Interpretation of User Evaluation for Emotional Speech Synthesis System

Ho-Joon Lee and Jong C. Park

Computer Science Department, KAIST
335 Gwahangno, Yuseong-gu, Daejeon 305-701 Republic of Korea
hojoon@nlp.kaist.ac.kr, park@cs.kaist.ac.kr

Abstract. Whether it is for human-robot interaction or for human-computer interaction, there is a growing need for an emotional speech synthesis system that can provide the required information in a more natural and effective manner. In order to identify and understand the characteristics of basic emotions and their effects, we propose a series of user evaluation experiments on an emotional prosody modification system that can express either perceivable or slightly exaggerated emotions classified into anger, joy, and sadness as an independent module for a general purpose speech synthesis system. In this paper, we propose two experiments to evaluate the emotional prosody modification module according to different types of the initial input speech. And we also provide a supplementary experiment to understand the apparently prosody-independent emotion, or joy, by replacing the re-synthesized joy speech information with original human voice recorded in the emotional state of joy.

Keywords: Emotional Speech Synthesis, User Evaluation, Emotional Prosody Modification, Affective Interaction.

1 Introduction

Speech is understood as the most basic and widely used communication method for the expression of one's thoughts during human-human interactions, and studied also for a user-friendly interface between humans and machines. The recent progress in speech synthesis has produced artificial voice results with very high intelligibility, but the quality of sound and the naturalness of inflection still remain a major issue. Recently, in addition to the need for improvement in sound quality and naturalness, there is a growing need for a method to generate spoken language expressions with appropriate emotions to provide the required information in a more natural and effective manner, as well as for the enhancement of an emotional speech synthesis system for effective human-robot interaction. The related work in the field confirms the common belief that prosody plays a key role for the task [1, 2]. However, during the development of our emotional speech synthesis system [3], we realized that, while there are emotions that can be easily perceived with simplified prosody structures, there are those that are very hard to express with prosody structures alone, even when we provide the most accurate prosody structure.

J.A. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2009, LNCS 5610, pp. 295–303, 2009.

In order to identify and understand the characteristics of these emotions and their effects, we propose in this paper a series of user evaluation experiments on an emotional prosody modification system that can express either perceivable or slightly exaggerated emotions as an independent module for general purpose speech synthesis systems.

2 Emotional Speech Synthesis System

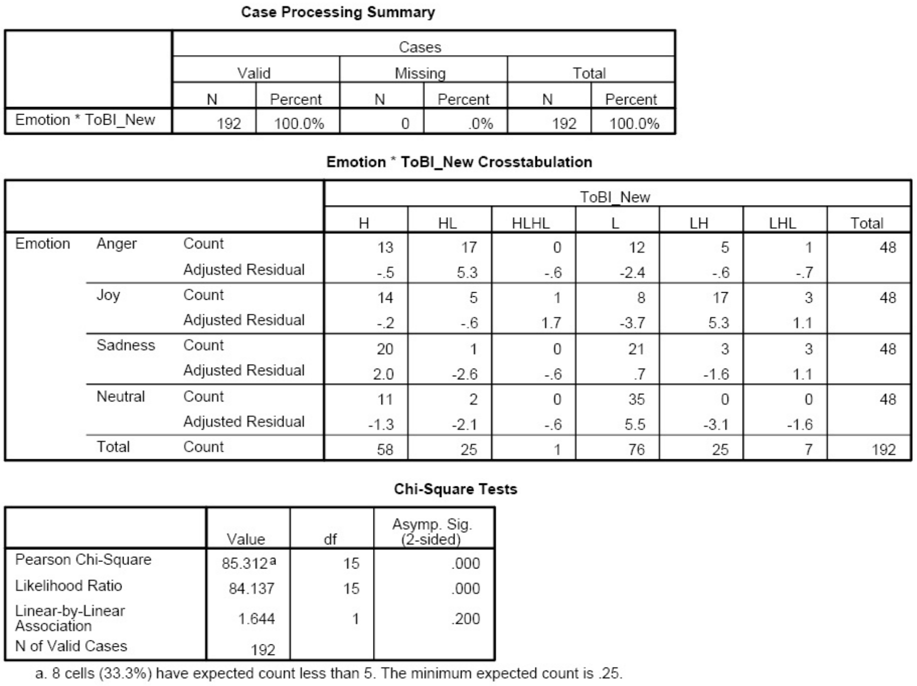
For the analysis of prosody structure through a more precise level of units, we annotated the Korean emotional speech corpus, distributed by the Speech Information Technology & Industry Promotion Center [4], with the K-ToBI labeling system. This speech corpus was recorded by six professional actors and actresses in a sound-proof room, and is composed of emotionally neutral ten sentences with six different emotions (joy, anger, sadness, fear, boredom, and neutral). An AKG C414-B ULS microphone was used with a 16KHz sample rate, and each speech was stored as a 16bit Windows wave format. We used eight sentences spoken by six speakers, as described in Table 1, considering four emotions (joy, anger, sadness, and neutral). The number of Ejeols (words separated by a space) was evenly distributed from 1 to 6.

Table 1. Eight sentences used for prosody structure analysis

Ejeol	Sentence
1	예. (Yes.)
1	아니요. (No.)
2	나도 몰라. (I don’t know either.)
3	야, 이제 그만하자. (See, let’s end it now.)
3	정말 그렇단 말이야. (It really is.)
4	지금 어디 가는 거야? (Where are you going now?)
5	이건 내가 원하던 게 아니야. (This is not what I wanted.)
6	난 가지 말라고 하면서 문을 닫았어. (I shut the door closed asking her not to leave.)

The Korean emotional speech corpus had passed manufacturer’s perception test performed by twenty subjects (eighteen males, two females), and Table 2 below shows the results. Among the emotions, anger turned out to be the most perceivable emotion (94.3%), and fear, the most confusing one (80.3%). However, the overall acceptance rate is more than 80%.

For the analysis of dominant emotional prosody patterns, we annotated eight sentences spoken by six speakers with four emotions, or 192 pieces of speech in total with the K-ToBI labeling system [5]. And for the statistical verification of the K-ToBI labeled data, we performed Pearson’s Chi-square tests. As shown in Fig. 1, the results support the null hypothesis that each emotion has distinct Intonational Phrase (IP) boundary patterns that can distinguish one emotional state from the rest. Then we calculated adjusted residuals to find the distinct pitch contour pattern or patterns. If the calculated value of the adjusted residual is bigger than 2, that feature can be statistically interpreted as the dominant pattern of a certain emotion. Pearson's Chi-square tests and adjusted residual were performed by SPSS software. From the statistical analyses of pitch contour patterns, we were able to find very strong tendencies between anger and HL%, joy and LH%, sadness and H%, and neutral and L%.



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Emotion * ToBI_New	192	100.0%	0	.0%	192	100.0%

Emotion * ToBI_New Crosstabulation

			ToBI_New						
			H	HL	HLHL	L	LH	LHL	Total
Emotion	Anger	Count	13	17	0	12	5	1	48
		Adjusted Residual	-.5	5.3	-.6	-2.4	-.6	-.7	
	Joy	Count	14	5	1	8	17	3	48
		Adjusted Residual	-.2	-.6	1.7	-3.7	5.3	1.1	
	Sadness	Count	20	1	0	21	3	3	48
		Adjusted Residual	2.0	-2.6	-.6	.7	-1.6	1.1	
	Neutral	Count	11	2	0	35	0	0	48
		Adjusted Residual	-1.3	-2.1	-.6	5.5	-3.1	-1.6	
	Total	Count	58	25	1	76	25	7	192

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	85.312a	15	.000
Likelihood Ratio	84.137	15	.000
Linear-by-Linear Association	1.644	1	.200
N of Valid Cases	192		

a. 8 cells (33.3%) have expected count less than 5. The minimum expected count is .25.

Fig. 1. Chi-square test and adjusted residual calculation results

Table 2. Perception test result done by twenty subjects

Speaker	Neutral	Joy	Anger	Sadness	Fear	Boredom
CWJ	89.5	93.5	88.5	85.5	59.0	93.0
KKS	62.5	90.5	92.0	80.5	85.5	82.0
LHJ	83.5	67.5	98.0	84.5	88.5	84.0
MYS	84.5	91.5	90.0	89.5	93.5	81.0
PYH	85.0	95.0	99.0	94.0	61.5	94.5
YSW	95.4	89.5	98.5	89.5	93.5	81.0
Average	83.3	87.9	94.3	87.3	80.3	85.9

To incorporate these analyzed and distinct Intonational Phrase boundary patterns for different emotional states, we propose a prosody-unit-level emotional prosody modifier that produces distinct pitch contour, intensity contour, and speech duration according to the three different emotional states: anger, joy, and sadness. The emotional prosody modifier is a simple, coarse-grained prosody re-synthesis module that consists of a pitch contour mapping function, a pitch exaggeration function, an intensity variation function, and a duration variation function. We set the empirical value of each prosodic parameter based on the previous findings in the literature [1, 2], also taking into account language specific phenomena for Korean including the speaker's gender information, short and long vowel sound disambiguation [6, 7], and prosodic structure of discourse markers [8], captured from various Korean speech corpora.

Equation 1 below shows the algorithm of our pitch contour modification function. This pitch contour modification function generates the base emotional pitch contour of speech including the synthesized results of Text-to-Speech (TTS) systems and recorded human voice for each emotion.

$$y'(t) = y(t)\cdot\left(1+a*\sin\left(b\pi+\frac{t-t_1}{t_2-t_1}*c\pi\right)\right)+d \tag{1}$$

where

t $\in [t_1, t_2]$;

y original pitch value as a function of time t;

y' modified pitch value;

a maximum / minimum pitch range ;

b initial position of pitch contour;

c final position of pitch contour (rising tone: 0.5, rising-falling: 1); and

d declination / ascent level.

After the modification of the base emotional pitch contour, we apply a pitch exaggeration function to characterize the difference in pitch variation according to the difference in emotion types. First, this module detects eight pitch points per unit. Then we exaggerate the difference in each pitch pair by adding 6Hz for joy and anger, and 40Hz for fear and sadness. Next, we adjust the intensity with the intensity contour modification function which is similar to the pitch contour modification function in Equation 1, but much simpler. Then we control the duration of each unit preserving the intrinsic value of f0. All these four modules are implemented in a PRAAT [9] script supporting not only commercial TTS systems, but recorded human voice also. We used the Python language for the interface of PRAAT software and TTS output or human voice, and therefore this module supports both Linux and Windows environments.

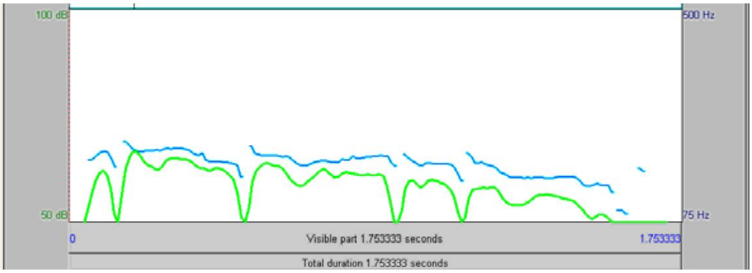


Fig. 2. Pitch and intensity traces of original speech, spoken in a neutral emotional state

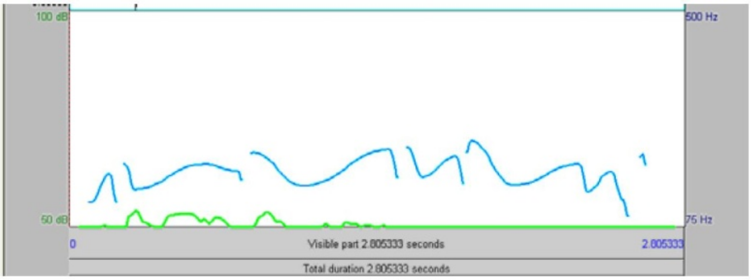


Fig. 3. Pitch and intensity traces of prosody modified speech to a sad emotional state

Fig. 2 shows the prosody trace of a recorded Korean utterance "이건 내가 원하던 게 아니야." which means in English "This is not what I wanted." spoken neutrally by a professional actress, and Fig. 3 shows its modified prosody trace as a sad emotional state produced by our emotional prosody modifier. The blue line (upper line) indicates the pitch contour, and the green line (lower line) the intensity. In Fig. 3, the entire duration is lengthened from 1.753 seconds to 2.805 seconds without any side effect such as f0 contour lowering. Pitch contour is spread more widely, and intensity is weakened.

3 Evaluation of Emotional Speech Synthesis System

For the identification and understanding of the characteristics of three basic emotions and their effects, we prepared three stages of experiments. The first and second experiments are designed to evaluate the emotional prosody modifier according to different types of the initial input speech, such as monotonous-prosody speech and excited-prosody speech. The supplementary experiment is performed to identify apparently prosody-independent speech. The subjects of these three experiments are fourteen kindergarten teachers, twelve of them females and two males. They are 29.6 years old on average. We did not carry out any prior training for the fourteen subjects, and answers were not notified to the subjects after the experiments.

At the beginning of the experiments, subjects were asked to choose one most likely emotion among anger, joy, sadness, and neutral. We used five semantically neutral sentences as show in Table 3. For the first experiment, five neutrally recorded speech files were used as a monotonous input speech, and the emotional prosody modifier produced fifteen results with three emotional states. The test sequences of first and second experiments were randomly organized.

Table 3. Input sentences for the evaluation of emotional prosody modifier

Sentence
야, 이제 그만하자. (See, let's end it now.)
정말 그렇단 말이야. (It really is.)
지금 어디 가는 거야? (Where are you going now?)
이건 내가 원하던 게 아니야. (This is not what I wanted.)
난 가지 말라고 하면서 문을 닫았어. (I shut the door closed asking her not to leave.)

Table 4 shows the evaluation results of the emotional prosody modification with monotonous input speech. From the analysis of the results of the first experiment, we find that anger is very sensitive to emotional prosody structure (80% of perception rate). And sadness also shows a strong relationship with prosody structure. It is rather surprising to note that none of the subjects perceived joy from the monotonous input speech, even though we modified the prosody structure of joy based on the analyses of real speech, exactly as we did for anger and sadness.

Table 4. Evaluation result for monotonous input speech

	Anger	Joy	Neutral	Sadness	Total
Anger	**56** **(80.0%)**	3 (4.3%)	6 (8.6%)	5 (7.1%)	70
Joy	12 (17.1%)	**0** **(0%)**	16 (22.9%)	42 (60.0%)	70
Sadness	4 (5.7%)	2 (2.9%)	23 (32.9%)	**41** **(58.6%)**	70

For the second experiment, we used five pieces of excited voice as the input for the emotional prosody modifier, and generated fifteen randomly organized test sets. Table 5 indicates the results of the second perception experiment.

Table 5. Evaluation result for excited input speech

	Anger	Joy	Neutral	Sadness	Total
Anger	**56** **(80.0%)**	7 (10.0%)	6 (8.6%)	1 (1.4%)	70
Joy	18 (25.7%)	**15** **(21.4%)**	15 (21.4%)	22 (31.4%)	70
Sadness	3 (4.3%)	38 (54.3%)	18 (25.7%)	**11** **(15.7%)**	70

Interestingly, anger preserved prosody sensitivity when the type of input was changed from monotonous-prosody speech to excited-prosody speech. From the second experiment, two major changes were observed: an increase in the perception rate of joy, and a decrease in the perception rate of sadness. The decrease in the perception rate of sadness can be caused by the sudden change of the test environment. In order to indentify the cause of this sudden change, we proposed the third experiment. However, the expected response of the perception rate of joy was still very weak.

To identify the characteristics of the emotional prosody structure of joy, and to validate the hypothesis above on a sudden change of sadness, we performed the third experiment with the same subjects and in the same sequence as the second experiment. The only difference between the second and third experiments was just the replacement of the modified joy speech with the original human voice recordings in the emotional state of joy, which had passed the manufacturer's perception test at the rate of 91.5%.

Table 6. Evaluation result for repeated test with human voice recordings

	Anger	Joy	Neutral	Sadness	Total
Anger	**58** **(82.9%)**	7 (10.0%)	4 (5.7%)	1 (1.4%)	70
Joy	32 (45.7%)	**12** **(17.1%)**	15 (21.4%)	11 (15.7%)	70
Sadness	10 (14.3%)	18 (25.7%)	19 (27.1%)	**23** **(32.9%)**	70

After the third perception test, we made three interesting interpretations from the results shown in Table 6. First, the same sequence in the repeated experiment did not seem to influence the perception rate of anger. There was only a slight movement from neutral to anger. This allows us to define anger as a primarily prosody-sensitive emotion.

Second, we found that some part of the decreased perception rate was due to the sudden change of the test environment. So it is a possible interpretation that there was a confusion of sadness in the second experiment. Despite the result of the second experiment, it appears that sadness is also a prosody-sensitive emotion.

Third and most important, we could not find any meaningful relationship between the prosody structure and the emotion of joy, even though we used real voice which had passed the manufacturer's perception test at the rate of 91.5%. This leads us to conclude that joy is not a prosody sensitive emotion, which forces us to find other, effective approaches to express the emotion of joy through an emotional spoken language generation system.

4 Discussion

For the accurate understanding of each evaluation result, a quantitative comparison method that can also describe the influence of wrong answers is called for. For example, the perception rate of the first experiment related to anger is just equal to that of the second experiment. But for the same category, it is very hard to figure out the influence of errors such as joy and sadness. For this kind of interpretation including error analysis, we suggest a Euclidean distance based quantitative comparison method. Fig. 4 describes a Euclidean distance model of tetrahedron designed for the analysis of four types of category.

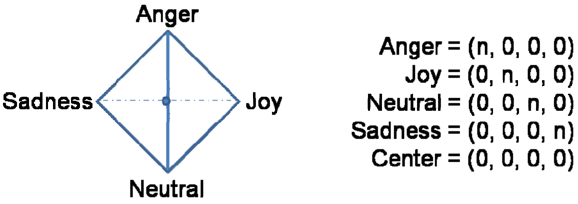


Fig. 4. Euclidean distance model for tetrahedron

From this point of view, we can calculate and compare each distance described in Table 4, Table 5, and Table 6. When the size of *n* is 70, the maximum distance of each category is approximately 98.99, and the minimum distance is 0.

Table 7. Euclidean distance of Table 4

	Anger	Joy	Neutral	Sadness
Anger	**16.31**	87.67	85.24	86.06
Joy	73.38	**84.05**	69.46	34.41
Sadness	81.06	82.76	62.53	**37.28**

Table 8. Euclidean distance of Table 5

	Anger	Joy	Neutral	Sadness
Anger	**16.79**	84.51	85.33	89.34
Joy	60.32	**63.70**	63.70	55.48
Sadness	79.86	38.44	65.41	**72.51**

Considering both correct answers and errors, we conclude that synthesized anger based on the monotonous input speech is slightly closer to the position of anger than that based on the excited speech, even though they have the same perception rate. And for the synthesis of anger, the change of initial input speech from monotonous to excited one increases the distance of joy by 3.16, but decreases the distance of neutral by 0.09 and sadness by 3.28.

5 Conclusion

In this paper, we proposed an emotional prosody modification system, and evaluated the performance of the system, in order to find a relationship between prosody structures and emotions.

First, we proposed a prosody-unit-level emotional prosody modification system that produces distinct pitch contour, intensity contour, and speech duration according to three different emotional states: anger, joy, and sadness.

And during the evaluation process, anger and sadness were identified as prosody sensitive emotions, whereas joy was not. Consequently, this difference led us to discover the possibilities and limitations of prosody modification for the generation of emotional spoken language expression systematically.

Further analyses of emotional speech data are necessary, taking into account various speakers, speaking environment, and speaking styles. And more organized evaluation and interpretation strategies are essentially needed for further work.

Acknowledgments. This research was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs, funded by the Korea Ministry of Knowledge Economy.

References

1. Schröder, M.: Emotional Speech Synthesis: A Review. In: Eurospeech 2001, vol. 1, pp. 561–564 (2001)
2. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18(1), 32–80 (2001)
3. Lee, H.-J., Park, J.C.: Customized Message Generation and Speech Synthesis in Response to Characteristic Behavioral Patterns of Children. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4552, pp. 114–123. Springer, Heidelberg (2007)
4. SiTEC Emotional Speech Corpus, `http://www.sitec.or.kr/English/index.asp`
5. Jun, S.-A.: K-ToBI (Korean ToBI) Labeling Convention. Korean Journal of Speech Science 7 (2000)
6. Lee, H.-J., Park, J.C.: Lexical Disambiguation for Intonation Synthesis: A CCG Approach. In: Korean Society for Language and Information, pp. 103–118 (2005)
7. Lee, H.-J., Park, J.C.: Vowel Sound Disambiguation for Proper Intonation Synthesis. In: 19th Pacific Asia Conference on Language, Information and Computation, pp. 131–142 (2005)
8. Lee, H.-J., Park, J.C.: Characteristics of Spoken Discourse Markers and their Application to Speech Synthesis Systems. In: 19th Annual Conference on Human and Cognitive Language Technology, pp. 254–260 (2007)
9. PRAAT, `http://www.praat.org`