robotic dog AIBO (Oudeyer, 2003). Table 2 illustrates another feature set used for the "segment based approach" (SBA) (Shami and Kamel, 2005). The size of the feature-vectors, provided as an input to the machine learning algorithm is practically not limited. One of the strategies applied for building a ML-classifier is to construct a feature vector with "everything that can be calculated" according to the reasoning that "the more information is collected from the raw signal, the better it is". This strategy is often used in the practice. There exist classifiers with feature vectors of hundreds of values. The big length of the input vector reduces the performance of the classifier. The next step in this strategy is to discover the features which discriminate the speech data (to the training labels) and to discard the non-discriminative features.

Table 2 Feature set used in the Segment-based approach (SBA) (Shami and Verhelst, 2007)

Pitch	Intensity	Speech Rate
-Variance -Slope -Mean -Range -Max -Sum of Abs Delta	-Variance -Mean -Max	-Sum of Absolute Delta MFCC -Var. of Sum of Abs. Delta MFCC -Duration

Speech research is already at a mature stage. Some studies focus on finding the most relevant acoustic features of emotions in speech as in (Fernandez and Picard, 2005; Cichosz and Slot, 2005). Other studies search for the best machine learning algorithm to use in constructing the classifier or investigate different classifier architectures. Lately, research has shifted towards investigating the proper time scale (utterances, segments) to use when extracting features as in (Shami and Kamel, 2005; Katz et al., 1996). Segment based approaches try to model the shape of acoustic contours more closely. There are also attempts to take into account phoneme-level prosodic and spectral parameters. (Lee S. et all., 2006(b), Lee, C.M. et all, 2004, Bulut et all 2005) All these efforts have lead to better and better ML-classifiers.

In all of the mentioned studies the classifiers were trained on one single speech corpus. It is known that ML-classifiers do not perform well on samples from other databases. There are no studies concerned with the problem of dependency of classifiers on the used speech corpora.

Multi-corpora recognition

A recent study, conduced at VUB-ETRO (Shami M., Verhelst W., 2007) treats the problem of multi-corpus training and testing of ML-classifiers. The study is based on the use of four emotional speech corpora: *Kismet, BabyEars* (both in American English), *Danish* (in Danish), and *Berlin* (in German). The four databases were grouped in two pairs: 1. Kismet-BabyEars pair, which contains infant directed affective speech, and 2. Berlin-Danish pair, containing adult directed emotional speech. The other difference between the two database pairs (DB-pairs) is in the length of the utterances (the infant-directed DB-pair contains shorter utterances).

Two approaches, corresponding to the two feature vectors (tables 1 and 2), were used - the segment based approach SBA and the utterance based approach AIBO. The two considered main questions have been: "When a classifier is trained to recognize a given emotion in one database, does it recognize the considered emotion in another database?" and "How does an ML-classifier perform if it is trained and tested on merged corpora, in other words – can it generalize?" The "behaviour" of the classifiers described in Shami and Verhelst (2007) lead to several fundamental questions concerning the recognition of emotion in speech.

The speech entities in the four corpora contain speech instances for different sets of basic emotions, some of them overlapping. Table 3 and table 4 give the emotion labels (E-labels) and the numbers of speech instances labelled with them in each of the databases.

For the multi-corporal testing of classifiers, first the speech instances of the non-corresponding E-labels in each pair were removed from the initial databases. In this way "reduced" databases were obtained with only the common E-labels for the pair classes. The following experiments were done:

Table 3 Emotion Classes in Kismet and BabyEars databases (Shami and Verhelst, 2007)

Kismet		Baby Ears	
*Approval	185	*Approval	212
*Attention	166	*Attention	149
*Prohibition	188	*Prohibition	148
*Soothing	143		
*Neutral	320		

Table 4 Emotion Classes in Berlin and Danish databases (Shami and Verhelst, 2007)

Berlin		Danish	
*Anger	127	*Angry	52
*Sadness	52	*Sad	52
*Happiness	64	*Happy	51
*Neutral	78	*Neutral	133
*Fear	55	*Surprised	52
*Boredom	79		
*Disgust	38		

Between-corpora experiment: Training on the one and testing on the other database of the pair. The results are not surprising: it seems that training on one database and testing on another database is not possible in general with the existing approaches.

Integrated corpus experiment: Merge databases into one “Integrated corpus” (for each pair).

First condition: Merge the classes from the corresponding E-label into a joint “common” class. For example, the instances from Kismet*Approval and from Baby ears*Approval were fused in a novel class: Integrated*Approval. The ML-classifiers were trained and tested on the fused classes. They “learned” them and “performed” the recognition task surprisingly well (classification accuracies: 74.60% for Kismet-Baby Ears and 72.2 % for Berlin-Danish[5]).

Second condition: Keep in the Integrated corpus the classes as they were in the initial databases of the pairs. The ML-classifiers were trained and tested in the integrated corpora on the old classes.

The classification accuracies obtained in the two “integrated” conditions were similar: the accuracy of the classifiers in an “integrated corpus” could be seen as average of the accuracies in the one and in the other databases of the pair. So, the use of a heterogeneous corpus does not lead to a notable deterioration in classification accuracy. This is a very good practical result, as it is known that the less uniform the training corpus is, the less accurate the classifier is. And, on the other hand, a classifier learned using heterogeneous corpora is more robust. One important conclusion, given in this study, is that the existing approaches for classification of emotions in speech are efficient enough to construct a single classifier, based on larger training data from different corpora. From the practical point of view, the result gives a solution for building classifiers in integrated corpora with shared emotion classes.

Here the results have been analysed from the point of view of another interesting finding, related to the representation of the emotion classes in the feature space. The result is seen in the Second “integrated” condition, were the emotion-classes have been preserved as they were in the initial databases of the pairs.

Table 5. Confusion matrix of Berlin-Danish Integrated corpus (Shami and Verhelst 2007)

A	B	C	D	E	F	G	H	←	classified as
74	1	2	1	0	0	0	0	A	Berlin*Neutral
3	36	0	25	0	0	0	0	B	Berlin*Happy
4	0	48	0	0	0	0	0	C	Berlin*Sadness
1	25	0	101	0	0	0	0	D	Berlin*Anger
0	0	0	0	106	2	17	8	E	Danish*Neutral
0	0	0	0	7	29	2	13	F	Danish*Happy
0	0	0	0	21	4	27	0	G	Danish*Sad
0	0	0	0	11	16	2	23	H	Danish*Angry

[5] The results were also compared for different machine-learning algorithms, not given here

For the Berlin-Danish integrated corpus, it turned out that classifiers never "confuse" for example Berlin*Anger and Danish*Angry. The confusion matrix of the Berlin-Danish pair is given in Table 5. It is seen that instances belonging to one of the databases are never "taken" as instances belonging to the other database. Automatic clustering (using the K-means clustering algorithm) showed that the same emotion-classes from the two databases are represented on different clusters and even that the entire databases doesn't share any cluster.

For the Kismet- BabyEars pair there was a small tendency of generalization over the emotions, as some instances of BabyEars were "confused" with the equivalent emotion in Kismet (but never the reverse). Automatic clustering showed that the two databases share four (of the six) clusters and that when there are classes from both databases on one cluster, these classes represent one and the same emotion.

Why these results? This could be linked to the language in which the emotions are expressed. Or to the nature of the emotions - Kismet/BabyEars contains infant directed communicative intents, generally regarded as culture and language independent (Fernald, 1992). In any case, the question which arises at this point is related to the recognition accuracy (RA) of humans on this task.

Comparison with cognitive processes

Human capacity to recognize emotions only from speech, reported in the literature, is between 60% and 85%, depending on the experiments, the emotion classes and other additional circumstances. For example, human listening recognition accuracy has been evaluated to be 79% for stimuli from the BabyEars database (Shami M., Verhelst W., 2006). On Danish database it is 67% (Engberg and Hansen, 1996). What about the vagueness of the different expressed emotions for the listeners? The reported experimental results in the literature show that, depending on the experiment, listeners recognize with unequal success the emotions Anger, Disgust, Fear, Happiness, Sadness, and Surprise, often supported as being basic for humans. For example, human RA is best for Anger and worse for Happiness in the experiment of Lee (Lee C.M. et all, 2004). In Danish database, humans recognise best Sad and worst Happy. The abundance of such examples leads to doubt that the target emotions are well expressed. One can also wonder whether participants share one and the same concept for the label "Sad".

The important point is that, in almost all last year's reported mono-corporal results, the recognition accuracy of the classifiers is comparable with the human categorization capacities for the samples, stored in the corresponding databases. The resemblances between the classifiers and the human evaluators within the same database goes further: as it has been reported by Shami and Verhelst in (Shami and Verhelst, 2007), the use of the SBA approach on the Danish database lead to a classifier which makes the same mistakes as humans. Listeners recognise best *Sad, the classifier does the same; listeners confuse *Surprise with *Happiness and *Neutral with *Sadness, the classifier does the same. From the modelling point of view, that means that the used feature-space is a good projection of the human cognitive space, which contains also models of acoustic parameters of speech emotions. The hope is that such a kind of mapping will be available for the multi-corporal experiment. Unfortunately, that is not the case.

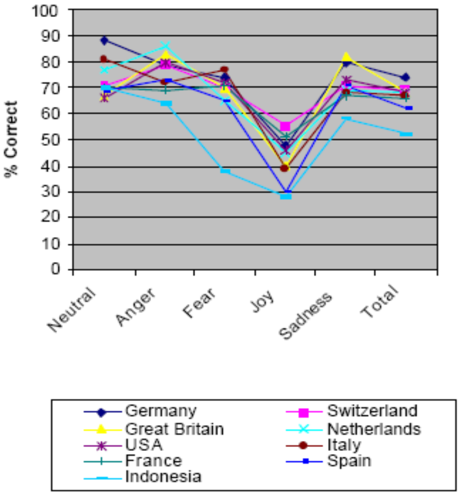

Fig.1. Result for human recognition of speech-emotion across languages and cultures (Scherer K., 2000)

Suppose that the aim is to build a multilingual emotional classifier. The corpus should include labelled classes of speech instances from several databases. A classifier will "learn" Danish*Anger, German*Angry, Polish*Angriness etc. These classes could be fused in one class; the classifier will learn the image of this composed class and will become more robust. As it is demonstrated with the multi-corporal experiments, classifiers "learn" quite well the images of composed emotion-classes, represented on non intersecting clusters in the feature space. One may speculate and fuse Danish*Sad with Berlin*Happy to train the classifier on the novel class "Integrated*Potatoes". The expectation, looking at the confusion matrix of Berlin/Danish pair, is that the classifier will "learn" that class.

A known work in the domain of speech and emotion is the study of Klaus Scherer (Scherer K., 2000, Scherer et all 2001), reporting results (fig. 1) of a multi-language emotion encoding/decoding experiment. Scherer used a set of basic emotions: {*fear*, *joy*, *sadness, anger and neutral}* and tested human recognition accuracy on samples of emotional speech, containing content-free utterances composed of phonological units from many different Indo-European languages. That was done in nine countries, on three continents. In all cases human recognition accuracy was substantially better than chance and showed an overall accuracy of 66% across all emotions and countries, suggesting the existence of similar inference rules from vocal expression across cultures. This key-suggestion is widely accepted in the speech-emotion scientific domain. So, it turns out that there are common acoustic images of emotional speech in the human cognitive space, and they are applied with a good result even for utterances of a never heard or even invented[6] language.

Scherer's study found differences in the results across the countries: the highest accuracies were obtained by native speakers of Germanic languages (Dutch and English), followed by Romanic languages (Italian, French, and Spanish). The lowest recognition rate was obtained for the only country studied that does not belong to the Indo-European language family, Indonesian.

Here a hypothesis could be made: the worse recognition result is obtained when using only the basic "perceptive" features which permit to categorize speech-emotions in the cognitive space.

The better recognition accuracy of the listeners from the other language groups can in this case be explained in two ways: 1. listeners perceive in the samples features *in addition* of the basic perceptive features; 2. the emotion categories in the cognitive space of these listeners were better fitting with the emotion-labels of the samples.

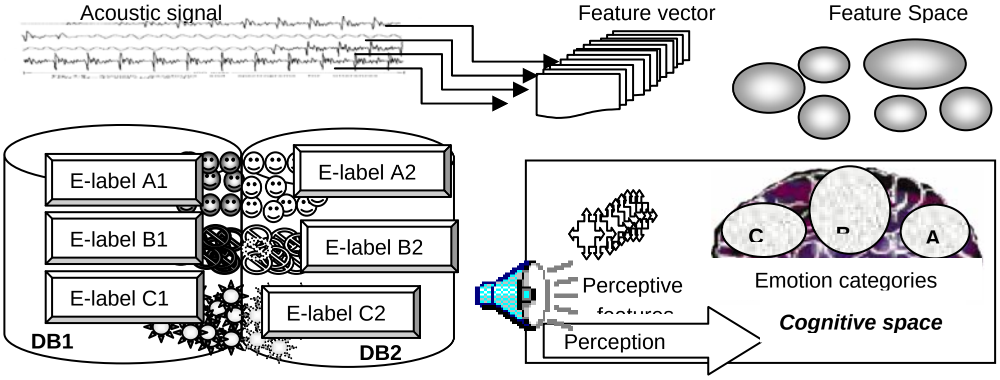

Fig 2. Scheme of the analogy between cognitive process and machine recognition.

The results of the multi-corporal machine leaning experiment are not comparable to the results in the Scherer's experiment. Figure 2 illustrates an analogy between the machine recognition and the human recognition. Classifiers depend exclusively on the labelled training data and humans perform the task without being trained. It is clear that the perceptive features used by humans permit generalisation and categorization of the signal, but the features extracted for the machine classifier do not allow that. If an ideal feature space could be employed, similar emotions belonging to different databases should be assigned to the same clusters, as humans do.

Several atomic hypotheses could be made at this point. For example:

A. There is not enough emotion-relevant information captured by the feature vector.
B. There is language dependent information captured by the feature vector.
C. The perceptual features allow humans to categorize to more general categories. The cognitive space has a structure which permits them to path the sub-category, used in the proposed label.

Concerning the first two hypotheses, a lot of efforts have been made to ameliorate the feature extraction and to find relevant feature vectors. Acoustic correlates of specific emotional categories are investigated in terms of pitch, energy, temporal and spectral parameters, on suprasegmental, segmental and even on phoneme level. This is in aim to extract more and more emotion-relevant information (HUMAINE, 2004a). The question about the language dependence of the used features stays open. Language dependent information is incorporated at all

[6] This is used also in the domain of synthesis of emotional speech – the produced speech is not in any language,

levels of speech prosody. Newborns discriminate different languages. Babies do that without relying on phonemic cues, but on the basis of rhythmic and intonational cues only (Ramus, F., 2002). We may expect that machine-learned classifiers do the same – they discriminate languages. So, the task is to present to the classifier only language independent information. A classic idea is to look for acoustic correlates of emotion in music, which corresponds a lot to Scherer's reasoning. There is a lot of research in this direction (Kim 2004; Kim et all 2004). But the speech signal is much more complex. How could the language dependent and the language independent ingredients of the extracted features be separated, and how to do this on the suprasegmental, the segmental and, why not, on the phoneme level in order to take only the features with pure information about emotions only? Humans can do that. So, it should be possible to do so. Obviously, such a task demands a lot of specific research.

Hypothesis C. requires a separate approach. The C hypothesis explains the good performance of humans in speech emotion recognition with the structure of the cognitive space of emotion categories.

Emotions

To study relations between speech and emotion, it is necessary to derive methods describing emotion. Although there have been numerous studies with regards to both the psychological and the engineering aspect of emotions, it is still not clear how to define and how to categorize human emotions. There are two basic approaches used.

The first approach is "discrete" (Fig. 3). Emotion categories are determined as entities with names and descriptions. Several theorists argue that a few emotions are basic or primary (Ekman, 1992; Izard, 1993). The emotions of anger, disgust, fear, joy, sadness, and surprise are often supported as being basic from evolutionary, developmental, and cross-cultural studies. That theoretical approach is convenient for the purposes of machine learning, as it provides directly labels for the training data. In speech emotion recognition, the attempts have mostly concentrated on a small number of discrete, extreme emotions, in aim to obtain maximally distinguishable prosodic profiles.

Fig. 3. Emotion categories -Labels

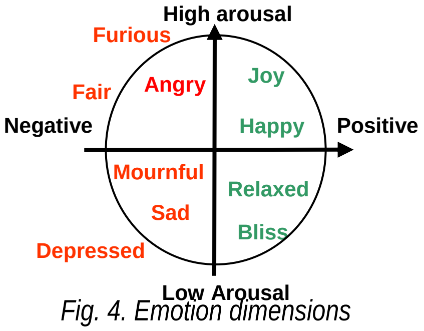

Fig. 4. Emotion dimensions

The other approach is "continuous". The basic properties of the emotional states are described in a continuous space of "emotion dimensions" (fig. 4). The most frequently encountered emotion dimensions are activation (the degree of readiness to act) and evaluation ("valence" in terms of positive and negative). They provide a taxonomy allowing simple distance measures between emotions.

The central question for the experts in the field of speech emotion is: what should be recognized, *emotional categories and/or dimensions*. The performance of human participants and the performance of an automatic recognition system are totally dependent on the number and the degree of differentiation of the emotion categories/dimensions that have to be discriminated. The consensus of the experts from HUMAINE is that "labelling schemes based on traditional divisions of emotion into 'primary' or 'basic' categories is not relevant" (HUMAINE 2004, b). So, the task has turned to cluster the emotional states with names in the continuous space. Several approaches have been developed for this purpose (Douglas-Cowie, et al. 2003; Devillers et al. 2005).

A large study was conducted within the international project AMI (Wan et al., 2005) to determine the most suitable emotion labels for the specific context of *meetings*. One of the contemporary labelling schemes *FeelTrace* (Cowie et al. 2000), which is based on the above mentioned emotion dimensions, was used. A listing of 243 terms describing emotions was compiled from the lists of three research centres. These emotion-labels were first *clustered by meaning* by the project's experts. After that, participants from various companies and professions evaluated the position of the separate emotions on the axes. Figure 5 gives the plot of the participant's evaluation of the emotions from one meaning-cluster.

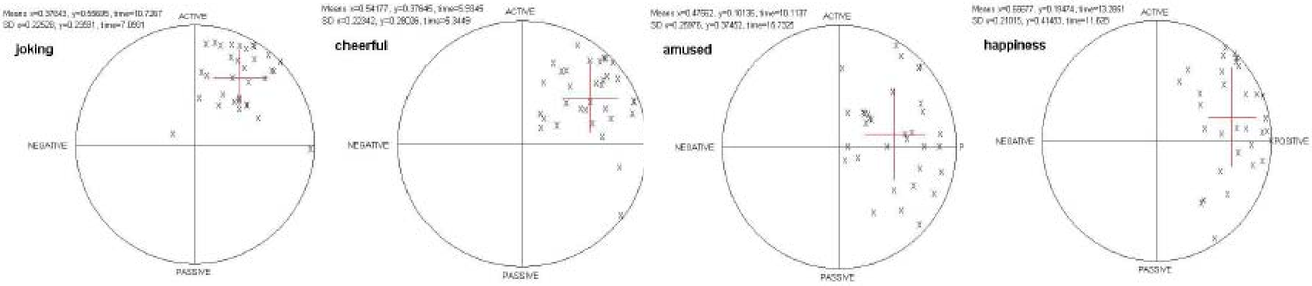

Fig. 5. Results of landmark placement survey for joking, amused, cheerful and happiness (Wan et all, 2005)

The first observation when analysing this experiment is that one can cluster emotion-names *by meaning*. The second observation is that the others agree on the same meaning-cluster, as they locate the emotions' names from the cluster on approximately the same place. The last observation is that the dispersion of participants' evaluations covers "semicircles" and quarters of the plane. One may suppose that in the cognitive space there exist "generalized" categories, in correspondence of the clusters. In any case, the agreement of the participants on the meaning of the axes is evident. Where meaning and categories appear, there should be an attempt to analyse the cognitive processes underlying emotion.

A Possible Cognitive Science Reasoning

At a first stage one should check if there are physiological phenomena, leading human beings to "innate" perception of the dimensions of emotion properties. Emotion-related biological changes are well documented. Recent studies (Kim 2004, Kim at all. 2004) also showed that parameters from measurements as cardiograms, encephalograms, respiration and skin conductivity, are highly correlated with the emotional dimensions. The study was conduced by provoking emotive states using music stimuli. As it is illustrated in fig. 6, on the Arousal axe there are two well distinguishable clusters, obtained when hearing songs inducing {joy and anger} for the right cluster and {sadness and bliss} for the left cluster.

So, there exists some innate knowledge about the emotion dimensions, as no-one learns how to feel when listening to music and what would be the heart rhythm at that moment. The hypothesis that in the cognitive space "general" emotional categories exist is supported by the results, as the obtained physiological clusters correspond to quadrants of the plane on figure 4. The set of stimuli and the reactions suggest that humans distinguish such general categories.

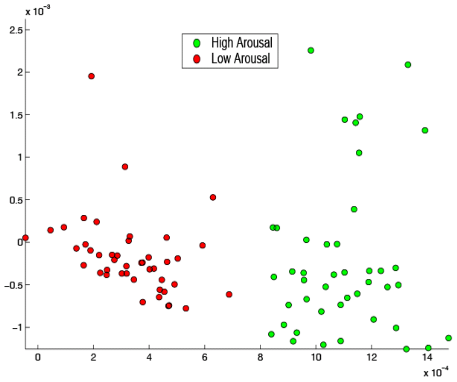

Fig. 6 Physiological clusters on the Arousal axe (Kim et all 2004)

These "general" categories do not obligatory have names. It is known in cognitive science that humans divide perceptual continuums intervals and then give names to the intervals. One example is the perception of colours and their names. The continuum of light frequencies is perceived in the same way by human being. But different cultures divide this continuum into intervals in different manner (and gave them names as "red" or "blue"). There are cultures in which the named-intervals for what we call "white" are nine and cultures which have only two names of colours for the entire spectrum.

The hypothesis that humans perceive features in emotive speech that allow them to categorize to more general categories seems reliable. These categories do not necessarily have names in the language(s). But when presenting to someone a sample of positive active speech and the labels {angry, sad, happy, fear and neutral}, she will certainly decide that it is "happy".

The problem is how to shape the feature space of multilingual classifiers of emotions.

The most convenient for machines are taxonomies and tree structures. Imagine the plane arousal-valence is covered with specific emotions, as it shown in figure 7, for example with the labels E1 to E8 (This precise positioning is purely geometrical; the labels are just covering the quadrants and the neutral positions). Assume the position of these labels corresponds to precise emotions like Anger, Happiness etc. By the way, the names of those places can be determined.

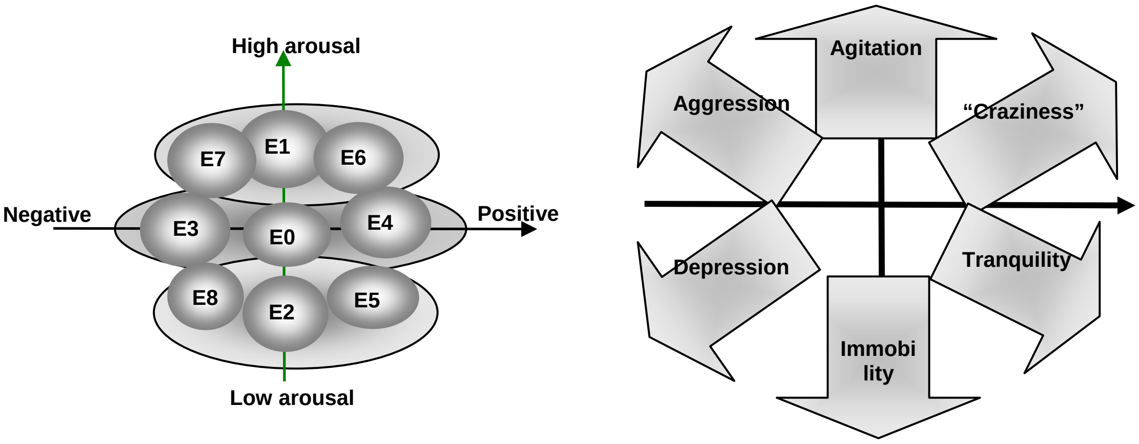

Figure 7 (a). Lower level of the taxonomical trees; (b) Tendencies of behaviour in the emotions' space.

Suppose these areas are leafs of a taxonomial tree. The upper level of the tree corresponds to general categories. As shown in figure 8, the taxonomic structure of general categories and more concrete emotions could be in two ways: 1. division to general categories depending on the arousal and to more concrete states following the valence – positive, neutral or negative (figure 8(a)); 2. division to general categories according to the valence and to concrete states following the arousal - positive, neutral or negative (figure 8(b)). As it is shown in figure 7(b), the 'general' level of classification could be useful for determining the tendency of the subject's behaviour.

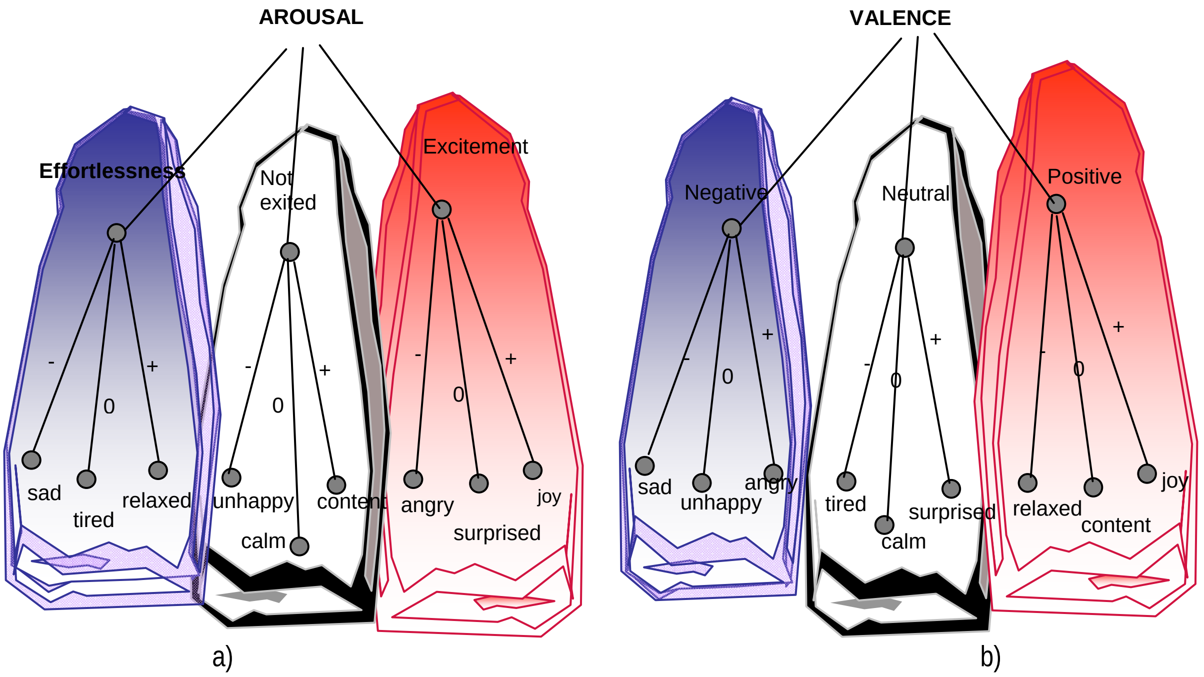

Figure 8. Taxonomies of general emotional categories and less general emotions.

Suppose that the set of language independent features, which leads to the classification to the general categories, is known. The proposal is to use the same strategy as humans seem to do. That leads to the following "algorithm":

- Take into account only language independent features.
- Classify to which general category belongs the speech signal.
- If we have information on the language, use additional information and classify to a leaf.

This strategy demands a kind of double classification – first to the general category and after that - to a leaf. But it avoids big mistakes. Such a classifier wouldn't need more and more data to be trained.

From a general point of view, the capacities of a machine-learned classifier are never as perfect as human capacities for recognition and categorisation. It is obvious that the use of additional channels of information for the machine recognition, such as visual (expression) features, will be very helpful.

Conclusion

In this paper it was assumed that language dependence is an important factor explaining why machine learned classifiers in (Shami and Verhelst, 2007) did not generalize from one database to the next. It should be noted, however, that the explanation more likely lies with the different ways in which the emotions are expressed in the different databases in general. Besides language and cultural differences, such differences could also have several other causes like the social setting in which the emotion occurs, the emotion as a permanent state of mind or induced by a specific exceptional event, etc. In (Shami and Verhelst, 2007), no generalization was found when the classifier was trained on the Danish database and tested on the German database or vice versa. However, in the Danish database, the emotions with a same label are usually more subtly expressed and more varied than in the German database, whose samples often sound over-acted, and it is not at all clear that this is language related. Further, there was only very little generalization between the two English databases even though, besides the English language, both databases shared a motherese style of expressiveness. Therefore, it is not proven that "existing approaches for supervised machine learning lead to database dependent classifiers which can not be applied for multi-language speech emotion recognition ... because they discriminate the emotion classes following the used training language".

The field of speech emotion recognition has achieved several promising results. However, the data-driven approaches lead to machine learned classifiers that are database dependent. The problem can be solved by means of merging emotion-speech corpora and training with more and more data.

Experimental results for human emotion recognition showed that the underlying cognitive mechanisms allow language independent categorisation although the information about the used language is deeply involved in the speech signal. The analysis suggested also that the cognitive process uses some internal structure of the emotional categories, existing in the cognitive space.

In this paper, we elaborated a general strategy for developing language independent emotion recognition, which does not need large amount of training samples in all languages. The proposed approach provides a basis for a future research and experimental work. The study should first consider the identification of language independent speech features and culture independent information from parallel modalities such as visual (expression) features. In a second step we would analyse several classifiers, by considering the general categories of emotion. Parallel to that we will investigate the relationship/dependencies between the emotion categories and language(s) for the classification of leafs (if necessary). A comparison with state-of-the art of automatic emotion classifiers will be made.

Bibliography

Breazeal, C., Aryananda L., 2002. Recognition of Affective Communicative Intent in Robot-Directed Speech. In: Autonomous Robots, vol. 12, pp. 83-104.

Bulut M., Busso C., Yildirim S., Kazemzadeh A., Lee, C.M., Lee S., and Narayanan S., (2005) Investigating the role of phoneme-level modifications in emotional speech resynthesis. In Proc. of EUROSPEECH, Interspeech, Lisbon, Portugal, 2005

Cichosz, J., Slot, K., 2005. Low-dimensional feature space derivation for emotion recognition. In: Interspeech 2005, p.p.477-480, Lisbon, Portugal.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M., 2000 'FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time, ISCA Workshop on Speech & Emotion, Northern Ireland 2000, p. 19-26

Devillers L., Vidrascu L., Lamel L., (2005) Challenges in real-life emotion annotation and machine learning based detection, Neural Networks, Volume 18 , Issue 4, Elsevier Science Ltd. Oxford, UK, UK, 407 - 422

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. Speech Communication, Speech Communication, Elsevier Science Publishers B. V. Amsterdam, The Netherlands, 40, 33-66

Dusan, S., Rabiner, L., 2005. On Integrating Insights from Human Speech Perception into Automatic Speech Recognition. In: Interspeech 2005, Lisbon, Portugal.

Ekman P., “Are there basic emotions?” Psychological Review 99 (3), pp.550–553, 1992.

Engberg, I. S., Hansen, A. V., 1996. Documentation of the Danish Emotional Speech Database (DES). Internal AAU report, Center for Person Kommunikation, Denmark..

Fernald, A., 1992. Human maternal vocalizations to infants as biologically relevant signals: an evolutionary perspective. In: Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, Oxford.

Fernandez, R., Picard, R. W., 2005. Classical and Novel Discriminant Features for Affect Recognition from Speech. In: Interspeech 2005, p.p. 473-476, Lisbon, Portugal.

HUMAINE 2004b, « Theories and Models of Emotion », June 17-19, 2004, Work Group 3 – Synthesis, online available on http://emotion-research.net/

HUMAINE, 2004a, “Emotions and speech - Techniques, models and results Facts, fiction and opinions”, Synteses of HUMAINE Workshop on Signals and signs (WP4), pr. by Noam Amir, Santorini, September 2004, online available on http://emotion-research.net/

Izard C., “Four systems for emotion activation: cognitive and noncognitive processes,” Psychological Review 100, pp.68–90, 1993

Katz, G., Cohn, J., Moore, C., 1996. A combination of vocal F0 dynamic and summary features discriminates between pragmatic categories of infant-directed speech. In: Child Development, vol. 67, pp. 205-217.

Kim K. H., Bang S.W. and Kim S. R. (2004), Emotion recognition system using short-term monitoring of physiological signals, in: Journal of Medical and Biological Engineering and Computing, Springer Berlin / Heidelberg. Volume 42, Number 3 / May, 2004

Kim, Jonghwa, 2004, Sensing Physiological Information, Applied Computer Science, University of Augsburg, Workshop Santorini, HUMAINE WP4/SG3, 2004, online available on http://emotion-research.net/

Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S., ”Emotion recognition based on phoneme classes”, ICSLP, 2004.

Lee, S., Bresch E, Adams J., Kazemzadeh A., and Narayanan S., 2006 (a). A study of emotional speech articulation using a fast magnetic resonance imaging technique. In Proceedings of InterSpeech ICSLP, Pittsburgh, PA, Sept. 2006.

Lee, S., Bresch E, and Narayanan S., 2006 (b). An exploratory study of emotional speech production using functional data analysis techniques. In Proceedings of 7th International Seminar On Speech Production, Ubatuba, Brazil, pp. 525-532. December 2006.

Picard, R., 1997. “Affective Computing”, MIT Press, Cambridge. 1997

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. Annual Review of Language Acquisition, 2, 85-115, 2002.

Rotaru, M., Litman, D., 2005. Using word-level pitch features to better predict student emotions during spoken tutoring dialogues. In: Interspeech 2005.

Scherer, K. R. “A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology,” in : Proc. ICSLP 2000, Beijing, China, Oct 2000.

Scherer, Klaus R., Rainer Banse, and Harald G. Wallbott. 2001. "Emotion Inferences from Vocal Expression Correlate across Languages and Cultures," Journal of Cross-Cultural Psychology 32/1: 76-92.

Shami M., Kamel, M., 2005. Segment-based Approach to the Recognition of Emotions in Speech. In: IEEE Conference on Multimedia and Expo (ICME05), Amsterdam, The Netherlands.

Shami, M., Verhelst, W., 2006. Automatic Classification of Emotions in Speech Using Multi-Corpora Approaches. In: Proc. of the second annual IEEE BENELUX/DSP Valley Signal Processing Symposium (SPS-DARTS 2006), Antwerp, Belgium.

Shami M., Verhelst W., 2007; An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech, to appear in: Elsevier Editorial System(tm) for Speech Communication, 2007

Shriberg, E., 2005. Spontaneous speech: How people really talk and why engineers should care. In: Eurospeech 2005, Lisbon, Portugal.

Ten Bosch, L., 2003. Emotions, speech and the ASR framework. In: Speech Communication, vol. 40, no. 1–2,.213–225.

Wan V., Ordelman R., Moore J., Muller R., (2005) “AMI Deliverable Report, Describing emotions in meetings”, internal project report, On line available http://www.amiproject.org/

Authors' Information

***Velina Slavova** - New Bulgarian University, Dept. of Computer Science; Bulgaria. e-mail: vslavova@nbu.bg*

***Werner Verhelst** – Belgium, Vrije Universiteit Brussel, Department of Electronics & Informatics; e-mail: wverhels@etro.vub.ac.be*

***Hichem Sahli** – Belgium, Vrije Unversiteit Brussel, Department of Electronics & Informatics; e-mail: hsahli@etro.vub.ac.be*