PRONUNCIATION VARIATIONS IN EMOTIONAL SPEECH

Thomas S. Polzin and Alexander Waibel

Interactive Systems Laboratories
Carnegie Mellon University
Pittsburgh, PA 15233, USA
{tpolzin,waibel}@cs.cmu.edu

ABSTRACT

In this paper we demonstrate how the emotional state of the speaker influences his or her speech. We show that recognition accuracy varies significantly depending on the emotional state of the speaker. Our system models the pronunciation variation of emotional speech both at the acoustic and prosodic level. We show that using emotion-specific acoustic and prosodic models allows the system to discriminate among four emotions (happy, sad, angry, and afraid) well above chance level. Finally, we show that emotion-specific modeling improves the word accuracy of the speech recognition system when faced with emotional speech.

1. INTRODUCTION

Many factors have an impact on the pronunciation of a given word, for example, the position of the word within the utterance, dialect, age, and gender of the speaker. An additional factor on the pronunciation has recently become the focus of research: How does the emotional state of the speaker influence his or her speech? We will show that using the JANUS system (Zeppenfeld et al. ([1997])), recognition accuracy varies depending on the emotional state of the speaker. The accuracy drops significantly when compared to the accuracy with which neutral speech is recognized. To build robust human-computer interfaces we have to account for this variation.

In order to address this issue it is crucial to be able to detect the emotional state of the speaker with a high degree of accuracy. Research in psycholinguistics indicates that prosodic information such as pitch and speaking rate is important in human recognition of underlying emotions in speech (Scherer et al. [1984, 1991]). Our system uses both acoustic (segmental) and prosodic (suprasegmental) information to model the pronunciation variation in emotional speech. We use a variation of hidden Markov models to integrate prosodic information into the recgonition process. We show that we can use emotion specific acoustic and suprasegmental models to detect the underlying emotional state of the speaker with an accuracy comparable to the performance of humans on this task. We then demonstrate that by using emotion-specific acoustic models we can improve the word accuracy of the recognition system significantly. We will conclude this investigation with a summary and extensions to the system we intend to incorporate in the near future.

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. 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. 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. To combine acoustic and suprasegmental information we linearly combine acoustic and suprasegmental probabilities. That is, each time we leave a suprasegmental model, for example a phone or syllable, we add the log probability that the suprasegmental observations given in the speech signal were produced by this suprasegmental model to the log probability that the acoustic observations given in the speech signal were produced

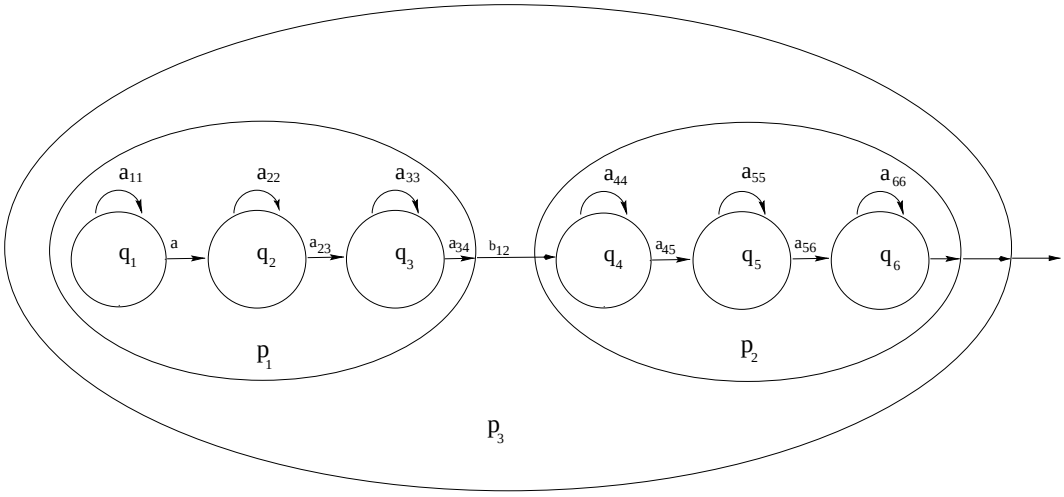

Figure 1: Suprasegmental Hidden Markov Model. The hidden Markov states q_1, q_2, and q_3 form a suprasegmental state p_1 (e.g. a phone). The states q_4, q_5, and q_6 form a different suprasegmental state p_2. These two suprasegmental states themselves constitute another suprasegmental state p_3, e.g. 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.

by the respective acoustic model. The weight factor 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 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 25% for all emotions except for "angry" when compared with the neutral pronunciation. The big discrepancies in word accuracy

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

Emotion	Happy	Afraid	Angry	Sad	Neutral
WA	51.6	46.0	64.2	45.6	71.9

depending on the underlying emotion – ranging from 46% to 71% – demonstrates the necessity of modeling the pronunciation variation in emotional speech.

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-specific models, i.e., emotion-specific 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-specific recognition system, i.e. a system based on emotion-specific acoustic or suprasegmental models, we looked for the highest probability that the sentence as recognized in step 1 was produced by the emotion-specific models (forced alignment), i.e.,

$$P(speech\ signal \mid sentence, models_i), \qquad (1)$$

   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 (1):

$$emotional\ State = arg \max_{i \in \{h,s,af,an\}} P(signal \mid sentence, models_i). \qquad (2)$$

To classify the detection accuracy we use precision/recall and the corresponding f1 value. For an emotion i we define:

$$precission_i = C_i/T_i \qquad (3)$$

$$recall_i = C_i/I_i \qquad (4)$$

$$f1_i = \frac{2 * precission_i * recall_i}{precission_i + recall_i} \qquad (5)$$

where C_i denotes the number of sentences in the test corpus judged correctly by the system to have the underlying emotional state i. Similar, T_i denotes the total number of sentences classified as i by the system and I_i refers to the actual number of sentences in the test corpus whose speaker was portraying emotional state i. Note that our system is not speaker independent. We tested and trained on the same speaker (multi speaker system).

Table 2: Emotion detection performance in percent using emotion dependent acoustic models

Emotion	Happy	Afraid	Angry	Sad
precession	0.80	0.56	0.90	0.56
recall	0.76	0.63	0.80	0.59
f1	0.78	0.59	0.85	0.57

3.5.1. Experiment 1 (Acoustic Models).

In the first experiment we developed emotion-specific acoustic models to obtain four emotion-specific 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:

$$emotional\ State = arg \max_{i \in \{h,s,af,an\}} P(signal \mid sentence, a\text{-}models_i). \quad (6)$$

The emotion detection performance is given in Table 2.

The overall f1 score is 0.69. 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.

3.5.2. Experiment 2 (Suprasegmental Models).

Starting with emotion-specific acoustic models we included emotion-specific 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-specifc acoustic and suprasegmental models we detected the emotional state following the procedure as described in Sect. 3.5, where:

$$emotional\ State = arg \max_{i \in \{h,s,af,an\}} P(signal \mid sentence, a\text{-}models_i,\ spm\text{-}models_i). \quad (7)$$

The resulting emotion detection performance is given in Table 3.

The overall f1 is 0.73, compared to an f1 of 0.69 in the previous experiment where we used only acoustic models. In particular, using prosodic information appears to help the detection of "happy" and "afraid'.

Table 3: Emotion detection performance using emotion dependent acoustic and suprasegmental models

Emotion	Happy	Afraid	Angry	Sad
precision	0.77	0.57	0.95	0.72
recall	0.88	0.76	0.80	0.49
f1	0.82	0.65	0.87	0.58

3.5.3. Experiment 3 (Recognizing Emotional Speech)

In our last experiment we investigated whether emotion-specific acoustic model are able to improve on the word accuracy. For this experiment we assumed that we knew the emotional state of the speaker and used acoustic models corresponding to this emotional state. The resulting word accuracy is given in Table 4. When we compare these word accuracies with the accuracies gained in the baseline experiment in section 3.3 we, first, see a general significant improvement in the overall performance and, second, the discrepancies in the recognition accuracy among the different emotions are reduced.

Table 4: Word accuracy in percent depending on the emotional state of the speaker using emotion dependent acoustic models

Emotion	Happy	Afraid	Angry	Sad	Neutral
WA	66.9	67.90	63.8	70.1	77.6

We then tried to rescore the word lattices produced by the previous recognition process with emotion dependent suprasegmental models but could not achieve an improvement.

4. CONCLUSIONS

Our investigation shows that the pronunciation variance in emotional speech allows the detection of the underlying emotional state of the speaker. We demonstrate that both acoustic and prosodic information carries important information about the encoded emotion. Our system was able to detect the emotional state well above chance level. Finally, we showed that by using emotion dependent acoustic models we improved the word accuracy of the recognition system.

We were not able to demonstrate that the improvement in emotion detection given by adding suprasegmental information to acoustic information led directly to a significant improvement in word recognition accu-

racy. However, we expect that as further development and refinement of the suprasegmental models lead to even greater increases in emotion detection accuracy, the corresponding jump in word recognition accuracy will become more evident.

5. REFERENCES

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

[1984] 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.

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

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

[1997] 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.