Detecting Emotions in Speech

Thomas S. Polzin[1] and Alex H. Waibel[1,2]

[1] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2] Fakultät für Informatik, University of Karlsruhe, Germany

Abstract. Human language carries various kinds of information. In human computer interaction the detection of the emotional state of a speaker as reflected in his or her utterances is crucial. In this investigation we will explore how acoustic and prosodic information can be used to detect the emotional state of a speaker. We will show how prosodic information can be combined and integrated with acoustic information within a hidden Markov model architecture, which allows one to make observations at a rate appropriate for the phenomena to be modeled. Using this architecture, we will demonstrate that prosodic information adds discriminative power to the overall system.

Keywords: multimodal communication, user modelling, emotion detection

1 Introduction

Human language carries various kinds of information. Merely considering strings of words without regard to the manner in which they are spoken, one might miss important aspects of an utterance. Additional information which can carry cues to the underlying emotional state of the speaker is encoded on the acoustic level (Scherer et al. (1984) and Scherer and Banse (1996)). Within human computer interaction, a system's ability to sense the emotional state of a speaker within a dialogue is crucial for several reasons. Depending on the emotional state of the speaker, any of the following might be true:

- the sentence uttered might have a meaning different from the meaning as derived by only considering the words uttered,
- the system might allocate special resources to process the input, or
- the system's output behavior might adapt.

While there is some research which attempts to detect emotions given visual cues (Kaiser and Scherer (to appear), Kaiser and Wehrle (1992), and Takeuchi and Nagao (1992)), the usage of acoustic cues for the detection of emotions in speech has not been widely explored (Frick (1985) and Dellaert et al. (1996)). The focus of most of the research to date has emphasized the synthesis of emotional speech (Murray and Arnott (1996)).

When studying emotions in speech one needs to make the following distinctions:

- the emotional attitude of the speaker towards the hearer,
- the emotional attitude of the speaker towards the message, or
- the emotional state of the speaker (Halliday (1970)).

Our investigation will focus on the emotional state of a speaker. This has been widely studied in psychology and psycho-linguistics in an attempt to determine which acoustic and prosodic features (e.g. speaking rate, intonation, intensity) encode the emotional state of a speaker (Scherer et al. (1991)). In this investigation we will show how both acoustic and prosodic information can be used to detect the underlying emotional state of a speaker.

In the next section, we will describe how acoustic information can be combined with prosodic information and integrated into a hidden Markov model (HMM) architecture, in which additional suprasegmental states allow dynamic observation rates depending whether we consider acoustic or prosodic information. This suprasegmental hidden Markov model (SPHMM) architecture permits observations at a time scale appropriate for the phenomena to be modeled. We will make use of this property of SPHMMs and introduce prosodic information into the overall processing.

In Sect. 3 we will describe two experiments we conducted with our system to detect the emotional state of a speaker. For these experiments we used a corpus in which we asked drama students to utter sentences pretending to be in one of five emotional states: happy, sad, angry, afraid, or neutral.

We will conclude this investigation with a summary, describe extensions to the system we intend to incorporate in the near future, and discuss the potential of an SPHMM based architecture to integrate information from different modalities.

2 Suprasegmental Hidden Markov Models

Suprasegmental hidden Markov Models (SPHMM) permit the summarization of several states within a hidden Markov model into what we will call a suprasegmental state. These suprasegmental states allow the consideration of the observation sequence spanned by their constituent states, i.e., these suprasegmental states can look at the observation sequence through a larger window. Suprasegmental states allow observations at rates appropriate for the phenomena they are intended to model. For example, prosodic information can not be observed at a rate which is used for acoustic modeling. Prosodic information applies, for example, to syllables, words, or phrases but can not be observed within a time window of 10ms, the time frame in which acoustic events are usually looked at. In our application acoustic events are modeled using conventional hidden Markov states, while prosodic events at the phone, syllable, word, and utterance level are modeled using suprasegmental states. The basic idea of an SPHMM is given in Fig. 1. To combine acoustic and suprasegmental information we use the following formula:

$$\begin{aligned}\log P(\mathit{acoustic\ model}, \mathit{suprasegmental\ model} \mid \mathit{speech\ signal}) = \\ (1-\lambda) * \log P(\mathit{acoustic\ model} \mid \mathit{speech\ signal}) + \\ \lambda * \log P(\mathit{suprasegemental\ model} \mid \mathit{speech\ signal}).\end{aligned} \tag{1}$$

That is, each time we leave a suprasegmental state (e.g. a phone or syllable) we add the log probability of this suprasegmental state given the respective suprasegmental observations within the speech signal to the log probability of the current acoustic model given the respective acoustic observations within the speech signal. The weight factor, λ, is determined empirically. For more details on the theory of SPHMMs see Polzin (to appear).

2.1 Suprasegmental Observations

We use suprasegmental states to capture the prosodic properties of phones, syllables, words, and utterances because they allow us to make observations at a time scale suitable for prosodic phenomena. Suprasegmental observations comprise information about the duration of the respective segment, information about fundamental frequency (pitch), and intensity. While it is possible to determine intensity and fundamental frequency for any given time point in the speech signal, this information becomes far more meaningful if we can observe intensity and fundamental frequency over the duration of, say, a syllable or word. This observation

Fig. 1. Suprasegmental Hidden Markov Model. The hidden Markov states q_1, q_2, and q_3 form a suprasegmental state p_1 (e.g. a phone). The states q_4, q_5, and q_6 form a different suprasegmental state p_2. These two suprasegmental states themselves constitute another suprasegmental state p_3, e.g. a syllable. Transition probabilities between hidden Markov states are represented by a_{ij} where i indicates the state we are leaving and j the state we are going to. Transition probabilities between suprasegmental states are represented with b_{ij} where i denotes the suprasegmental state we are leaving and j denotes the suprasegmental state we are going to.

then allows the derivation of additional observations, such as mean and variance, the correlation between intensity and fundamental frequency, and whether intensity or fundamental frequency is steady, falling, or rising over the segment in question. Note that within a conventional HMM, observations have to be at a constant rate and, thus, it is not possible to look at the dynamic behavior of intensity and fundamental frequency over the course of a syllable or word if, at the same time, we want to observe acoustic events at a much smaller time scale.

The choice of suprasegmental observations has to reflect two issues:

1. In principle, these observations have to be computed for every possible segmentation within the Viterbi. In order to get an acceptable run time behavior, computing these observations should not be unreasonably computationally expensive.
2. These observations have to be robust with respect to noise and idiosyncrasies of speakers.

3 Experiments

3.1 The Corpus

We hand generated 50 sentences for the corpus. These sentences were comprised of questions, statements, and orders. The sentence length varied from 2 to 12 words; the mean sentence length was 5.8 words. The corpus was comprised of 291 word tokens (87 types).

We asked 5 drama students to pronounce the sentences according to the emotional label given in square brackets at the beginning of the sentence on a computer screen. The students were asked to portray each of these sentences in each emotional mood (happy, sad, angry, and afraid). In addition, we asked for a neutral pronunciation for all 50 sentences. Thus we have a maximum of 250 sentences for a given student.

SennHeiser HMD 410 or SennHeiser HMD 414 microphones were used for all recordings. The recording system used was the Gradient Desklab Model 14, with a sampling rate of 16 khz. All recordings were transcribed by hand.

3.2 Human Performance

We conducted a small informal experiment to determine the human performance on detecting the underlying emotional state of the speaker. The subjects had to listen to the utterances of one speaker played back in random order. The task of the subject was to choose one emotion out of four (happy, sad, angry, or afraid). Human performance was at about 70% accuracy. Note that the baseline is 25% (random guessing).

3.3 Baseline

For the following experiments we used the Janus speech recognition system (Zeppenfeld et al. (1997)) which was trained independently on a different corpus of spontaneous speech (English Spontaneous Scheduling Task, ESST). The word accuracy (WA) on this corpus was about 80%. We used this recognition system to determine the influence of emotional speech on word accuracy. The resulting word accuracy is given in Table 1. The word accuracy dropped about 10% for all emotions except for "angry" when compared with the neutral pronunciation.

Table 1. Word accuracy depending on the emotional state of the speaker in percent

Emotion	Happy	Afraid	Angry	Sad	Neutral
WA	55.8%	54.0%	65.9%	55.6%	65.9%

3.4 Training

Training of the SPHMMs is very similar to training of conventional HMMs. The only addition is that it is necessary to train suprasegmental models on top of acoustic models.

For this investigation, we derived emotion-dependent models, i.e., emotion-dependent acoustic and suprasegmental models. For example, for a word, we had four different suprasegmental word models: "happy" , "sad" , "angry", and "afraid". We used about 70% of the corpus for training acoustic and suprasegmental models. The rest of the corpus was used for testing.

3.5 Testing

The underlying emotional state was determined the following way:

1. The utterance was recognized using an emotion-independent recognition system as described in Sect. 3.3.

2. Using an emotion-dependent recognition system, i.e. a system based on emotion-dependent acoustic or suprasegmental models, we looked for the highest probability that the sentence as recognized in step 1 was produced by the emotion-dependent models (forced alignment), i.e.,
$$P(speech\ signal \mid sentence, models_i), \tag{2}$$
where h, s, af, and an stand for happy, sad, afraid, and angry, respectively. We tested for all four emotions and, thus, obtained four probabilities, one for each emotion.
3. The four probabilities as returned in step 2 were compared. We took the highest probability to be indicative of the actual emotional state of the speaker, i.e. we maximized (2):
$$emotional\ State = arg \max_{i \in \{h,s,af,an\}} P(speech\ signal \mid sentence, models_i). \tag{3}$$

Experiment 1 (Acoustic Models). In the first experiment we developed emotion-dependent acoustic models to obtain four emotion-dependent speech recognition systems. Based on these models, we determined the emotional state of the speaker following the procedure as outlined in the section above, where:

$$\begin{aligned} &emotional\ State = \\ &\quad arg \max_{i \in \{h,s,af,an\}} P(speech\ signal \mid sentence, acoustic\ models_i). \end{aligned} \tag{4}$$

The emotion detection accuracy is given in Table 2. The overall emotion detection accuracy is 65.4%. Using acoustic models enabled the system to detect the correct emotional state well above chance level. The high accuracy with which the underlying emotion of utterances spoken in an angry state were recognized seems to correlate with the high word accuracy for these utterances as given in Table 1.

Table 2. Emotion detection accuracy in percent using emotion dependent acoustic models

Emotion	Happy	Afraid	Angry	Sad
Emotion Accuracy	60.0%	73.0%	83.0%	46.0%

Experiment 2 (Suprasegmental Models). Starting with emotion-dependent acoustic models we included emotion-dependent suprasegmental models to see whether prosodic information would add discriminative power to our system. The influence of suprasegmental information on the overall probability computation was regulated by a weight factor, λ, mentioned in Sect. 2. We determined λ empirically on an independent development set. Using emotion-dependent acoustic and suprasegmental models we detected the emotional state following the procedure as described in Sect. 3.5, where:

$$\begin{aligned} &emotional\ State = \\ &arg \max_{i \in \{h,s,af,an\}} P(speech\ signal \mid sentence, acoustic\ models_i, suprasegmental\ models_i). \end{aligned} \tag{5}$$

The resulting emotion detection accuracy is given in Table 3.

The overall emotion detection accuracy is 72.25% which amounts to an absolute improvement of about 7%. In particular, using prosodic information appears to help the detection of “happy” and “sad”.

Table 3. Emotion detection accuracy in percent using emotion dependent suprasegmental models

Emotion	Happy	Afraid	Angry	Sad
Emotion Accuracy	93.8%	60.0%	77.9%	59.6%

4 Conclusions

Our investigation shows how acoustic and prosodic information can be combined and integrated into a HMM-based speech recognition system using suprasegmental states. Using both, acoustic and prosodic information the system nearly achieves human performance levels when trying to detect the emotional state of a speaker. We show that prosodic information is essential for a reliable detection of the underlying emotional state of a speaker (7% absolute improvement).

To test our system's capabilities to detect the emotional state of a speaker we certainly have to test it on additional real world data. We intend to apply these techniques to human-to-human telephone conversations.

The choice of words is most probably another indication of a speaker's emotional state. A straightforward extension to our system would be to use emotion-dependent language models. Another extension which comes to mind is to combine audio and visual information. This combination proved to be quite useful in speech recognition where information about lip movements was combined with acoustic information and successfully reduced the word error rate of the overall system (Bregler et al. (1993) and Stiefelhagen (1997)).

We consider SPHMMs to be a suitable tool to integrate observations from different modalities because they allow for observation at a rate appropriate for each modality. We demonstrated this possibility by the way we integrated prosodic information into the overall processing using an SPHMM architecture. Following up on this approach, it is conceivable to integrate additional modalities, such as gestures, facial expressions, heartbeat, or body temperature. The SPHMM architecture allows the consideration of each modality at an appropriate rate and synchronizes it with observations from other modalities. For example, we could develop word models which have access to information about the dynamic behavior of the fundamental frequency (pitch) and intensity, about accompanying hand gestures, and facial expressions.

References

C. Bregler, S. Manke, H. Hild, and A. Waibel. Improving connected letter recognition by lip reading. In *ICASSP*, Minneapolis, 1993. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing.

F. Dellaert, T.S. Polzin, and A. Waibel. Recognizing emotions in speech. In *ICSLP*, 1996.

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

M.A.K. Halliday. *A course in spoken English.* Oxford University Press, London, England, 1970.

S. Kaiser and T. Wehre. Automated coding of facial behavior in human-computer interactions with FACS. *Journal of Nonverbal Behaviour*, 16(2):67–83, 1992.

I.R. Murray and J.L. Arnott. Synthesizing emotions in speech: Is it time to get excited? In *ICLSP*, 1816–1819, 1996.

T.S. Polzin. Suprasegmental hidden Markov models. Technical report, School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213, USA, to appear.

R. Stiefelhagen, U. Meier, and J. Yang. Real-time lip-tracking for lip reading. In *Eurospeech 9*, 1997.

K.R. Scherer, D.R. Ladd, and K.E.A. Silverman. Vocal cues to speaker affect: Testing two models. *Journal of the Acoustic Society of America*, 76:1346–1356, 1984.

K.R. Scherer, R. Banse, H.G. Wallbott, and T. Goldbeck. Vocal cues in emotion encoding and decoding. *Motivation & Emotion*, 2(15):123–148, 1991.

K.R. Scherer and R. Banse. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70, 614–636, 1996.

S. Kaiser and K.R. Scherer. Models of 'normal' emotions applied to facial and vocal expressions in clinical disorders. In: W.F. Flack and J.D. Laird (Eds.), *Emotions in psychopathology*, New York:Oxford University Press, to appear.

A. Takeuchi and K. Nagao. Communicative facial display as a new conversational display. Technical report, SCSL-TR-92-019, Sony Computer Science Laboratory, Tokyo, 1992.

T. Zeppenfeld, M. Finke, K. Ries, M. Westphal, and A. Waibel. Recognition of conversational telephone speech using the Janus speech engine. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Munich , Germany, 1997.