Analysis of emotion in speech using perceived and automatically extracted prosodic features

Suk-Myung Lee, Jeung-Yoon Choi

School of Electrical & Electronic Engineering, Yonsei University,
134 Sinchon-dong, Seodaemun-gu 120-749, Seoul, Republic of Korea
pooh390@dsp.yonsei.ac.kr, jychoi@yonsei.ac.kr

Abstract

This study investigates the relationship between emotional states and prosody. A prosody detection algorithm was applied to emotional speech to extract accents and intonational boundaries automatically and these were compared with hand-labeled prosodic units. The measurements used in the detection algorithm are derived from duration, pitch, harmonic structure, spectral tilt, and amplitude. The utterances are part of a Korean emotional database subset in which 10 sentences were spoken by 6 speakers over 4 emotions (neutral, joy, sadness and anger). By comparing the probabilities of occurrence and temporal patterns of events that were detected prosodic events between neutral speech and emotional speech, our experiments find different distributions for each emotion. Overall, joy and anger tended to have more events classified as accents compared to other emotions. Also, sadness had more events corresponding to boundaries. In addition, joy had more events classified as accents at the beginning of utterances, while anger had more accents at the ends of utterances. These results indicate that prosodic characteristics can be useful for classification of emotion and in synthesizing emotional speech.

Index Terms: accent, boundary, prosodic feature, emotion speech

1. Introduction

Speech, which comprises both lexical and nonlexical information, is the easiest and most natural way to exchange information between humans. Over the past few decades, much research has been carried out in speech recognition systems to analyze lexical messages spoken by humans. More recently, one of the major goals in this research area is the detection of nonlexical information such as emotion.

Emotion expression through speech is commonly expressed through changes in prosody and voice quality. Previous research shows that pitch, energy and durational features are important cues in representing each emotion [1, 2, 3]. Recently, studies have found correlation between emotion and global features, such as pitch contours, and between prosodic units, such as accents and boundaries [4].

The purpose of this work is to make a first approach towards finding the relationship between emotion and prosody. More precisely, the objective is to compare the prosody pattern for emotion using perceptual labels and results from automatic prosody detection. We use 4 units (accent, boundary, accent plus boundary, and neutral) in a simplified version of TOBI [5] to analyze prosody.

The paper is organized as follows. First, a description of the prosodic labels and the database used for the experiments is given. Then, the automatic prosody detection algorithm is summarized. Next, the results obtained in these experiments are presented, and discussions follow.

Table 1: *The list of simplified prosodic units used in this study*

a	accent	**b**	boundary
ab	accent & boundary	**n**	neutral

2. Experiments

2.1. Prosodic labels

A set of four simplified labels, which describe the broad prosodic classes, were chosen: they are **a** (accent), **b** (boundary), **ab** (accent and boundary), and **n** (neutral). Each vocalic segment is assumed to belong to one of the four labels. These units are simplified versions of units in the TOBI system. All accents, including rising or falling accents (which are separately grouped into classes based on primary accent tones such as H*, !H* and L* in the TOBI system), are represented by **a**, and **b** represents intonational boundaries which consist of rising and falling boundaries (indicated by H-, !H-, L-,H-H%, H-L%, !H-H%, !H-L%, L-H% and L-L% in the TOBI system). In this paper, we grouped intermediate and full intonational boundaries into one class. For the cases when an accent and a boundary tone coincide, the composite label **ab** is used. The rest are assigned to the neutral vocalic segments **n**.

2.2. Data corpus

In this study, we used the KED database (Korean emotion database) [6] to study the relation between prosody and emotion in speech. The emotions targeted by the database include joy, sadness, anger, and neutral. The sentences are 6 declarative and 4 interrogative Korean sentences, each comprising 3 phrases, and each is uttered by three males and three females who are all Korean amateur actors and actresses. The utterances are presented in Table 2. Each utterance is recorded in a quiet office environment using a DAT (digital audio taperecorder) with a 16kHz sampling rate and quantized to 16 bits. A subjective listening test conducted on a Korean audience of 30 using this emotion speech database resulted in a perceptual classification score of 78.2% [6]. In order to analyze the relationship between emotion and prosody, the database was labeled by 4 students familiar with speech and prosody. The database was labeled on a syllable basis, with each syllable marked with the simplified prosodic labels for accent and boundary.

Table 2: *10 Korean sentences used in this study*

Declarative sentence:
1. 같은 걸로 먹자. - Let's eat the same thing. (Kat'eun keol'ro meok'ja.)
2. 햇볕이 쨍쨍 비친다. - The sun is shining brightly. (Het'byeot'chi jjeng'jjeng pi'chin'da.)
3. 준비 다 됐어요. - We are all ready. (Choon'bi ta twet'seo'yo.)
4. 철수가 큰일 냈다. - Cheol'soo caused a big problem. (Cheol'soo'ga keun'il net'da.)
5. 어서 차를 불러요. - Please call a car quickly. (Eo'seo cha'reul bool'leo'yo.)
6. 다음에 다시 얘기하자. - Let's talk again later. (Ta'eum'e ta'shi ye'gi'ha'ja.)
Interrogative sentence:
1. 아직 전화 안 왔니? - Hasn't the call come yet? (A'jik cheon'hwa an wat'ni?)
2. 함께 갈 수 있겠니? - Can you go together? (Ham'kke kal soo it'get'ni?)
3. 숙제 다 했니? - Did you finish your homework? (Sook'je ta het'ni?)
4. 오늘 시간 있니? - Do you have time today? (O'neul shi'gan it'ni?)

2.3. Acoustic measurements

Acoustic cues related to the voice source examined in this paper in order to identify the presence of prosodic events can be divided into five categories: duration, pitch, harmonic structure, spectral tilt, and amplitude [7]. First, durational measurements include length of following pause and speech rate. Pitch measurements include fundamental frequency measured at the end of a vocalic segment, and the slope and the convexity of the fundamental frequency.

The slope was calculated as

$$\text{slope} = \frac{s(t_2) - s(t_1)}{t_2 - t_1}.$$

And the convexity was calculated as the sum of the difference between each signal point and the linear interpolation between the start and end values of the segment. That is,

$$\text{convexity} = \frac{\sum_{t=t_1}^{t_2} s(t) - h(t)}{t_2 - t_1},$$

where t_1 and t_2 are respectively the start and end times of the vocalic segments, $s(t)$ is the value of the measurements at time t, and $h(t)$ is the linear interpolated function,

$$h(t) = \frac{s(t_2) - s(t_1)}{t_2 - t_1} \cdot (t - t_1),$$

for $t_1 \leq t \leq t_2$, and $t_1 < t_2$, respectively.

Harmonic structure measurements include end value, slope and convexity of h1-h2, where h1 and h2 are the amplitudes of the first and second harmonics, respectively. Spectral tilt measurements include end value, slope and convexity of h1-a1, h1-a3, and a1-a3, where a1 and a3 are the amplitudes of the first and third formants, respectively. Finally, amplitude measurements included the end value, slope, and convexity of the rms.

2.4. Automatic prosody detection algorithm

The 23 voice source acoustic cues described in the study cited above were used to detect accent and boundary for the training and test sets from the KED database labeled with the simplified prosodic labels. The data set was used to find means and covariance matrices for 23 dimensional Gaussian distributions for nonaccented versus accented tokens (accent detection), and for nonboundary versus boundary tokens (boundary detection). This algorithm reported that detection experiments on the Boston University Radio Speech Corpus yielded equal error detection rates around 70% for accent and boundary detection.

3. Results

3.1. Experiment 1: Probabilities of occurrence for prosodic events within an utterance

The results comparing accents and boundary across emotions are shown in Figure 1 and in Figure 2. The probabilities were computed by counting the number of the instances of accents and boundaries and dividing the total number of syllables in utterances for each emotion.

$$\textit{probability of accent or boundary} = \frac{\textit{total number of accents or boundaries}}{\textit{total number of syllables}} \quad (1)$$

Analyzing the results for accents across emotions, emotional utterances detected more accents compared to neutral utterances. Especially, joy and anger had more accents than neutral emotions. These results indicate that even though speakers utter the same lexical sentence, they use more accents when they are joyful or angry. In the case of detecting boundaries, neutral, joy and anger emotions detected almost the same probabilities for boundaries. However, sad emotions detected slightly more boundaries compared to other emotions. These patterns can also be observed in the perceptual results. This behavior implies that people tend to produce more syllables which are classified as boundaries when they are sad.

It is well known that emotions of speech are expressed through pitch, intensity or voice quality. However, these results indicate that these measurements may be expressed through prosodic events, such as accents and boundaries, as cues for emotions. Next, in order to observe the relationship between emotions and prosody in more detail, we analyze temporal patterns of prosodic events across all emotions in Experiment 2.

3.2. Experiment 2: Temporal distribution of probabilities of occurrence for prosodic events

In Experiment 2, we divided utterances into 3 parts, the initial, middle and final phrases, to analyze temporal distributions of prosodic events for different emotions. The probabilities were also computed by counting the number of the instances of accents and boundaries and dividing the total number of phrases in utterances for each emotion.

$$\textit{probability of accent or boundary} = \frac{\textit{total number of accents or boundaries}}{\textit{total number of phrases}} \quad (2)$$

First, perceptual results are shown in Figure 3 and in Figure 4. From the results for accent, it was observed that joy and anger

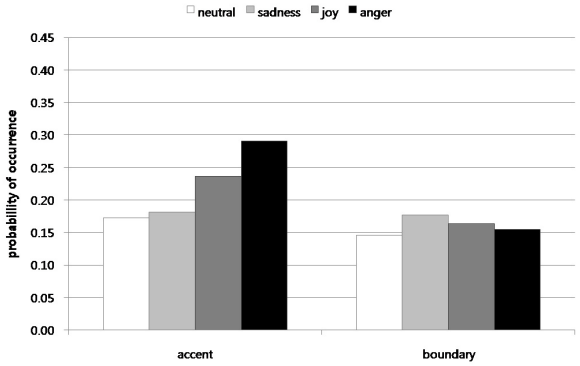

Figure 1: Perceptual results for occurrence of accent and boundary in an utterance for 4 emotion classes.

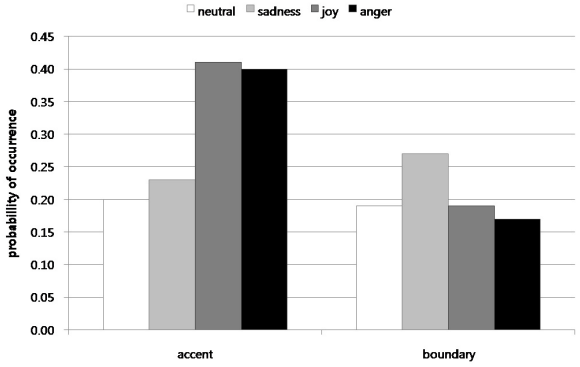

Figure 2: Automatic detection results for occurrence of accent and boundary in an utterance for 4 emotion classes.

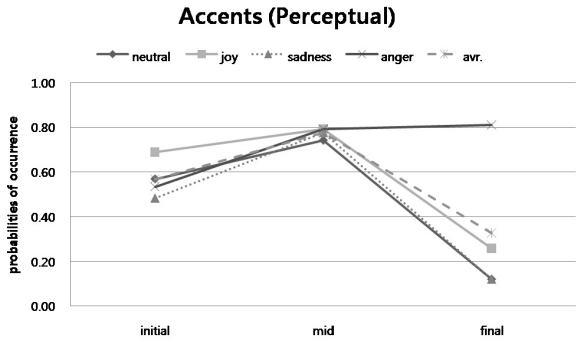

Figure 3: Perceptual results of probabilities of occurrences for accent among all emotions through initial, middle and final parts of utterances.

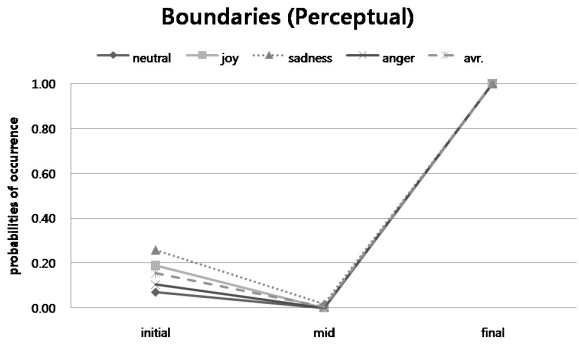

Figure 4: Perceptual results of probabilities of occurrences for boundary among all emotions through initial, middle and final parts of utterances.

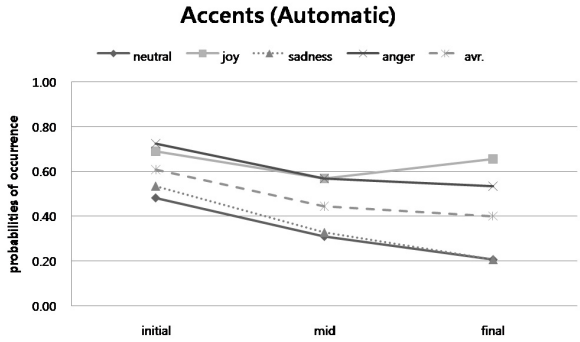

Figure 5: Automatic detection results of probabilities of occurrences for accent among all emotions through initial, middle and final parts of utterances.

have more syllables classified as accents. Especially, joy tended to display more accents at the initial phrases of the utterances. In contrast, much more accents were detected at the final phrases of angry utterances. In addition to the results in Experiment 1, these results indicate that the distribution of accents within the utterance are also important cues for emotion classification. Figure 4 shows that for boundaries, more events are detected at initial phrases for sad utterances. Also, there are almost no boundaries at the middle phrases of utterances, in contrast to the final phrases containing full intonational boundaries.

Figure 5 and Figure 6 show the distributions of accent and boundary across emotions from automatic prosody detection. From Figure 5, it can be seen that joyful and angry utterances have more accents throughout utterances compared with other emotions. However, the temporal patterns are quite different from perceptual results. This suggests that there may be acoustic cues that contribute to accent detection that are not being exploited in automatic detection. In fact, the cited results for accent detection are much lower than for boundary detection. Investigation into such cues may be the subject of future work. On the other hand, boundary detection results showed similar trends with perceptual results. Especially, sad utterances yielded more syllables detected as boundaries as in Experiment 1.

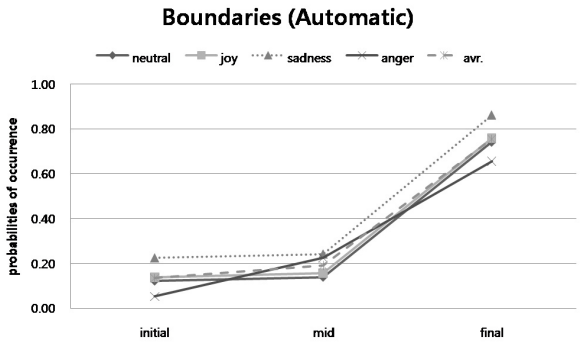

Figure 6: Automatic detection results of probabilities of occurrences for boundary among all emotions through initial, middle and final parts of utterances.

4. Summary and discussions

In this paper, we analyzed characteristics of prosody patterns for perceptual and automatic results for 4 emotions using a Korean emotional database with neutral, sad, joyful and angry utterances. It was shown that probability of occurrence for accents and boundaries were different for each emotion, and also temporal patterns of occurrence varied across the position of syllables within utterances. Especially, joyful and angry utterances tended to have more events perceived and classified as accents compared to other emotions. Sad utterances have more events corresponding to boundaries. These behaviors imply that prosodic events such as accents and boundaries can be used as cues for emotion expressions.

Future work will be application of our results to include more detailed observations between prosodic events and emotions, as well as including observations for larger databases and for other languages. These results are expected to be useful for classification of emotions as well as synthesizing emotional speech.

5. Acknowledgements

This work was supported in part by a grant from Microsoft Research Asia.

6. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction", IEEE Signal Processing Magazine, 2001

[2] Carl E. Williams, Kenneth N. Stevens, "Emotions and spech: some acoustical correlats", J. Acoust. Soc. Am. 52, 1972.

[3] D. Ververidis, C. Kortopoulos and Ioannis Pitas, "Automatic emotional speech classification", Procedings of Int. Conf. Acoust., Speech, Signal Process, 2004.

[4] V. Auberge, N. Audibert, and A. Rilliard, "Acoustic morphology of expressive speech : What about contours?", Proceedings of Speech Prosody 2004.

[5] M. Beckman and G. Ayers, "Guidelines for TOBI labeling (version 3.0)", The Ohio State University, 1997.

[6] K. Bong-Seok, "Text independent emotion recognition using speech signals", Master Thesis,The Yonsei University, 2000.

[7] J. -Y. Choi, M. Hasegawa-Johnson, and J. Cole, "Finding intonational boundaries using acoustic cues related to the voice source", J. Acoust. Soc. Am. 118, 2579-2587, 2005.

[8] Z. Inanoglu and S. Young, "Intonation modelling and adaptiation for emotional prosody generation", Proceedings of ACII, 286-293, 2005.

[9] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence", IEEE Trans. Audio, Speech, and Language Processing 16, 216-228, 2008.

[10] D. -N. Jian,W. Zhang, L. -Q. Shen, and L. -H. Cai, "Prosody analysis and modeling for emotional speech synthesis", Procedings of Int. Conf. Acoust., Speech, Signal Process, 281-284, 2005.