ISCA Archive
http://www.isca-speech.org/archive

ITRW on
Speech and Emotion
Newcastle, Northern Ireland, UK
September 5-7, 2000

EMOTION-SENSITIVE HUMAN-COMPUTER INTERFACES

Thomas S. Polzin and Alexander Waibel

Interactive Systems, Inc.
Pittsburgh, PA

Carnegie Mellon University
Pittsburgh, PA

ABSTRACT

People are polite to their computers. They are flattered by them, form teams with them and even interact emotionally with them. In their experiments, Reeves and Nass (The Media Equation, 1996) showed that humans impose their interpersonal behavioral patterns onto their computers. Thus, the design of human-computer interfaces should reflect this observation in order to facilitate an effective communication.

In order to build a human-computer interface that is sensitive to the user's expressed emotion, we investigated spectral, prosodic, and verbal cues in the user's utterance. Based on these cues, we showed that the classification system achieved accuracies comparable to human performance.

Finally, we demonstrate how to integrate information about the expressed emotion into a dialog system. The dialog system employs different discourse strategies depending on the expressed emotion allowing for a natural and effective communication between the user and the system.

1. INTRODUCTION

In several experiments Reeves and Nass [1] show that humans extrapolate their interpersonal interaction patterns onto their computers. Humans are polite to computers, they are flattered by them, form teams with them, and have emotional experiences when interacting with them. As a result, there are even good and bad computers.

If we indeed try to interact with computers in the same way as we interact with other humans, then the nature of this interaction should be reflected in the design of human-computer interfaces to facilitate a more “natural”, more human-like interaction. Studies on human computer interaction (HCI) recognize this similarity and also consider emotions to be an important factor in the communication between humans and machines.

However, most investigations in HCI focus on the computer’s synthesis of emotional expressions, both visual and acoustic [2]. We already have cartoon-like characters popping up if we do something wrong or ask for help. Depending on the situation these characters smile or frown at us. Soon these characters will speak to us, have facial expressions, and use gestures to make a point [3,4,5].

But if Reeves and Nass [1] are right about their thesis then this kind of interface design is dangerously one sided since this design ignores the emoting user. It might even be a little bit confusing when we have to interact with characters emoting heavily while our emotional state is - quite impolitely - ignored. We do need interfaces that not only express emotions but also detect emotions in the user.

The problem of an emoting computer that is unaware of the emotional state of the human user becomes even more evident when we allow the human user to speak to the computer using a speech recognition system in the “front end"” of the interface. When a speech recognition interface only pays attention to what is said but ignores how it is said, the interface fails to pick up information that is essential for an effective communication. For instance, certain word or syntactic choices might indicate that the speaker is angry or sad. Certain acoustic features might indicate that the speaker is bored or interested. It should be obvious that this kind of information is important for natural interaction with a computer and essential for successful communication. Therefore, the search for cues that allow for the detection of the speaker's expressed emotion in an utterance becomes an important topic of research.

Emotions can be communicated in various ways by relying both on verbal and non-verbal means. Non-verbal means comprise body gestures, facial expressions, the modifications of prosodic parameters, and changes in the spectral energy distribution.

This investigation is confined to information within the utterance. We show that verbal and non-verbal information within the signal allows an effective decoding of the expressed emotion.

Finally, decoding the expressed emotion within an utterance is, of course, only the first step in the realization of emotion-sensitive human-computer interfaces. When sensing a certain emotion in the person we communicate with, we adjust our linguistic strategies. We might avoid certain words, ask certain questions, talk about certain topics, and avoid others. Thus, the integration of information of the user's expressed emotion into a dialog system becomes an essential part in building effective human-computer interfaces.

The following section describes previous research regarding the detection of emotions expressed in speech. This section also specifies the modeling assumptions of our approach. The third section reports the results of our experiments. We describe briefly the underlying corpus that consisted of several thousands of sad, angry, or neutral speech segments from English movies. In our experiments, we explored, in particular, the role of word choice and prosodic information. Here, based on this kind of information our classification system achieved an accuracy comparable to the accuracy of human listeners performing the same task. The fourth section describes the integration of information about the user's expressed emotion into a dialog system. Depending on the expressed emotion the dialog system adopts its overall strategy. We conclude this investigation with a summary and point to possible extensions.

2. MODELING VERBAL AND NON-VERBAL INFORMATION

It is, of course, possible to explicitly verbalize ones emotional state, experience, or upbringing:

I'm angry!

I'm experienced!

I'm frustrated!

I'm sad.

However, as the above examples suggest, the explicit verbalization of these facts renders the dialog quite unnatural. It interrupts the natural flow of the conversation and slows down the overall communication. Other circumstances might prevent the explicit verbalization as well. For instance, the person might not be aware of his or her current emotional state or level of experience. Another reason might be a situation where the person is embarrassed about certain facts, such as his or her origin, level of experience, or age.

However, the inference of these background facts is crucial to adjust the computer's response behavior and to allow for an effective communication. This is the case for computer-mediated and human-to-computer communication.

Several non-verbal means such as certain facial expressions, body postures, or gestures are also used to communicate emotions. Our investigation explores only one possible way to express an emotion relying on non-verbal cues: the modification of prosodic parameter in someone's speech. Similarly, there are several ways to communicate an emotion by verbal means, such as, certain syntactic constructions [6] or discourse strategies [7]. For our purposes, this investigation is confined to explore word choice as one way to express emotions verbally.

2.1. Verbal Information

In this investigation we did not try to model processes at the discourse or semantic level. Instead, we modeled lexical processes that signal an emotional involvement of the speaker. We modeled emotion-specific word choice information by computing the probability of a certain word given the previous word and the speaker's expressed emotion. The idea behind computing these probabilities is that certain word combinations are more probable for the expression of certain emotions. Computing the probability of a certain word given a history of previous words is a technique widely used in speech recognition (language modeling) and was also applied to text categorization [8] or to infer discourse structure [9]. We used emotion-specific back-off language models to detect the expressed emotion in an utterance [10] and trained these language models on emotion-specific data. Function words were deleted prior to training and testing.

2.2. Non-Verbal Information (Prosody)

The role of prosodic information within the communication of emotions has been studied extensively in psychology and psycho-linguistics [11,12]. This kind of research focuses mainly on the question of how humans express emotions by modifying prosodic parameters of their speech.

Research in the automatic detection of the expressed emotion has been quite limited [13,14]. In this investigation, we explored the following prosodic features in more detail:

- The mean and the variance of the fundamental frequency within an utterance. We normalized the fundamental frequency with respect to the speaker's gender.
- Two features, which approximated jitter information (small perturbations in the contour or the fundamental frequency).
- The mean and the variance of the intensity within an utterance normalized with respect to the respective context.
- Two features which approximated tremor information (small perturbations in the intensity contour).

Prosodic features are multi-functional. They not only express emotions but also serve a variety of other functions as well, such as word and sentence stress or syntactic segmentation. More important, in particular, fundamental frequency and intensity vary considerably across speakers and have to be normalized properly.

We trained emotion-specific prosodic models using Gaussian mixtures based on the above features that were normalized with respect to the context and the speaker's gender.

2.3. Non-Verbal Information (Spectral)

The role of spectral information in the communication of emotions was demonstrated in an experiment by [15], in which they resynthesized only pitch and intensity information from the signals of emotional speech segments, thus removing basically all spectral information. While the emotions of the original segments were recognized by human listeners with an accuracy of 85%, this accuracy decreased dramatically to just 47% for the sentences in which the spectral structure was filtered out and only prosodic information was preserved. Thus, information other than prosodic information was also able to signal the emotions originally expressed in the sentences. Voice quality – for instance, its clarity or pleasantness – can carry information about the vocal emotional expression. If changes in the voice quality correlate with the expression of certain emotions, then particular distributions of spectral energy may indicate these very emotions.

In order to account for the properties of the auditory system, we modeled spectral information by means of cepstral coefficients. We used a 30 dimensional melscale filterbank and derived 16 cepstral coefficients. We added the first and second order derivative of these coefficients. In addition, we considered log power and its first and second derivative. Thus, we had a total of 51 features, which we reduced to 32 coefficients by linear discriminative analysis. For the following experiment, we used triphone acoustic models comprising a total of 9358 states each with 70 mixture components per state.

In order to model emotion-specific spectral information, we started with an existing recognition system and performed emotion-specific adaptations. In our investigation, we used an maximum likelihood liner regression adaptation technique that is frequently used to adapt a speech recognition system to novel speakers [16,17].

This approach adapts only the mean parameters of acoustic models to novel speech data by a set of linear transformations that are found using the maximum likelihood training algorithm. In the test phase, we computed the probabilities that a given utterance was generated by the emotion-specific acoustic models and took the highest probability to be indicative of the expressed emotion.

2.4. Training and Testing

We essentially used the same training and testing procedure for verbal and prosodic models. We trained models on emotion-specific subsets of the training corpus.

In order to test the models' detection accuracy of an expressed emotion in an utterance, we computed the probabilities that the respective emotion-specific models generated the current observations (verbal, prosodic, or spectral). We took the highest probability to be indicative of the underlying emotion.

We used f1-scores that combine precision and recall to report the results in the following experiments. For some emotion class **i**, we define *precision* as the ratio of the number of segments classified correctly as **i** and the number of segments in the corpus classified as **i**, regardless of whether they were classified correctly or not. We define *recall* as the ratio of the number of segments classified correctly as **i** and the total number of segments in the respective corpus belonging to class **i**. Intuitively, the closer the precision, the recall, and the corresponding f1-score are to 1, the more accurate the classification of the respective classification system.

3. EXPERIMENTS

3.1 The Corpus

We chose speech segments from English movies as our corpus. This choice was motivated by several factors. First, in order to accurately estimate statistical models, an extensive quantity of training data was needed. By using a film's close captions as a first approximation for segmentation, transcription, and tagging, we were able to collect a large supply of emotional speech samples in a relatively short amount of time. Second, previous experiments indicated that the emotions in acted speech could be consistently decoded by humans and automatic systems [12,13]. Third, although we did not pursue visual cues in this investigation, this corpus allows the integration of visual information with spectral, prosodic, and verbal information in future work. Note that this corpus differs considerably from corpora used in other studies [18,19,13,20] in which several speakers uttered the same sentence in different emotional variations.

The movies were down loaded from a Toshiba VCR (M-752) to a Pentium II personal computer equipped with a Crystal Audio System with a sampling rate of 16kHz and 16 bits. The close captions from the video stream were extracted with the Text Grabber VBI Line 21 Video Decoder (GP-500). For the segmentation, transcribing, and tagging we employed five students who were instructed in several training sessions.

Three major steps were involved in tagging these talk shows and movies:

- Transcribers were told to find segmentations, which coincided with sentence or utterance boundaries.
- The expressed emotion within a segment had to be constant. Initial or final noises or silences were to be excluded from the segment.
- The close captions were the starting point for the transcriptions. Missing noises (human and non-human) or words were added by the transcribers.
- Each segment was annotated with three tags. The first tag indicated the gender of the speaker and the second tag the amount of background noise. For the last tag, the emotion expressed by the speaker, we told the transcribers to be as specific as possible and to choose from a given set of emotion tags.

For the tagging process the transcribers relied on either CoolEdit 96 (Syntrillium) or Sound Forge 4.0 (Sonic Foundry) and Sennheiser headsets.

The distribution of emotions within this corpus is given in **Table 1** in which we considered all segments regardless of the amount of background noise or music.

Tag	Neutral	Angry	Sad	Afraid
#	2991	1586	1076	203
Tag	Disgusted	Ironic	Happy	Surprise
#	26	28	347	19

Table 1: Distribution of segments according to the expressed emotion

Since only neutral, angry, and sad segments occurred frequently enough to warrant a reliable estimation of model parameters, the subsequent experiments were confined to exploring differences among these categories. The corpus was divided into training and test sets. The test set consisted of segments randomly drawn from the movies. Note that within this test set the expressed emotions were distributed uniformly. Thus, the base line classification accuracy is 33% achieved by either random guessing or consistently classifying the same emotion. Since we used this corpus to train prosodic models only speech segments without any excessive background noise were considered in the training and test procedures.

3.2 Assessing Human Performance

We conducted two experiments involving human listeners. The motivation behind these experiments was to validate the quality of the tagging procedure and to interpret the accuracy of the automatic classification.

Four human subjects were asked to classify all speech segments from the test corpus according to the expressed emotion. Subjects were allowed to listen to the segments as many times as they wanted. The corresponding accuracies are given in **Table 2**. Angry segments were classified most accurately with an f1-score of 0.8, followed by neutral (0.7) and sad (0.6). Overall, about 70% of the segments were classified correctly.

	Sad	Angry	Neutral
F1-score	0.6	0.8	0.7

Table 2: F1-scores of human subjects classifying the segments in the test set.

In the following experiment, we asked five different subjects to classify all segments from the test corpus given only a textual representation of what was said. The corresponding accuracies are given in **Table 3**. Angry segments were classified best with an f1-value of 0.65, followed by neutral segments (0.54) and sad segments (0.43). Note that the order is the same as in the experiment above. However, overall only about 55% of the segments were classified correctly, which is 15% less than in the experiment above.

	Sad	Angry	Neutral
F1-score	0.43	0.65	0.54

Table 3: F1-scores for human subjects classifying speech segments based on a textual representation.

3.3. Classification Using Verbal and Non-Verbal Information

We trained emotion-specific prosodic models based on the eight prosodic features described above. Using these models to classify the segments in the test corpus, the system classified 60.4% of the segments correctly. The accuracies are given in **Table 4**. Angry segments were classified most accurately with an f1-score of 0.69, followed by neutral segments (0.61) and sad segments (0.54). Note that the order is identical to the order of the accuracies achieved by human listeners in the experiment above. However, the overall accuracy of human listeners was about 10% better than the overall accuracy of the classification system.

	Sad	Angry	Neutral
F1-score	0.43	0.65	0.54

Table 4: F1-scores for combined prosodic information.

We also explored emotion-specific spectral information using the adaptation technique as described above. The results of this experiment are given in **Table 5**.Overall, 63.9% of the segments were classified correctly. Angry segments were classified most accurately with an f1-score of 0.74. Sad and neutral segments achieved an accuracy of about 0.6. Emotion-specific adaptation of spectral information lead to a higher classification accuracy than classification based on prosodic information (63.9% and 60.4%). Remember that human subjects achieved an accuracy of about 70%.

	Sad	Angry	Neutral
F1-score	0.54	0,69	0.61

Table 5: F1-scores for spectral information.

In the final experiment, we trained emotion-specific back-off bigram models and used them to classify the segments in the test corpus. The corresponding accuracies are given in **Table 6**. Neutral segments were classified most accurately with an f1-score of 0.5, followed by angry (0.47), and sad (0.42).

	Sad	Angry	Neutral
F1-score	0.6	0.74	0.58

Table 6: F1-scores for verbal information.

Overall, 46.7% of the segments in the test set were classified correctly, which is about 10% less than the corresponding accuracy achieved by humans in the corresponding experiment above. Note that the accuracies of all three classification tasks (prosodic, spectral, and verbal) lay significantly above chance.

4. EMOTION-SPECIFIC DIALOG DESIGNS

Detecting the expressed emotion in a user's utterance is, of course, only the first step required to build emotion-sensitive human-computer interfaces. Once the expressed emotion is recognized, the system's behavior has to adapt to this information. A system able to decode the emotion expressed in an utterance but subsequently ignoring this insight would either be highly confusing to the user or it would be considered very rude. In the following we illustrate some areas in the design of a dialog manager that allow specifying emotion-specific responses and strategies.

Note that the above adjustment problem is not confined to interfaces being aware of the user's emotional state. Within human communication, insights into a person's idiosyncrasies are used to adjust the way we talk to that very person: we converse differently to a child than to an adult, differently to a woman than to a man, we change our discourse strategies when we realize that the person we are talking to is pressed for time or is experienced in the topic. All these adjustments allow for an effective communication and avoid misunderstandings.

Our dialog management module is implemented as a state transition network in which the user's utterance triggers the next state transition. Each state specifies how the user is prompted for information, what kind of information is fed back to user, and the way the user's utterance determines the state to transition to. Within our dialog management module adjustments to the user's expressed emotion take place in these three areas:

1. **Prompting:** Depending on the inferred emotion the system interacts with the user with different prompts. For instance, in case the user seems to be annoyed, the prompts sound more apologetic. When the user seems rushed, the prompts are succinct and to the point.
2. **Feedback:** Similar to the prompting, feedback to the user varies depending on the inferred emotion. For instance, in case of frustration, the feedback could be more explicit.
3. **Dialog flow:** Varying the prompts and feedback depending on the expressed emotion, adjusts only the concrete interaction with the user. Our dialog management module allows also the specification of different dialog flows to accommodate specific needs as imposed by the user's expressed emotion.

Note that these emotion-specific adjustments are specified explicitly during the design phase of the dialog management and vary depending on the concrete application.

Typically, we try to determine the emotion expressed by a user within his or her first utterances and react accordingly during the rest of the interaction. In other words, the emotion has scope over the remaining conversation. This is certainly a shortcoming of our approach, since it does not model any emotional changes in the user. We are currently evaluating an approach in which the user's emotional state is evaluated at each dialog state.

5. SUMMARY

This investigation explored whether language models and prosodic models based on eight prosodic features allow for an accurate detection of the expressed emotion in an utterance. Our experiments were based on a corpus of several thousands of sad, angry, and neutral speech segments from English movies. While the accuracies based on automatic classification fell short when compared to the accuracies achieved by human subjects, they were significantly above chance level (33%). For instance, the automatic classification based on prosodic models achieved an accuracy of about 60% while humans were able to classify about 70% of the segments correctly. Using spectral information, we were able to achieve an accuracy of about 64% correctly classified segments.

The results were similar for verbal cues. A classification system based on emotion-specific bigram probabilities achieved an accuracy of about 47%, well above chance level. However, the accuracy of human subjects was 55%.

We also argued that an adequate emotion-sensitive human-computer interface has to meet at least two requirements:

1. **Detection** of the user's expressed emotion.
2. **Adjustment** of system behavior with respect to the expressed emotion.

We described how our dialog management module met these requirements by the adjustments of its interaction with the user (prompting, feedback, dialog flow) once his or her emotion was recognized.

Future work entails investigating the combination of prosodic, spectral, and verbal information. Additional emotion categories require more sophisticated prosodic and verbal features, such as speaking rate, or stress distribution. Additional work has to be spent on the adjustment of the interaction with the user once his or her emotion is recognized.

Refererences

1. B. Reeves and C. Nass. *The Media Equation*. Cambridge University Press, 1996.

2. L. McCauley, B. Gholson, X. Hu, and A. Graesser. *Delivering smooth tutorial dialogue using a talking head*. In Proc. of wecc-98, Workshop on Embodied Conversational Characters, Tahoe City, California, 1998. AAAI, ACM/SIGCHI.

3. J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. *ANIMATED CONVERSATION: Rule-based Generation of Facial Expression, Gesture and Spoken Intonation for*

Multiple Conversational Agents. In Proc. of SIGGRAPH '94, 1994.

4. J. Cassell, M. Bickmore, M. Billinghurst, L. Cambbell, K. Chang, H. Vihjálmsson, and H. Yan. *An architecture for embodied conversational characters.* In Proc. of wecc-98, Workshop on Embodied Conversational Characters, Tahoe City, California, 1998. AAAI, ACM/SIGCHI.

5. N. Tosa and R. Nakatsu. *Life-like communication agent - emotion sensing character "mic" and feeling character "muse"*. In Proc. of Multimedia '96, Hiroshima, Japan, 1996.

6. A. Hübler *The Expressivity of Grammar.* Mouton de Gruyter, Berlin, New York, 1998.

7. R. Fiehler. *Kommuikation und Emotion. Theoretische und empirische Untersuchungen zur Rolle der Emotionen in der verbalen Interaktion*. Mouton de Gruyter, Berlin, 1990.

8. K. Seymore and R. Rosenfeld. *Using story topics for language model adaptation*. In Proc. of Europseech, 1997.

9. M. Finke, M. Lapara, A. Lavie, L. Levin, L. Mayfield Tomokiyo, T. Polzin, K. Ries, A. Waibel, and K. Zechner. *Clarity: Inferring discourse structure from speech*. In Proc. of the AAAI 98 Spring Symposium. 1998.

10. S. Katz. *Estimation of probabilities from sparse data for the language model component of a speech recognizer*. IEEE Transactions on Acoustics, Speech and Signal Processing, 35(4):400-401, 1987.

11. R. Frick. *Communicating emotion. The role of prosodic features*. Psychological Bulletin, 97(3):412-429, 1985.

12. K.R. Scherer. *Vocal affect expression: A review and a model for future research*. Psychological Bulletin, 99:143-165, 1986.

13. F. Dellaert, T.S. Polzin, and A. Waibel. *Recognizing emotions in speech*. In Proc. ICSLP, Philadelphia PA, USA, 1996.

14. N. Amir and S. Ron. *Towards an automatic classification of emotions in speech*. In Proc. of ICLSP, Sydney, 1998.

15. P. Lieberman and S.B. Michaels. *Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech*. Journal of the Acoustical Society of America, 34:922-927, 1962.

16. M.J.F. Gales. The generation and use of regression class trees for mllr adaption. Technical Report CUED/F-INFENG/TR 263, Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, England, August 1996.

17. C.J. Legetter and P.C. Woodland. *Speaker adaptation of HMMs using linear regression*. Technical Report CUED/F-INFENG/TR 181, Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, England, June 1994.

18. B. Tischer. *Die vokale Kommunikation von Gefűhlen*, volume~18 of Fortschritte der psychologischen Forschung. Psychologie Verlags Union, Weinheim, 1993.

19. G.S. Katz. *A Quantitative Study of Vocal Acoustics in Emotional Expression*. PhD thesis, University of Pittsburgh, 1997.

20. K.R. Scherer. *Randomized splicing: A note on a simple technique for masking speech content*. Journal of Experimental Research in Personality, 5, 1971.