MIT Media Lab Technical Report No. 585, January 2005. Appeared in:Intelligent user Interfaces (IUI 05), 2005, San Diego, California, USA.

Emotive Alert: HMM-Based Emotion Detection In Voicemail Messages

Zeynep Inanoglu
Speech Interfaces Group
MIT Media Lab
20 Ames st
Cambridge, MA 02139, USA
zeynepinan@post.harvard.edu

Ron Caneel
Human Dynamics Group
MIT Media Lab
20 Ames st
Cambridge, MA 02139, USA
rcaneel@media.mit.edu

ABSTRACT

Voicemail has become an integral part of our personal and professional communication. The number of messages that accumulate in our voice mailboxes necessitate new ways of prioritizing them. Currently, we are forced to actively listen to all messages in order to find out which ones are important and which ones can be attended to later on. In this paper, we describe Emotive Alert, a system that can detect some of the significant emotions in a new message and notify the account owner along various affective axes, including urgency, formality, valence (happy vs. sad) and arousal (calm vs. excited). We have used a purely acoustic, HMM-based approach for identifying the emotions, which allows application of this system to all messages independent of language.

Keywords

Affective computing, machine learning.

INTRODUCTION

The affective contents of a voicemail message is only available to us once we listen to a significant portion of the message. According to [6] , one method employed by voicemail users is to listen to the first few seconds of each message for qualities such as speaker identity and speaker's tone of voice to decide if a message needs immediate attention. In this paper, we look at the first ten seconds of voice mail messages to extract salient acoustic features. Based on these features, we have trained emotion models for eight emotional states: happy, sad, calm, excited, urgent, not urgent, formal, informal. We have integrated our classifier with PhoneShell, the voicemail service for the Speech Interfaces group in the Media Lab.

Previous work has been undertaken to assess urgency and business-relevance based on the automatic or manual transcription of contents of the message. [5] Our approach was to set aside content and try to classify the affective qualities of the message based only on acoustic features. The goal of the project was threefold: to converge on a set of acoustic features that discriminate along the four axes of valence, arousal, urgency and formality, to investigate the best statistical model for the task and to have an end-to-end emotive alert system that can be used by a group of voice mail owners on a regular basis.

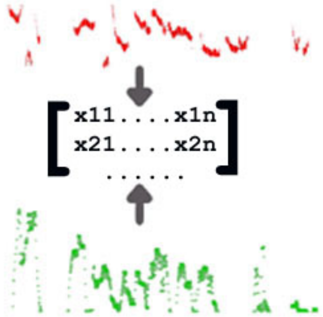
TRAINING AND LABELING

A total of 361 message segments were available for training. Three sources of training data were used:

- **PhoneShell voicemail messages:** We had access to 219 voicemail messages left for members of the Speech Interfaces Group in the Media Lab. The messages consisted of a good mix of personal and business content.
- **CallHome English Speech Corpus:** This corpus, available through the Linguistic Data Consortium, is a collection of natural telephone conversations, mostly of informal nature, between friends and family. We have extracted 90 segments with strong affective undertones to aid in our training.
- **Oasis Database:** We have used 52 segments of speech data from the Oasis Database which is derived from conversations recorded by British Telecom(BT) between operator service agents and customers in the United Kingdom. (see [2])

Two experimenters independently labeled each message along four axes. The first ten seconds of each message was used for labeling. A binary labeling scheme was employed. For each of the eight emotional states, a 0 or 1 was assigned to each message, indicating the presence or absence of the emotion. Labels were then compared and

Table 1: List of Features

	Pitch	**Loudness**	**Speaking Rate**
	• **25th percentile** • **75th percentile** • **f0 range** • **slope of the regression line through contour** • **total number of local optima**	• **mean loudness** • **25th and 75th percentile of perceived loudness** • **25th and 75th percentile of rms** • **mean spectral loudness in 4-14th barks**	• **duration of the voiced segment**

only those messages with labels that were in agreement were used as training data.

EMOTION MODELS AND ACOUSTIC FEATURES

The significance of prosody in conveying emotions has been illustrated by many studies including [1] . The amount of affective information in a speaker's prosody increases particularly when speech is the only channel used to convey the message (i.e. there is no visual information). In the long term, it would make sense to make use of both prosodic features and the actual contents of the message to infer the emotional state. However, in this paper, we're interested in finding out how far we can get by using prosodic features without automatic recognition and processing of the contents. We have experimented with a set of features based on pitch, speaking rate and perceived loudness. Each of these features were extracted from a single voiced segment. Subsequent to a segmentation scheme, which extracts all the voiced segments in the first ten seconds of speech, feature vectors are formed for each voiced segment possible. We use the sequence of feature vectors to train hidden markov models for each emotion. Table 1 summarizes the 23 features used in our final HMM-based system.

Perceived loudness was extracted using code provided by Raul Fernandez of MIT Media Lab and is based on Zwicker's model of perceived loudness, which accounts for filtering effects of the human auditory system. For detailed information on Zwicker's model, see [4], [3].

As a baseline, we have also extracted a global feature set to train a set of Gaussian Mixture Models with varying number of mixtures and covariance matrix types. To train the GMMs, a single feature vector is formed per message, instead of a sequence of feature vectors. The same features were used for this case, with the exception of absolute duration. Instead, the average duration of voiced segments was used as well as the total number of voiced segments averaged by message length (in case the message is shorter than 10 seconds). These two features are good indicators of speaking rate, which is somewhat lost when the features are extracted from voiced segments only.

CLASSIFICATION OF VOICEMAIL MESSAGES

When a new message is received, the Emotive Alert system executes the segmentation and feature extraction on the first 10 seconds of the message. The feature vector sequence is fed through the eight emotion models, resulting in log likelihoods for each emotion. The first step in the classification process is to choose the appropriate axes. It is quite common that the message may be neutral when assessed along one axis (neither happy, nor sad) but may clearly be in one extreme of another axis (a message that is neither happy nor sad may clearly be formal). Forcing decisions on insignificant axes would merely provide the voicemail user with useless information. In order to avoid this, we look at the difference between the log likelihoods of opposite emotions and only pick the two axes that have the highest difference. Once the axes are identified, the appropriate pole of that axis is reported to the user.

RESULTS

We have performed leave-one-out cross validation on all the labeled data. Table 2 provides a summary of the results given the different models we have used. The values indicate the percentage of correctly classified messages for all the data. The manual labeling by the two experimenters were taken as the benchmark with which all the results were compared. N represents the number of states or mixtures, while Diagonal or Full represents the type of covariance matrix used.

Our results clearly indicate that the baseline GMMs were not able to discriminate along the axis of valence and urgency. On the other hand, using three state HMMs with segmental prosodic feature sets resulted

Table 2: Results of Leave-One-Out Cross Validation. Percentage of correctly identified emotions in each training group, where system performance was checked against manual labels.

	Valence		Arousal		Urgency		Formality	
Model	Happy	Sad	Excited	Calm	Urgent	Not Urgent	Formal	Informal
GMM N=1 Diag	75	44	52	76	35	70	73	44
GMM N=1 Full	87	11	68	79	17	92	60	59
GMM N=2 Diag	65	37	67.5	75	37	63.5	49	68
GMM N=4 Diag	70	37	62	74	37	68.5	64	57
GMM N=2 Full	NA	NA	57	74	NA	NA	37	65
HMM N=3 Fully connected	79	73	67	67	66	76	68	64

in consistent discrimination of the data. In the case of formality, HMMs also improved the performance of the GMMs. The only case where most GMMs outperformed the HMM model was arousal but the difference was not significant. This may be due to the dominant role of speaking rate features in representing arousal which were only available in the global feature set. We are in the process of analyzing such feature-axis dependencies.

APPLICATION

We have integrated our HMM models with PhoneShell, the voicemail system of the speech interfaces group. New voicemail messages are processed by the segmentation and feature extraction schemes and classification takes place along all axes. The application is currently implemented in Matlab. Once the new message is classified, a multimedia message (MMS) is formed with a subject line consisting of the two emotions on the two most significant axes and an emoticon representing the dominant emotion. The message generation and dispatch is done using metasend from within a perl script. For instance, if the classifier identifies the axis of formality and urgency as the two most significant axes, only the results of classification along these axes are reported to the account owner. (i.e. "You have an urgent and informal message"). Currently we are able to send MMS messages to mobile phones supporting MMS and to email accounts.

DISCUSSION

The results of the HMM-based system indicate that Emotive Alert identifies emotions with better than chance performance for all four axes. During informal interviews, where users were asked to pick two of the eight available emotional states upon listening to a message, the axes of choice were very frequently identical to the axes selected by the system. It seems that whenever a human being is unsure of labeling a message along one axis (i.e. the message is neither happy nor sad), the system seems to successfully discard that axis and report more significant ones. We believe that in its current condition, the Emotive Alert system has fulfilled its goal of providing the voicemail user with information on the affective content of the message.

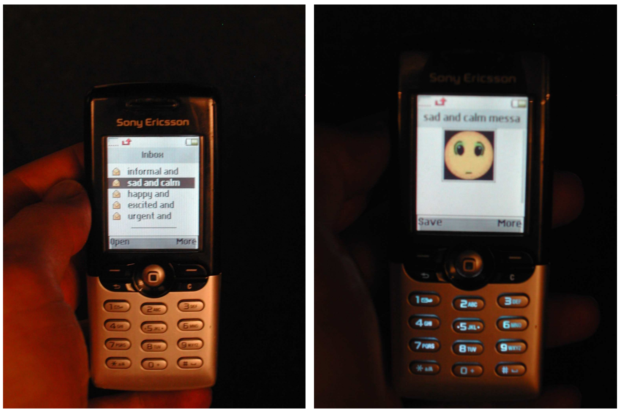


Figure 1: An example of an MMS message received by the voicemail owner.

REFERENCES

1. R. Cowie, D. Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human computer interaction. *IEEE, Signal Processing Magazine*, 2001.

2. P. J. Durston, M. Farell, D. Attwater, J. Allen, H.-K. J. Kuo, M. Afify, E. Fosler-Lussier, and L. C.-H. Oasis natural language call steering trial. In *Proceedings Eurospeech*, pages 1323–1326, Aalborg, Denmark., 2001.

3. R. Fernandez. *A Computational Model for the Automatic Recognition of Affect In Speech.* PhD thesis, MIT Media Lab, 2004.

4. H. Quast. Absolute perceived loudness of speech. *Joint Symposium on Neural Computation*, 2000.

5. M. Ringel and J. Hirschberg. Automated message prioritization: Making voicemail retrieval more efficient. *CHI*, 2002.

6. S. Whittaker, J. Hirschberg, and C. Nakatani. All talk and all action: Strategies for managing voicemail messages. *CHI*, 1998.