An Analysis of the Effect of Emotional Speech Synthesis on Non-Task-Oriented Dialogue System

Yuya Chiba, Takashi Nose, Mai Yamanaka, Taketo Kase, Akinori Ito
Graduate School of Engineering,
Tohoku University, Japan

Abstract

This paper explores the effect of emotional speech synthesis on a spoken dialogue system when the dialogue is non-task-oriented. Although the use of emotional speech responses has been shown to be effective in a limited domain, e.g., scenario-based and counseling dialogue, the effect is still not clear in the non-task-oriented dialogue such as voice chat. For this purpose, we constructed a simple dialogue system with example- and rule-based dialogue management. In the system, two types of emotion labeling with emotion estimation are adopted, i.e., system-driven and user-cooperative emotion labeling. We conducted a dialogue experiment where subjects evaluate the subjective quality of the system and the dialogue from multiple aspects such as richness of the dialogue and impression of the agent. We then analyze and discuss the results and show the advantage of using appropriate emotions for expressive speech responses in the non-task-oriented system.

1 Introduction

Recently, spoken dialogue systems have been becoming popular in various applications, such as a speech assistant system in smartphones and smart speakers, an information guide system in public places, and humanoid robots. There have been a variety of studies for developing spoken dialogue systems, and the systems are roughly grouped into two categories, task-oriented and non-task-oriented systems, from the aspect of having a goal or not in the dialogue. Although the task-oriented dialogue systems (Zue et al., 2000; Kawanami et al., 2007) are important as practical applications, e.g., ticket vending and information guidance, the role of the non-task-oriented systems is increasing for more advanced human-computer interaction (HCI) including voice chat.

There have been many studies related to the non-task-oriented dialogue systems. Nakano et al. (2006) tried to incorporate both task-oriented and non-task-oriented dialogue functions into a humanoid robot using a multi-expert model. Dybala et al. (2010) proposed an evaluation method of subjective features of human-computer interaction using chatbots. Yu et al. (2016) proposed a set of conversational strategies to handle possible system breakdowns. Although these studies enhance the performance of the dialogue systems, an important role is still missing from the viewpoint of the system expressivity. Specifically, the system cannot perceive and express para-linguistic information such as emotions, which is completely different from our daily communication.

Several studies have been presented where emotions were taken into consideration in spoken dialogue systems. MMDAgent (Lee et al., 2013) is a well-known open-source dialogue system toolkit where emotional speech synthesis based on hidden Markov models (HMMs) (Yoshimura et al., 1999) is incorporated and style modeling and style interpolation techniques can be used for providing expressive speech (Nose and Kobayashi, 2011). Su et al. (2014) have combined situation and emotion detection with a spoken dialogue system for health care to provide more warming feedback of the system. Kase et al. (2015) developed a scenario-based dialogue system where emotion estimation and emotional speech synthesis were incorporated. However, the use of emotional speech synthesis was not investigated in a non-task-oriented dialogue system, and the effect of the emotions on the dialogue is still unclear.

In this study, we develop a Japanese simple non-task-oriented expressive dialogue system with text-based emotion detection and emotional speech synthesis. We then conduct a dialogue experiment in which participants chat with the system and evaluate the performance in terms of multiple subjective measures such as richness and pleasantness of the conversation and analyze the result. We also examine the change of the pitch variation of the users in the dialogue to investigate the acoustic effect of the system expressivity on the utterance of the users.

2 Overview of the Dialogue System

Figure 1 shows the flow of the dialogue system constructed for the experiment in Section 5. The speech input is decoded to the text using a speech recog-

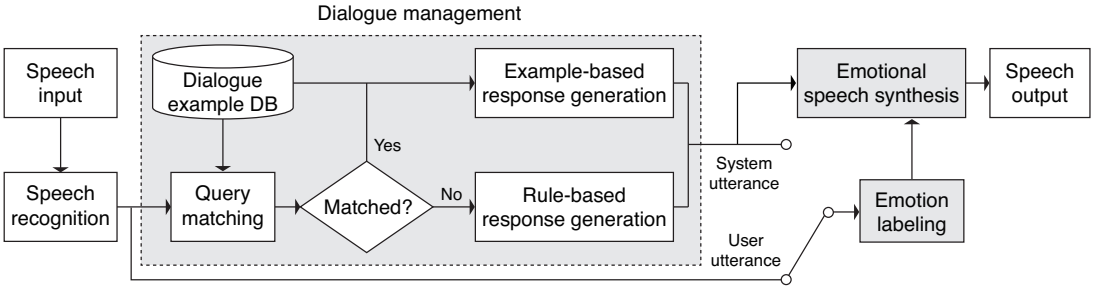

Figure 1: Overview of our non-task-oriented dialogue system with system-driven/user-cooperative emotional speech synthesis. The system or the user utterance is used alternatively for the emotion labeling in the case of (a) the system-driven or (b) the user-cooperative systems, respectively.

Table 1: Dialogue example between a user and the system. U and S denote user and system, respectively, and E and R denote the types of the response generation, i.e., example- and rule-based, respectively.

U	Hi!	
S	Hello! Let's talk about ball games.	E
U	All right!	
S	Is the basketball exciting?	R
U	I like watching basketball.	
S	That's good, me too.	R
U	Do you like baseball?	
S	I don't like baseball.	R
U	Why?	
S	I've never thought the reason.	E
U	I see.	

nizer, Julius (Lee and Kawahara, 2009). In the dialogue management part, system responses are generated by combining example-based and typical rule-based (Weizenbaum, 1966) response generation methods. First, query matching for the example-based response generation is applied to the text using a dialogue example database that is constructed in advance. Specifically, the decoded text is converted to a vector using a bag of words, and cosine similarity is calculated between the text and the questions in the database. If the similarity score is larger than or equal to a predetermined threshold, the answer corresponding to the question having highest similarity is adopted as the system utterance. Otherwise, the system utterance is generated by applying the prepared rules to the decoded text, i.e., the user utterance. For the rule-based response generation, nouns (e.g., baseball, pasta) and subjective words (e.g., like, dislike, happy) are extracted from the user utterance and are used for the response generation based on the rules. Table 1 shows an example of the dialogue between a user and the system where the system responses are generated using both example- and rule-based methods.

After the response generation, emotion estimation, in other words, emotion labeling, is performed using either the system or the user utterance to choose the emotion to be used in the succeeding speech synthesis. We call the emotion labeling with the system and the user utterances "system-driven" and "user-cooperative" labeling hereafter, which was also discussed in the previous study on scenario-based dialogue (Kase et al., 2015). Finally, emotional speech synthesis based on HMMs is performed using the emotion label and the corresponding emotion-dependent acoustic model trained in advance. The details of the emotion estimation and the emotional speech synthesis are described in Sections 3 and 4, respectively.

3 Emotion Labeling Using System or User Utterance

In both system-driven and user-cooperative emotion labeling, the emotion category is estimated from the content of the text (Guinn and Hubal, 2003), i.e., the system or the user utterance in Figure 1, which was previously used in (Kase et al., 2015). Basically, the estimation of emotion category is based on matching between words in a sentence and a database of emotional expression words. Two data sources are exploited, one is an evaluation polarity dictionary of verbs (Kobayashi et al., 2004), and the other is a sentiment polarity dictionary of nouns (Takase et al., 2013), both are for Japanese words. The expressions and words in those dictionaries have either positive or negative polarity. Thus, if a sentence has a word or an expression (a phrase) with positive or negative polarity, we give the sentence "happy" or "sad" emotion, respectively. If no such words and phrases are found, we give a "neutral" emotion label. Several rules are employed for complicated situation in the expression matching, as follows.

1. If the emotional expression in the database is a phrase, the phrase is adopted only when all words of the phrase coincide with the text.

2. If two or more expressions are matched, the last expression is adopted.

3. If a negative expression is found such as "not (nai in Japanese) " after the match, we reverse

the polarity. Note that the negative expressions in Japanese succeed the modified word, e.g., "tanoshiku nai (happy not)" means unhappy.

4 Emotional Speech Synthesis

In this study, we use emotional speech synthesis based on HMMs which are widely used in the various research fields. The choice is mainly because of the computation cost in speech synthesis. The computation cost of HMM-based speech synthesis is relatively low compared to the other existing synthesis methods such as synthesis techniques based on unit selection (Hunt and Black, 1996) and deep neural networks (Zen et al., 2013). The low computation cost is essential to achieve the spoken dialogue system with smooth interaction between the system and users. In addition, a variety of expressive speech synthesis techniques have been proposed in the HMM-based speech synthesis (Nose and Kobayashi, 2011), which will enrich the dialogue system also in the future work.

In the HMM-based speech synthesis, speech samples are modeled by the sequences of context-dependent phone HMMs. Phonetic and prosodic contextual factors are used for the context. In the model training, the HMM parameters are tied using state-based context clustering with decision trees for each acoustic features, i.e., spectral, excitation, and duration features. The HMMs are then optimized using the EM algorithm. In this study, we adopted style-dependent modeling (Yamagishi et al., 2005) for the emotional speech synthesis. In the synthesis phase, the input text is converted to a context-dependent label sequence using text analysis, and the corresponding HMMs are concatenated to create a sentence HMM. Finally, the speech parameters are generated from the sentence HMM using speech parameter generation algorithm (Tokuda et al., 1995), and a waveform is synthesized using a vocoder.

5 Dialogue Experiment

We conducted a dialogue experiment using several systems to confirm and investigate the effect of emotional speech synthesis on the non-task-oriented dialogue system.

5.1 Experimental Procedure and Conditions

Ten subjects participated in the dialogue experiment and evaluated the subjective quality of the system and the dialogue. Each subject conducted a dialogue whose topic was "ball game" twice. The duration was about 60 to 90 seconds in each dialogue. We constructed the following four systems where different emotion labeling was adopted.

Baseline No emotion labeling (neutral)

System System-driven emotion labeling

User User-cooperative emotion labeling

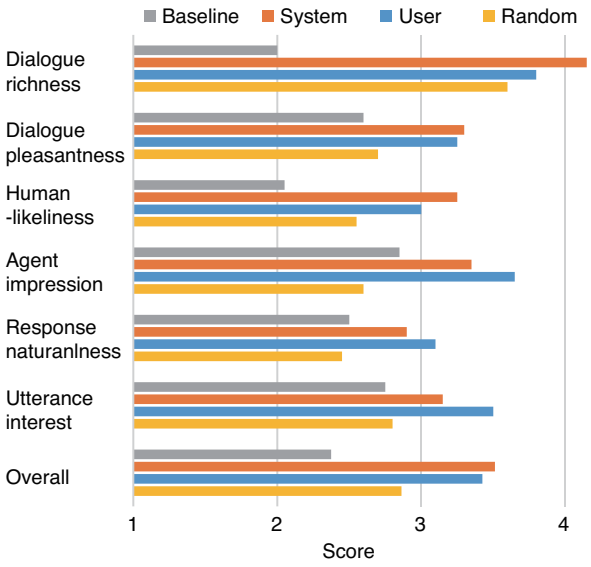

Figure 2: Average subjective scores of the participants for non-task-oriented dialogue to the four systems.

Random Random emotion labeling

Participants sat on a chair in a soundproof room and conducted a dialogue with an agent in a laptop PC. The visual of the agent was 10 cm high and 4 cm wide, and only lip-sync was implemented with no facial expressions and motions. After the dialogue, participants were asked about 1) richness of the dialogue, 2) pleasantness of the dialogue, 3) human-likeliness of the agent, 4) impression of the agent, 5) naturalness of the response, and 6) interest in the response. The rating score is 1 for the lowest and 5 for the highest.

For the emotional speech synthesis, we used emotional speech data of a professional female narrator who uttered 503 phonetically balanced Japanese sentences with neutral, joyful, and sad emotional expressions. The other basic conditions of the training and synthesis were the same as the previous study (Yamagishi et al., 2005).

5.2 Results and Discussions

Figure 2 shows the average scores of the subjective rating for the four systems. From the results, we first found a clear increase of the overall scores in the cases of using system-driven and user-cooperative emotion labeling compared to the baseline (no emotions) and random emotion labeling. This result indicates that the use of appropriate emotions in the synthetic speech response improves the subjective performance also for the non-task-oriented dialogue system.

Next, we conducted one-way ANOVA for the four systems, where emotion labeling methods was a factor. We found significant differences at a 5% level for the richness of the dialogue ($p < 0.001$), pleasantness of the dialogue ($p = 0.025$), human-likeliness of the agent ($p = 0.005$), and impression of the agent ($p = 0.001$).

Table 2: p-values of the multiple comparison test by t-test with Bonferroni correction. The results with a significant difference at 5% level are in a bold font.

Richness of the dialogue

	System	User	Random
Baseline	**<0.001**	**<0.001**	**<0.001**
System		>1.000	0.239
User			>1.000

Pleasantness of the dialogue

	System	User	Random
Baseline	0.094	0.147	>1.000
System		>1.000	0.224
User			0.335

Human likeliness of the agent

	System	User	Random
Baseline	**0.006**	0.050	0.951
System		>1.000	0.298
User			>1.000

Impression of the agent

	System	User	Random
Baseline	0.482	**0.035**	>1.000
System		>1.000	0.057
User			**0.002**

From these results, we verified that the type of the emotion labeling method actually affected the impression of subjects to the agent and conversation. In contrast, there are no significant differences in the naturalness of the response ($p = 0.242$) and interest in the response ($p = 0.062$) from the result of the one-way ANOVA. We then conducted a multiple comparison test by t-test with Bonferroni correction. Table 2 shows the p-values of the test.

In the rating of the richness of the dialogue, the three systems with emotions gave higher scores than the baseline. This result indicates that the richness is related to the variation of the emotions of the synthetic speech responses. Although there is no significant difference in the pleasantness of the dialogue, several scores had the same tendency as the previous study (Kase et al., 2015) in which the systems with the emotion labeling based on emotion estimation gave higher scores than the other systems. On the other hand, in human-likeliness of the agent, the tendency was different from the result in (Kase et al., 2015), and the system-driven labeling gave the highest score. A possible reason for this mismatch is that dialogue breakdown can occur in the non-task-oriented dialogue differently from the scenario-based one. About the impression of the agent, the user-cooperative system gave a better score than the baseline and the random labeling systems. Users tend to prefer the system that understands the users' emotional state and sympathizes with them.

5.3 Prosodic Analysis of User Utterances

In the dialogue experiment, we recorded the user utterances with 16 kHz sampling and 16 bit quantization.

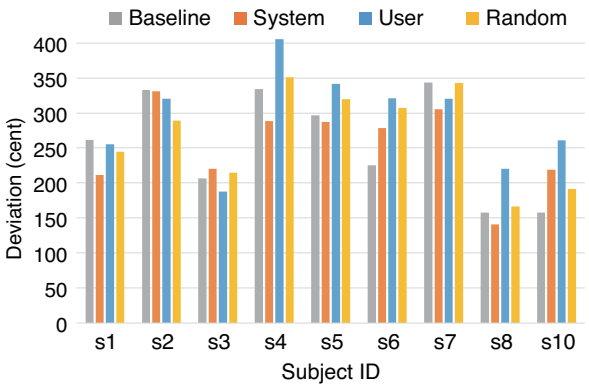

Figure 3: Mean values of the F0 deviations in each utterance for respective subjects, s1 to s10, (except s9).

The utterances of one subject (s9) had a problem in the recording, and hence we analyzed the 511 utterances of nine subjects. In this study, we focused on the fundamental frequency (F0) which is known to be the most important speech parameter for emotional expression. F0s were extracted using the SWIPE algorithm (Camacho, 2007) with 10-ms frame shift.

We calculated the deviations of F0s for the utterances of respective subjects in each system. Figure 3 shows the mean values of the deviations for each subject. We conducted one-way ANOVA where the labeling method was a factor. Although we expected that the emotional speech responses in a non-task-oriented dialogue more affect the user utterances than that in the scenario-based dialogue, there was no significant difference ($p = 0.613$) between the systems. One possible reason is that the naturalness of the system response is still insufficient to draw out emotions of the users.

6 Conclusions

In this paper, we discussed the effect of emotional speech synthesis on the non-task-oriented spoken dialogue system. We constructed dialogue systems with system-driven and user-cooperative emotion labeling and compared the subjective performance with the systems with no emotion and random emotion labeling. Experimental results showed that the use of emotional speech responses clearly improves the subjective scores such as richness of the dialogue and impression of the agent even when the dialogue is non-task-oriented. Improving the performance of the emotion estimation using both system and user utterances is our future work. The use of the acoustic information in the emotion estimation is also a remaining issue.

Acknowledgments

Part of this work was supported by JSPS KAKENHI Grant Numbers JP15H02720, JP16K13253, and JP17H00823.

References

Arturo Camacho. 2007. *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*. Ph.D. thesis, University of Florida Gainesville.

Pawel Dybala, Michal Ptaszynski, Rafal Rzepka, and Kenji Araki. 2010. Evaluating subjective aspects of HCI on an example of a non-task oriented conversational system. *International Journal on Artificial Intelligence Tools*, 19(06):819–856.

Curry Guinn and Rob Hubal. 2003. Extracting emotional information from the text of spoken dialog. In *Proceedings of the 9th International Conference on User Modeling*.

Andrew J Hunt and Alan W Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP*, volume 1, pages 373–376.

Taketo Kase, Takashi Nose, and Akinori Ito. 2015. On appropriateness and estimation of the emotion of synthesized response speech in a spoken dialogue system. In *Proceedings of the International Conference on Human-Computer Interaction*, pages 747–752.

Hiromichi Kawanami, Hiroshi Saruwatari, Kiyohiro Shikano, et al. 2007. Development and portability of ASR and Q&A modules for real-environment speech-oriented guidance systems. In *Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 520–525.

Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. 2004. Collecting evaluative expressions for opinion extraction. In *Proceedings of the International Conference on Natural Language Processing*, pages 596–605.

Akinobu Lee and Tatsuya Kawahara. 2009. Recent development of open-source speech recognition engine julius. In *Proc. APSIPA ASC*, pages 131–137.

Akinobu Lee, Keiichiro Oura, and Keiichi Tokuda. 2013. MMDAgent—A fully open-source toolkit for voice interaction systems. In *Proc. ICASSP*, pages 8382–8385.

Mikio Nakano, Atsushi Hoshino, Johane Takeuchi, Yuji Hasegawa, Toyotaka Torii, Kazuhiro Nakadai, Kazuhiko Kato, and Hiroshi Tsujino. 2006. A robot that can engage in both task-oriented and non-task-oriented dialogues. In *IEEE-RAS International Conference on Humanoid Robots*, pages 404–411.

Takashi Nose and Takao Kobayashi. 2011. Recent development of HMM-based expressive speech synthesis and its applications. In *Proc. APSIPA ASC*, pages 1–4.

Bo-Hao Su, Ping-Wen Fu, Po-Chuan Lin, Po-Yi Shih, Yuh-Chung Lin, Jhing-Fa Wang, and An-Chao Tsai. 2014. A spoken dialogue system with situation and emotion detection based on anthropomorphic learning for warming healthcare d. In *Proceedings of the IEEE International Conference on Orange Technologies*, pages 133–136.

Sho Takase, Akiko Murakami, Miki Enoki, Naoaki Okazaki, and Kentaro Inui. 2013. Detecting chronic critics based on sentiment polarity and users behavior in social media. In *Proceedings of the Student Research Workshop in 51st Annual Meeting of the Association for Computational Linguistics*, pages 110–116.

Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. 1995. Speech parameter generation from HMM using dynamic features. In *Proc. ICASSP*, pages 660–663.

Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Junichi Yamagishi, Koji Onishi, Takashi Masuko, and Takao Kobayashi. 2005. Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. *IEICE Trans. Inf. Syst.*, E88-D(3):503–509.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. Eurospeech*, pages 2347–2350.

Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. 2016. Strategy and policy learning for non-task-oriented conversational systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 404–412.

Heiga Zen, Andrew Senior, and Mike Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *Proc. ICASSP*, pages 7962–7966.

Victor Zue, Stephanie Seneff, James R Glass, Joseph Polifroni, Christine Pao, Timothy J Hazen, and Lee Hetherington. 2000. JUPITER: a telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1):85–96.