Exchanging Emotions — SOM Approach

Heikki Hyötyniemi and **Juha Hemánus** and **Vuokko Lantz**

Abstract. Today's cellular phones are still based on the age-old views of what a telephone can do, that is transmit speech. However, in human interaction, communication is not based just on the actual words but also on the way how the words are spoken. How to better take this into account in the design of future phones? How to make the phones less technical, more personal? In this paper, we try to answer such questions by discussing the possibilities of applying *affective computing* in cellular phones and presenting experiments that have been carried out with Self-Organizing Map (SOM) algorithm and visualization of emotional aspect of speech.

1 INTRODUCTION

It has been claimed that only a fraction of human interaction takes place on linguistic level — in addition to that, there are unintentional and subconscious hints in speech and gestures revealing your emotions, motives, etc. If the yesterday's technology only facilitated the transfer of speech information as a sound, why should the future devices be limited in this way? If the non-verbal content in speech was exploited more explicitly, the future phones could support human communication in a much richer way than the phones of today.

When the crucial communication needs are fulfilled, one could perhaps look at the phone more like a playful thing or social instrument rather than as a tool. Indeed, we are going from effective to *affective computing.*

Affective Computing has gained widespread interest in the late 1990's. This is due to a great need of developing information systems that are increasingly able to adapt themselves to the user's moods in order to provide user-friendlier and easier interaction styles to a wide range of users, independent of their background in system usage. If the system knew more about the user and his/her mental state, the action required for the desired goal could be differentiated much faster (with varied prompts or modes: Fast short-cuts for the experienced, more thorough basic info supply for the first-time users), and thus the usability could be increased remarkably. This aprroach could, in the first phase, be applied for exmple to basic information searches such as library services, transport schedules and show info and ticket reservations.

On the basis of prior work of Scherer [9],[10], Murray and Arnott [6], Pereira and Watson [7], Dellaert, Polzin, and Waibel [2], and Picard [8]) we already know that the basic dimensions of emotion are arousal (or intensity) and valence. Sometimes a third dimension — quality — is added. Usually, the basic six universal human emotions are listed as follows: *Happiness, sadness, fear, anger, disgust,* and *surprise.* It has been assumed that of these six basic emotions, sadness and anger are best recognized, followed by fear and joy. Arousal dimension of emotion seems to be best communicated by the pitch and loudness of the speech whereas the valence (positive or negative attitude) is said to be composed of subtler and more complex patterns of inflection and rhythm. Murray and Arnott [6] discern three basic types of voice parameters affected by emotion: Voice quality, utterance timing and utterance pitch contour.

It needs to be remembered that detecting emotion in voice is a huge task. Scherer mentions [9] that in studies of lay judges' ability to recognize emotions from purely vocal stimuli, an accuracy of about 60 % is found (if the listener judgments were based purely on guessing, that is on a random chance, the rate would be 12 %). Picard mentions [8] a study conducted at the Massachusetts Institute of Technology Media Lab in 1996 where the researchers developed a method for a computer to classify sentences as approving or disapproving. The resulting recognition accuracy was 65–88 % for speaker-dependent, text-independent classification of approving versus disapproving sentences. It is noteworthy that the same sentences were also classified by people and with a similar classification accuracy. The classification accuracy differed from speaker to speaker: The computational model successfully recognized the valence differences between three speakers and this pattern of success was observed also in the case of human judges.

If a computer is trying to adapt its behavior to better suit its current user, then an ability to sense user's approval or disapproval would aid this process remarkably. But, as talking of such a subtle and private human dimension as expressing emotions and having them recognized, it seems that a recognition accuracy of 65–88 % is far from satisfactory. Instead of increased usability, an adaptive system making wrong decisions about the user's emotions and thus wrong adjustments to itself is plainly annoying. How could we improve the recognition accuracies – or could such vague intepretations of emotions and attitudes still be usefull in some applications?

2 MODELING EMOTIONS

Indeed, when analyzing speech, we are facing interesting challenges and dilemmas. For example in speech recognition, emotional variations are regarded as noise that should be filtered out from the signal in order to increase the recognition rate for the sopken words. On the other hand, in affective computing and in the analysis of emotional aspects of speech, these variations are of the main interest.

In the experiments reported here, the six above mentioned emotions were selected as a starting point. Some of the emotions were renamed in order to have a research problem better suited to discerning and solving problematic situations in human-computer interaction applications in mobile communication terminal. The old categorization of emotions has emerged from more traditional psychological and socio-prychological (human-human interaction) problems setting. So, sadness was re-defined as *anguish*, anger as *aggression* and disgust as *frustration.* All this to get more situational, contextual and acute expressive material from the test subjects – they could imagine themselves having problems in the interaction with the system. Happiness, fear and surprise remained as they are. In addition to these, *neutral* speaking voice was included for providing a calibration, or a baseline, for the other emotions.

For the first experiment, altogether 32 people were chosen to the test group. Most of them were at least a little used to prompted vocal

[1] Nokia Research Center, P.O. Box 407, FIN-00045 Nokia Group, Finland. Email: heikki.hyotyniemi@hut.fi

expression due to experiences in amateur theatre. Of these 16 were male and 16 female. The ages of the subjetcs ranged from 17 to 67.

As research material, 13 sentences were chosen from a well-known speech database called TIMIT (collected by Texas Instruments and Massachusetts Institute of Technology in the late 1980's). The expressive and exclamatory strength of the sentences, as opposed to their semantic meaning, was used as a choice criterion. In fact, sentences were selected so that they would be as open as possible to any of the basic emotions, not leaning to any specific interpretative direction.

The analyzed material consisted of 32 (number of test subjects) times 7 (number of emotions) times 13 (number of sentences), altogether 2912, utterances. The data was recorded digitally (sampling frequency was 32 kHz, high frequencies were filtered away later on). The utterances were divided into 1/64 second long periods, so that a typical sentence consisted of a set of some hundred such a sequences. Feature analyses were carried out separately for all these sequences in a Matlab environment.

The sequences were divided into classes representing silent, transitory, and voiced periods. The class of (inter-sentence) silent periods were used only for estimating the fragmentariness of the speech, whereas the voice power characteristics were analysed from the samples belonging into the two other classes. The voiced samples were analyzed for finding the spectral properties, specially the pitch frequency, and its time-dependent patterns. No syntactical or semantical analysis were carried out.

To begin with, altogether 30 features were extracted. Most of the features were related to the spectrum but also dynamic AR parameters and cepstrum parameters were experimented with. The features were collected into real-valued data vectors. The speech samples assumedly containing some emotion were used to define the *prototypes* for the emotions, that is, the corresponding feature vectors determined an *emotion cluster* in the feature space. Three first sentences were reserved for testing purposes and they were not used for model construction.

For managing the high-dimensional feature space, efficient mathematical tools were needed. It turned out that the *Fisher Discriminant Analysis*, FDA, (originally presented in [3]) was a practical choice. It is computationally light weight being based on linear theory but it is still capable of producing easily analyzable and intuitively understandable results.

In its simplest form Fisher Discriminant Analysis constructs an *axis* in the feature space, along which the between-emotions feature variance is maximized while the within-emotion variance is minimized. When speech samples are projected onto this axis, the emotions are optimally distinguishable from each other in terms of linear theory. When a model for distinguishing between two emotions is constructed applying FDA, the most characteristic samples of the other emotion are projected in the other end (discriminant function value, or the projection length, being high and positive), whereas the samples representing the other emotion are projected in the different end (discriminant function being high but negative) of the axis. Therefore, the *loadings* of the projection axis reveal what is the relevance of each feature in the comparison of two different emotions: High weight (positive or negative) in the vector means that the corresponding feature is specially significant when distinguishing between the emotions, values near zero mean that the feature is irrelevant.

A graphical user interface was constructed for the visualization of the analysis results (see Fig. 1). In the figure, data samples, that is, the 1/64-second periods of the spoken sentence, are represented by black dots. There are six plots showing how the samples within one

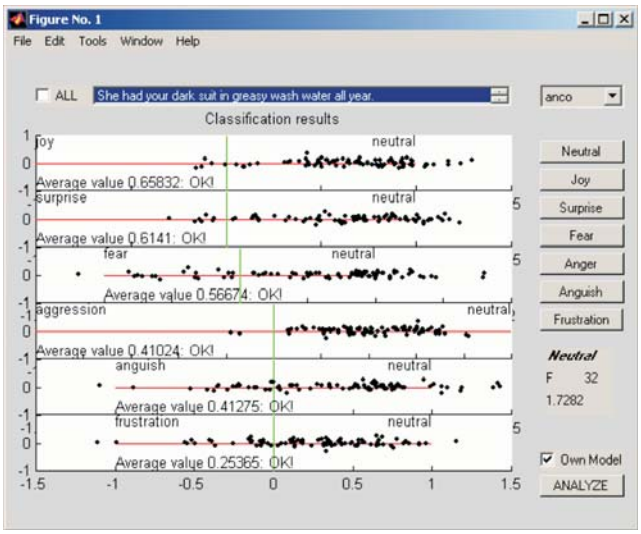

Figure 1. Graphical emotion analysis environment

sentence are located when projected onto different pair-wise emotion axes; in the figure, *non-neutrality* is being detected from the samples. The more there are samples on the left-hand side (to the left of the green line) of a given axis, the more certainly the sentence can be classified to the corresponding emotion.

It turned out — in accordance with prior studies, for an example see [8] — that the pitch frequency and its time variation characteristics, amplitude ratios between the first harmonics, signal power variation characteristics, and fragmentarity of the speech are the most important features for capturing emotional cues in speech. The sheer loudness of speaking voice was not regarded as relevant feature due to the difficulties of measuring it independently from the recording conditions. For example, the distance between the speaker and the microphone could not be controlled accurately in practice.

When applying the simplest form of FDA, separate axes are needed for each pair-wise analysis between two emotion classes; the final multi-class selection among emotions has to be based on some appropriate *voting scheme.* If all pair-wise comparisons return discriminant function values that are positive, then it is easy to make conclusions, the classification being unanimous. In more complex cases some cleverer strategy needs to be employed: For example, a linear multivariate regression model can be constructed for mapping the voting results into final classifications results.

There were some especially good actors and actresses among the test subjects and the emotion analysis results were very clear in their cases. However, usually unanimous decisions on the emotions of the spoken sentences could not be made. Unfortunately, it seems that some of the test speakers were less adept at acting and expressing prompt emotions verbally. Was this their failing or is our emotion classification technique inappropriate? This question is elaborated in the next section.

Whereas the analyses that were based on personal emotion models were rather succesful, attempts to use some generic, multi-speaker emotion models failed. As explained in [8], the way how people manifest their feelings is a very personal matter. It is also a question of culture and language:

- Some people just express their feelings differently: For example, while the pitch typically goes up when agitated (so called hot

anger), for some people it goes *down* (cold anger).
- In some cultures, anger is not expressed openly in general: Such feelings are reflected in irony and in changing vocabulary instead.
- In different languages, the nature of speech and thus the feature contents may be so different that analysis that is based on simple features may give very biased results.

Therefore, it seems evident that there is no universal emotion model; the analyses need to be carried out personwise. More powerful classification schemes (for example, the Support Vector Machine SVM [1]) would not solve the problem that is clearly beyond the data and feature-space division level. However, to be truly useful component of a user interface, the emotion analysis system cannot be too user-dependent: The user cannot be asked to supply the system with enough samples of all his moods to train it from a scratch: First, this would decrease the initial usability of the system significantly; secondly, it would take quite much time and talking to have a representative dataset of all the user's emotions.

Thinking positively — the personal nature of the emotion models should be seen as a strength rather than as a problem. Personal models do not just classify emotions but they also tell something about how individual people express their emotions. Evidently, unsupervised approaches are needed for the analysis of emotions. Similarly, different kinds of applications, not just hard classification, need to be created for emotion analysis.

3 EMOTION MAP

If it is taken as a starting point that *emotion analysis cannot be truly reliable and general*, what is left, is there any reason to continue research? The claim here is that *yes, indeed.*

After the basic needs have been covered — in the case of cellular phones, the need of communication and exchanging information — there are *new uses* for the devices. It has been claimed that games and other forms of entertainment are the new rising trend; but there are also other views to this leisure society.

As pointed out, no crucial basic functionalities should be based on emotion analysis. On the other hand, different kinds of fun and leisure applications are very difficult to foresee. To entrap possible users, something new and special needs to be offered, something that makes the phone less ordinary and different from the other phones. For example, the personal phone could be made *more* personal, more intimate — and affective computing is just what can be applied here. What is more, such approaches can also offer tools for inter-intimacy, closer connection between good friends, family members, and people in love.

Affective features in a phone are — as assumed above — only additional gimmicks on top of conventional phone functionalities. If these new features offer no real, practical benefits, they must not be a burden to the user. The new features should be self-sustained. In the case of a phone, the affective applications should be listening to the user's voice and adapting themselves in the background. This can be achived with self-organizing, unsupervised methods.

What is more, new interaction and communication channels could be employed. New modalities could be incorporated, so that in addition to the voice, other senses would also be activated, like vision. However, in a cellular phone there are constraints what comes to computing capacity, resolution, and the size of the display. A *Self-Organizing Map* seems to be a very promising environment to implement the above ideas — the algorithm is of low computational complexity and enables inter-personalized affective computation in a visually fancy setting in an unsupervised fashion.

The Self-Organizing Map (SOM) [5] is an artificial neural network structure, in which complex data can be made visually better understandable. A SOM projects data from a high-dimensional feature space onto a (typically) two-dimensional manifold so that the local topological feutures of the data set are preserved. When this manifold is unfolded for visualization, the projection is like a *map*. Nearby points in the high-dimensional space tend to be near each other also on the final low-dimensional map. There are various applications of this method, for example in the analysis of industrial data. Perhaps the most ambitious application is the modeling of textual documents [4]. These applications have shown that such a map form really helps to understand high-dimensional data sets and complex systems.

SOM is an intuitive and somewhat heuristic method. Results obtained with it are not unique as the outcome of the algorithm depends on the random initialization and the random presentation order of the data in the learning phase. Also, its convergence has not been proven generally. The most important and interesting property of SOM is that it converts high-dimensional data into a visually better understandable form — in a way it utilizes human pattern recognition capability for understanding large data sets by arranging data items into more general groups on the basis of their similarity.

Without going into algorithmic details (see [5]), only following needs to be recognized here: when a high-dimensional feature vector $f(k)$ is input in the converged SOM at time k, the outcome is its horizontal and vertical on the map, that is $x(k)$ and $y(k)$. Whereas the elements in f can be continuous-valued, x and y have discrete integer values as they refer to a map unit most similar to the input vector.

The intuitivity of SOM can be utilized for making emotions represented by high-dimensional feature vectors better manageable: different ways of speaking are mapped onto different regions of the map. During on-line operation, the visualization of speaker's emotions can be made dynamic with a traversing *speaking voice fingerprint*, that is the feature vector of the latest speech sample projected on the map. Different regions on the map can be assigned to different emotions using labelled speech samples. With such a map, changes in speaker's mood are reflected on the movements of the speaking voice fingerprint . The annotation of speech samples this is the only task where user interaction is needed.

From here on, it can be assumed that the phone automatically adapts to its owner's emotions and ways of speaking. There are no predefined, correct emotions. Instead, the system learns the user's personal moods and modes. This is accomplished by modelling the user's voice characteristics, pitch frequencies, etc. during the normal phone-using situations; only relevant, person-oriented emotions will be present in the model. These emotions will then be, for example, something like "formal/neutral" (when speaking with work mates), "casual" (when speaking with friends), or "lisping" (when speaking with kids). On the other hand, for some people "tired" can be one characteristic way of speaking.

This emotional SOM idea was tested in natural situations: Four test persons imitated their own phone use, imagining discussions with their friends, relatives, bosses, and vacuum cleaner salesmen. The Self-Organizing Maps were constructed for each speaker independently. The map dimensions were in all cases 10×10, that is, $1 \leq x \leq 10$ and $1 \leq y \leq 10$. In Figs. 2 and 3, two example maps (note — the best ones) are shown with appropriate labeling; three sentences are shown in both cases. The map color can be used as a third dimension when visualizing the map; now the standard visualization method called *U matrix* is employed. Because of the high

noise level, the feature vectors were calculated from only 1/64 second long speech samples, the map coordinates of the moving voice fingerprint needed to be low pass filtered. The final points that are shown on the map were calculated as an exponentially weighted sum of map locations of the current and previous feature vectors. The location of the moving speaking voice fingerprint can be calculated recursively as follows:

$$\begin{aligned} x_{\text{out}}(k) &= \lambda \cdot x_{\text{out}}(k-1) + (1-\lambda) \cdot x(k) \\ y_{\text{out}}(k) &= \lambda \cdot y_{\text{out}}(k-1) + (1-\lambda) \cdot y(k), \end{aligned} \tag{1}$$

where the forgetting factor λ was 0.9.

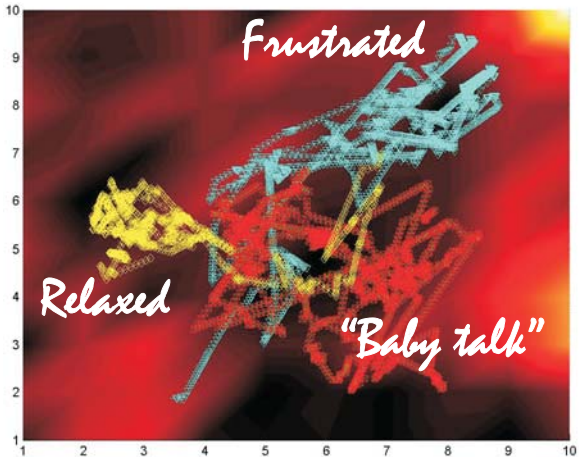

Figure 2. The mood map for one of the test persons

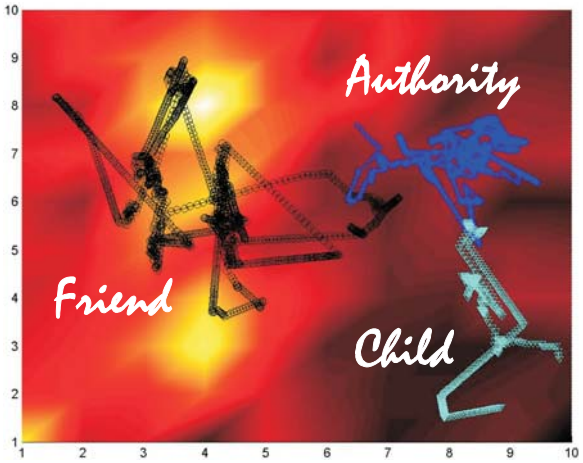

Figure 3. The mood map of another test person

It should be noted that if the SOM creation is carried out from scratch, the results will not always be the same. The resulting maps can just as well be mirror images of each other. However, if an existing map is being refined with new data, no abrupt changes in the map sructure will take place.

According to the results, it seems that different ways of speaking were distributed rather nicely on the map. However, it needs to be recognized that results are by no means always so clear; sometimes the results seemed contradictory. After all, it is only some technical features that are utilized — for example, it seemed that sound periodicity, that is the ratio between silent and voiced samples, was the most distinguishing feature among samples. And changes in this feature can be a result of many different things (boredom, relaxed mood, etc). So, seeing interesting and meaningful structures in a SOM is a bit like with horoscopes: People just seem to believe in them even though they know deep down that horoscopes are just for fun and should not be taken too seriously. After all, when interpreting emotions, there is nothing absolutely correct or incorrect. Indeed, when looking at the mood map, some people trust the computer more than their own feelings: "I did not recognize I was speaking angrily!". Indeed, such maps have not similar problems with objectivity as human — how many of us like to admit being jealous?

One potential field of application for such a mood map would be enhancing intimacy of phone conversation. As a proof of their sincere intentions, good friends could, for example, exchange each other's emotion maps: The listener can detect the movements of the speaker's "emotion pointer" on the map a after receiving his personal SOM-based emotion model. Revealing emotions — whether real or imagined — is like sharing secrets.

4 CONCLUSIONS

Experiments reveal that recognition of emotions is by no means easy. How emotions are expressed is not culture independent; neither is it language independent, or speaker independent. These things make automatic recognition of emotional aspects of speech a very challenging problem. It seems that nothing too crucial can be implemented based on automated emotion analysis. However, there can be some leisure and amusement related functionalities that work well even with less than perfect recognition results. If emotion functionalities are just additional gimmicks, not too much active concentration and effort can be assumed from the user to make the system work. The emotion analysis should be autonomous, with no or very little user interaction needed. It is not realistic, for example, that each user should explicitly train all of his or her personal ways of expressing emotions to the phone. This is the spirit of the *emotion SOM* proposed her.

Another point is that the nominal basic feelings may be irrelevant for some individual speaker. There may exist more characteristic ways of speaking, like "business mode", "leisure style", "tired", "cuddling" etc. Taking this into account is a key to truly personalized affective devices.

There are big promises, but also risks. If such emotion oriented features are implemented in the cellular phones, how will people react? Would you like the computer to analyze your feelings and would you trust on the results?[2]

Anyhow, it needs to be emphasized that this kind of applications are not really crucial and they should not be advertised as such. But, you never know what kind of innovations become *hot*, at least without trying them. And this kind of things — revealing such intimate personality twists like emotions — just might become a must among best friends. Visuality, intuitivity and non-technical outlook of the application are essential success factors. Perhaps one could speak of information age horoscopes, or cellular crystal balls!

[2] Note that if one really would like the phone to spy on you, and to read your mind, it would be easiest to add new sensors: Measuring the minor changes in the perspiration level, or measuring the humidity on the fingertips holding the phone, would make it possible to implement real lie detectors!

REFERENCES

[1] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines.* Cambridge University Press, 2000.

[2] F. Dellaert, T. Polzin, and A. Waibel, “Recognizing Emotion in Speech”. *Proceedings of ICSLP,* 1996.

[3] R.A. Fisher, “The use of multiple measurements in taxonomic problems”. *Annals of Eugenics*, **7**, pp. 179–188, 1936.

[4] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen, “WEBSOM — self-organizing maps of document collections.” *Neurocomputing,* **21**, 1998, p. 101–117.

[5] T. Kohonen, *Self-Organizing Maps.* Springer, Berlin, 1995.

[6] I.R. Murray and J.L. Arnott, “Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion”. *J. Acoust. Soc. Am.*, **93**(2), 1993, pp. 1097–1108.

[7] C. Pereira and C. Watson, “Some Acoustic Characteristics of Emotion”. *Proceedings of ICSLP,* 1998.

[8] R. Picard, *Affective Computing.* MIT Press, Cambridge/London, 1997.

[9] K.R. Scherer, “Expression of emotion in voice and music”. *J. Voice,* **9**(3), 1995, pp. 235–248.

[10] K.R. Scherer, “Adding the Affective Dimension: A New Look in Speech Analysis and Synthesis”. *Proceedings of ICSLP,* 1996.