Analyzing Dialogue Data for Real-World Emotional Speech Classification

Ryuichi Nisimura, Souji Omae, Hideki Kawahara, Toshio Irino

Faculty of Systems Engineering, Wakayama University, Japan

nisimura@sys.wakayama-u.ac.jp

Abstract

In order to obtain an understanding of the user's emotion in human-machine dialogues, an analysis of dialogical utterances in the real world was performed. This work comprises three major steps. (1) The actual conditions of 16 basic emotions were evaluated using Japanese child voices, which were collected through the field test of the public spoken dialogue system. (2) Two factors were derived by a factor analysis. The factors were defined as fundamental psychological factors representing "delightful" and "hateable" emotions. (3) The relationships between the factors and the physical acoustic features were investigated to establish a capability to sense a user's mental state for the dialogue system. In the experimental discriminations between the delightful and hateable emotions, a correct rate of 98.8% was achieved in classifying child's utterances by the SVM (Support Vector Machine) with 11 acoustic features.

Index Terms: emotional speech, classification, factor analysis, dialogue system, real world

1. Introduction

We have been analyzing emotional voices in order to improve the human-machine dialogue. At some future date, spoken dialogue systems will be utilized in our daily lives. As the spoken language is one of the most familiar means of communicating with others, an intelligent human-machine dialogue will provide increased opportunities for using computers. A capability to sense a user's mental state will be useful in establishing familiarity between a system and a user. However, it is inadequate to investigate the real-world conditions of emotional voices because many of the previous works targeted acting utterances[1, 2]. Due to the difficulty of data collection, real emotional speech has not been explored, except in a few cases such as call centers[3, 4]. This paper examines the actual emotional conditions in natural dialogical utterances, which were recorded in the field test of our dialogue system.

We also propose fundamental psychological factors that represent user's mental status. A factor analysis revealed two factors showing "delightful" or "hateable" emotions. In order to classify the emotional voices automatically, the relationships between the psychological quantities and the physical acoustic features were investigated. The previous works have not provided articulate definitions of the psychological quantities. Further, we have carried out experimental discriminations of actual child voices by using the SVM (Support Vector Machine)[5].

This paper comprises five major sections. In Section 2, we will discuss the emotional evaluation of Japanese utterances collected through the public spoken dialogue system. The factor analysis is described in Section 3 and the fundamental psychological factors are proposed. Section 4 presents experimental discriminations that classify actual child's utterances while investigating the relationships of the factors and the physical acoustic features. Section 5 concludes the paper and also describes the future works.

2. Emotional evaluations for natural dialogical utterances

Dialogical utterances were analyzed by 5-grade evaluations using basic emotions with 16 dimensions listed in Table 1. An operator (adult male) rated whether the 16 emotions were included in the segmented speech. Human feelings comprise a blend of some basic emotions. It is possible to observe two or more emotions simultaneously. For example, the voice by a user who was experiencing strong feelings of anger and sadness was marked as five on "Anger" and "Sadness." The utterance of the excitement feeling was marked as three on "Excitement." This multidimensional evaluations were carried out with one operator's subjectivity. The operator's sensation was reflected on the evaluated values. The system developed on this results would be able to succeed to one regular man's sense.

The 16 emotions shown in Table 1 were determined by referring to the basic emotions given by Ekman[6], Schlosberg[7], and Russell[8]. It should be noted that "Neutral" implies a state of the user maintaining his/her composure like a poker face. Its meaning differs from the complete absence of emotions.

The target data were extracted from natural utterances which were recorded by the public spoken dialogue system. Figure 1 shows our dialogue system, which is called the "Takemaru-kun" system[9, 10]. This practical speech-oriented guidance system is permanently located at the en-

Table 1: Sixteen basic emotions for the 5-grade evaluations.

Pleasure	Unpleasure	Fear	Surprise
Anger	Contempt	Sadness	Joy
Excitement	Mirth	Tiredness	Tension
Contentment	Depression	Pressure	Neutral

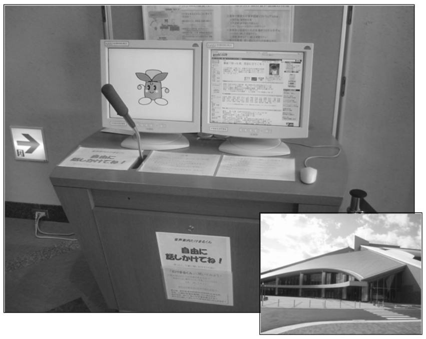

Figure 1: The speech-oriented guidance system "Takemaru-kun" and the Ikoma Community Center.

trance hall of the Ikoma Community Center in order to inform visitors about the center and Ikoma City. We have been examining the spoken dialogue interface with graphical Web animations through a long-term field test that commenced in November 2002[10].

A total of 2,699 sentences, which were uttered by children from April 1 to April 20, 2003, were evaluated. Hereafter, we focus on the voices of children because 58.1% of the voices in the Takemaru-kun system were uttered by children[10]. And, an extraction of emotional samples required frank and natural voices uttered by artless children, as opposed to educated adults who feign composure.

3. Factor analysis

After the evaluations, a factor analysis with the varimax rotation was performed in order to estimate the fundamental factors from the 5-grade values of the 16 emotions. The number of underlying factors was set to 8. We selected the fundamental factors on the basis of the factor loadings, which were computed by the maximum likelihood method. The factor scores were obtained by the Bartlett scoring method. Figure 2 shows the results in the factor loadings. The y-axis indicates the factor loadings for the 16 emotions. Figure 3 illustrates the cumulative contribution rates.

Evident connections were obtained between the factors and the emotions, as shown in Figure 2. It was certain that factor 1 represented negative emotions, such as "Anger," "Unpleasure," "Pressure," and "Contempt." Factor 2 showed specific correlations with positive emotions, such as "Pleasure," "Mirth," and "Joy." Therefore, factor 1 was named as a "hateable" factor, and factor 2 was termed as a "delightful" factor.

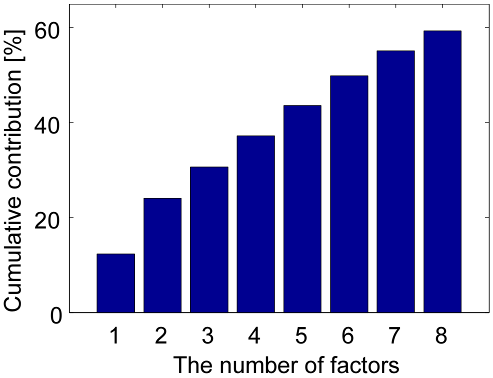

Figure 3: Cumulative contribution rates.

Some slight tendencies were also observed as follows: "Tension" of factor 3, "Contentment" of factor 4, "Tiredness" of factor 5, "Contempt" of factor 6, "Sadness" and "Depression" of factor 7, and "Pressure" and "Excitement" of factor 8. Almost all the factors showed interesting results for "Neutral." We considered "Neutral" to have independent tendencies, unlike the other emotions.

We now define the fundamental psychological factors that comprise the first two factors, delightful and hateable. Detailed investigations would be necessary in the future studies. Figure 3 shows that the cumulative contribution rates have stayed low. Further examination is required to raise the contribution rates by improving the analysis method

4. Experiments of automatic emotional classification

We have examined the automatic emotional classification of child's utterances between delightful and hateable. It implies that the relationships of the proposed factors and the physical acoustic features were investigated. As for a classifier, we used the SVM[5], which performs two-value classifications and has been applied to natural language processing, such as text categorization[12].

4.1. Acoustic features

The 16 acoustic features were adopted as follows.

- Maximum, median, interquartile range, and variance of pitch
- Maximum, median, interquartile range, and variance of pitch rising slopes
- Maximum, median, interquartile range, and variance of power
- Maximum, median, interquartile range, and variance of power rising slopes

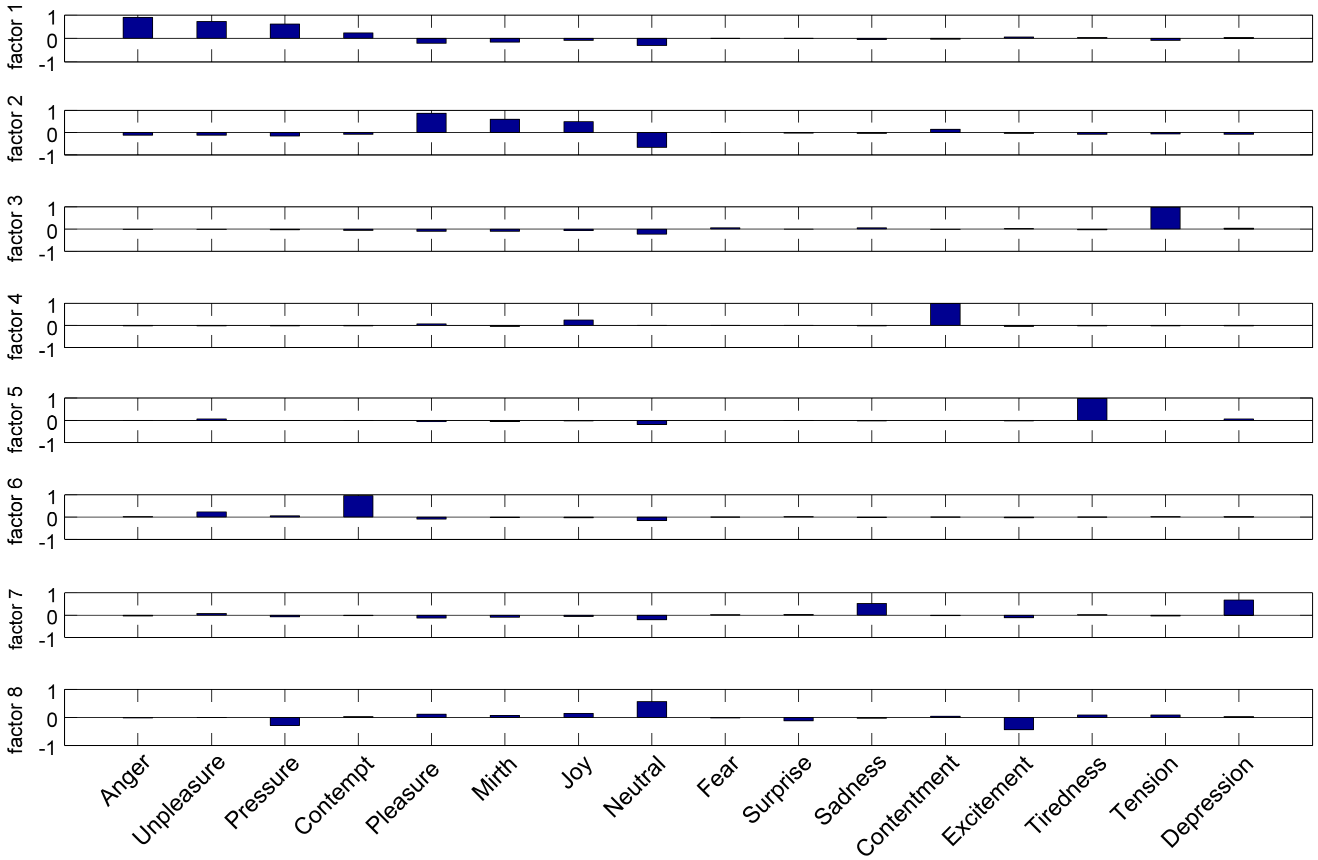

Figure 2: Results of the factor analysis. The y-axis indicates factor loadings to each emotion.

The pitch and the power were estimated by using the YIN estimator[11], and smoothed by a median filter with a 17 msec window length. We have dealt with the pitch extractions only in those voiced parts whose power was limited to -30 dB from the maximum in the segmented speech.

We considered this configuration of acoustic features based on the 87 features proposed by Ververidis[1], which include statistical features of pitch, spectrum, and energy.

4.2. Regression analysis

We applied regression analysis to predict the correlations between the fundamental psychological factors and the 16 features. The partial regression coefficients are shown in Tables 2 and 3. Each cell indicates the coefficient after normalization by the maximum value. The features with large absolute values would provide high mutuality, which would enable discriminating between the factors.

4.3. Experimental results

Figure 4 shows the experimental results of the SVM with the ANOVA kernel. The SVM performed two-class classifications among the two factors. The correct rates for discriminating between the delightful and hateable emotions are shown on the y-axis. The x-axis presents the thresholds of the partial regression coefficients, which were ordered by the absolute value in Tables 2 and 3. The features having large coefficients were examined sequentially after sorted.

A total of 600 samples were provided for the tests from the child utterances analyzed in Section 3. They contain 300 delightful and 300 hateable samples so that the SVM could cover the same number of positive and negative data. The correct rates for discriminating the two-class samples were averaged by the jackknife method with 600 repeats, where one sample was evaluated against the SVM model built from the other 599 samples.

A correct rate of 98.8% was obtained at the threshold of 0.1, as shown in Figure 4. We confirmed the peak performance with 11 acoustic features instead of using all the features. The addition of the variances of power and those of power rising slopes induced the largest change. We can conclude that our emotional discrimination was related to the following 11 features.

- Maximum, median, and interquartile range of pitch rising slopes

Table 2: Partial regression coefficients of the scores of the hateable factor. (factor 1)

	Pitch	Power	Rising slope	
			Pitch	Power
Maximum	0.00	1.0	-0.52	0.28
Median	0.00	0.26	-0.85	-0.62
Interquartile range	-0.01	0.23	0.65	-0.15
Variance	0.00	-0.18	0.00	0.10

Table 3: Partial regression coefficients of the scores of the delightful factor. (factor 2)

	Pitch	Power	Rising slope	
			Pitch	Power
Maximum	0.00	-0.35	0.15	-0.1
Median	-0.01	0.44	-0.83	0.74
Interquartile range	0.01	-0.14	0.18	-0.53
Variance	0.00	0.13	0.00	0.10

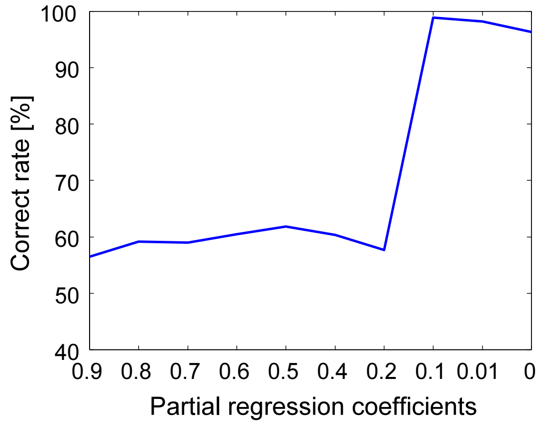

Figure 4: The correct discrimination rates for the delightful and hateable emotions.

- Maximum, median, interquartile range, and variance of power
- Maximum, median, interquartile range, and variance of power rising slopes

5. Conclusions

The dialogical utterances were analyzed in order to realize a capability to sense the user's emotional state in human-machine dialogues. Emotional conditions with 16 dimensions were evaluated subjectively using actual dialogue data collected through the Takemaru-kun system. Two factors were derived by the factor analysis. As a result, fundamental psychological factors representing "delightful" and "hateable" emotions were proposed. The relationships between the proposed factors and the physical acoustic features were investigated. Experimental discriminations using natural voices uttered by artless children were examined. It showed 98.8% correct rate for discriminating between the delightful and hateable emotions with 11 acoustic features.

In our future studies, we plan to carry out a further analysis that considers the other factors presented in Section 3. We also plan an investigation that makes use of adult's utterances. And, we will carry out evaluations of the proposed scheme in the real spoken dialogue system.

6. Acknowledgements

This work was partially supported by the e-Society project of MEXT (Ministry of Education, Culture, Sports, Science and Technology), Japan.

7. References

[1] D. Ververidis et al., Automatic emotional speech classification, *Proc. ICASSP2004*, vol. 1, pp.593–596, 2004.

[2] I. Luengo et al., Automatic emotion recognition using prosodic parameters, *Proc. Interspeech2005 - Eurospeech*, pp.493–496, 2005.

[3] L. Vidrascu et al., Detection of real-life emotions in call centers, *Proc. Interspeech2005 - Eurospeech*, pp.1841–1844, 2005.

[4] S. Steidl et al., "of all things the measure is man" - automatic classification of emotions and inter-labeler consistency, *Proc. ICASSP2005*, pp.317–320, 2005.

[5] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995

[6] P. Ekman, An argument for basic emotions, *Cognition and Emotion*, vol. 6(3–4), pp. 169–200, 1992.

[7] H. Schlosberg, Three dimensions of emotion, *Psychological Review*, vol.61(2), pp.81-88, 1954.

[8] J.A. Russell, A circumplex model of affect, *Jounal of Personality and Social Psychology*, vol.39, pp.1161–1178, 1980.

[9] R. Nisimura et al., Public speech-oriented guidance system with adult and child discrimination capability, *Proc. ICASSP2004*, vol.1, pp.433–436, 2004.

[10] R. Nisimura et al., Operating a public spoken guidance system in real environment, *Proc. Interspeech2005 - Eurospeech*, pp.845–848, 2005.

[11] A. de Cheveigne et al., Yin, a fundamental frequency estimator for speech and music, *Journal of the Acoustical Society of America*, vol.111, no.4, pp.1917–1930, 2002.

[12] T. Kudo et al., Fast methods for kernel-based text analysis, *Proc. ACL2003*, pp.24–31, 2003.