INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Evaluating Intention Communication by TTS using Explicit Definitions of Illocutionary Act Performance

Nobukatsu Hojo[†], *Noboru Miyazaki*

NTT Communication Science Laboratories
[†]nobukatsu.hojo.cd@hco.ntt.co.jp

Abstract

Text-to-speech (TTS) synthesis systems have been evaluated with respect to attributes such as quality, naturalness and intelligibility. However, an evaluation protocol with respect to communication of intentions has not yet been established. Evaluating this sometimes produce unreliable results because participants can misinterpret definitions of intentions. This misinterpretation is caused by the colloquial and implicit description of intentions. To address this problem, this work explicitly defines each intention following theoretical definitions, "felicity conditions", in speech-act theory. We define the communication of each intention with one to four necessary and sufficient conditions to be satisfied. In listening tests, participants rated whether each condition was satisfied or not. We compared the proposed protocol with the conventional baseline using four different voice conditions; neutral TTS, conversational TTS w/ and w/o intention inputs, and recorded speech. The experimental results with 10 participants showed that the proposed protocol produced smaller within-group variation and larger between-group variation. These results indicate that the proposed protocol can be used to evaluate intention communication with higher inter-rater reliability and sensitivity.

Index Terms: TTS evaluation, spoken dialogue systems, speech synthesis, dialogue act, felicity conditions

1. Introduction

In the pragmatics field, e.g. works by Grice [1] and speech-act theory [2, 3], the conveyance of speakers' intentions to hearers is considered to play a central role in communication. It is reasonable to assume that communication will likely fail if hearers infer different intentions, e.g., "CRITICISM", from what speakers are trying to convey, e.g., "APOLOGY", on the basis of an utterance, e.g., "excuse me.". For human-machine interaction, it is also considered essential for a system to generate utterances in a way that users will recognize and identify the system's exact intentions. Since intention recognition is affected by prosody [4, 5] as well as utterance surfaces and conversational contexts [6], text-to-speech (TTS) synthesis systems play a role in conveying a system's intentions. These demonstrate the necessity to evaluate TTS with respect to accurately conveying intentions ("intention communication" in this work).

The most commonly used evaluation protocols for TTS are listening tests with respect to quality, naturalness, intelligibility, similarity and expressiveness [7, 8, 9, 10]. Application-dependent measures are also used, such as those for audiobook reading [11] and spoken dialogue systems or human-robot interaction [12, 13, 14, 15, 16, 17, 18]. For example, Su et al. showed that emotional TTS for healthcare systems can be used to enable systems to provide warmer feedback of the system [16]. Chiba et al. applied emotional TTS to non-task oriented spoken dialogue systems and showed an improvement in richness of the dialogue and impression of the agent [18]. However, an evaluation protocol with respect to intention communication has not yet been established.

A possible candidate for an evaluation protocol is "speech-act rating" [5]. In this test, participants are asked, "How does the sample sound like the target intention (e.g. CRITICISM)?", given a description of the definition of the target intention. In general, participants need to comprehend precisely what is being asked in questions in order for reliable and valid evaluation results to be obtained. However, this is difficult because descriptions of intentions are colloquial and implicit, and they leave room for misinterpretation. For example, interpretation about what is and what is not "CRITICISM" can be inconsistent among participants and researchers. This problem may not sound serious when a limited number of categories [4, 5] or low-dimensional representation [10] are of interest. However, it becomes severe when using larger number of intentions for evaluation because it will be harder to precisely comprehend their difference. Addressing this problem leads to establishing a general protocol that can be used for a wide range of utterances and intentions.

To address this problem, we propose providing participants with explicit definitions of intentions. Since, for several intentions, their detailed definitions have already been formulated as "felicity conditions" in speech-act theory [3], we extend this formulation to all of the intentions we evaluate in our experiments. We expect and investigate two possible benefits of our proposal. One is a decrease in within-group variation (i.e. between-participants variation). Explicit definitions can suppress individual difference in interpreting questions. The other is an increase in between-group variation. Diffidence regarding question interpretation often results in neutral and less informative ratings. We expect that explicit definitions of intentions can prevent this. These changes in variance are beneficial since they will decrease the number of participants needed for a statistically significant result. The experiments in Sec. 3 and Sec. 4 investigate the changes in variance of ratings.

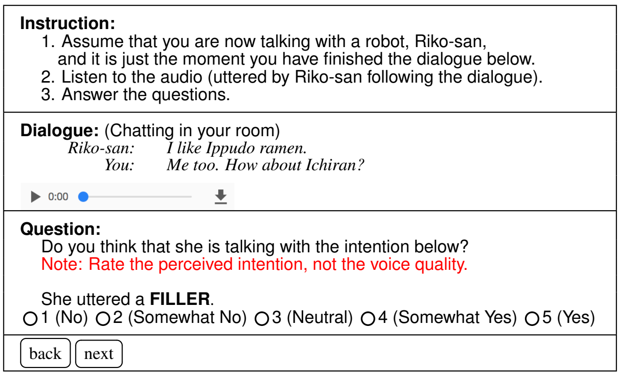
2. Methodology

2.1. Baseline

In this section, we first define the baseline evaluation protocol referring to the previous research [5]. There are two experiments in [5] for evaluating intention communication by speech. One is an alternative forced choice (AFC) categorization test where participants are forced to choose the perceived intention from a given set of choices. The other is a speech-act rating test where participants rate the extent to which a sample sounds like the reference intention. In this study, we define the baseline protocol on the basis of the speech-act rating test. We believe that the AFC test will be infeasible when using a large intention set because choosing an answer from a lot of choices is difficult.

While the conventional speech-act rating test evaluates context-free speech samples, our baseline displays context for

http://dx.doi.org/10.21437/Interspeech.2019-2188



Instruction:

1. Assume that you are now talking with a robot, Riko-san, and it is just the moment you have finished the dialogue below.
2. Listen to the audio (uttered by Riko-san following the dialogue).
3. Answer the questions.

Dialogue: (Chatting in your room)

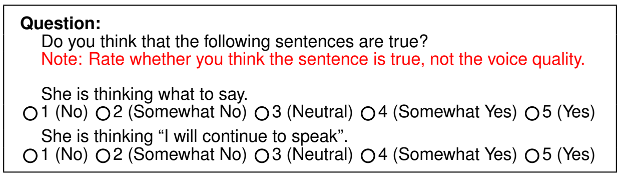
Riko-san: *I like Ippudo ramen.*
You: *Me too. How about Ichiran?*

Question:

Do you think that she is talking with the intention below?
Note: Rate the perceived intention, not the voice quality.

She uttered a **FILLER**.
○1 (No) ○2 (Somewhat No) ○3 (Neutral) ○4 (Somewhat Yes) ○5 (Yes)

Figure 1: *Example of GUI for baseline protocol. Text was translated from Japanese by authors.*



Question:

Do you think that the following sentences are true?
Note: Rate whether you think the sentence is true, not the voice quality.

She is thinking what to say.
○1 (No) ○2 (Somewhat No) ○3 (Neutral) ○4 (Somewhat Yes) ○5 (Yes)

She is thinking “I will continue to speak”.
○1 (No) ○2 (Somewhat No) ○3 (Neutral) ○4 (Somewhat Yes) ○5 (Yes)

Figure 2: *Example of GUI of questionnaires for the proposed protocol. Same instructions as the baseline are used. Text was translated from Japanese by authors.*

each rating. This is because inference for some intentions, such as ACKNOWLEDGEMENT, REPEAT and PARAPHRASE, is strongly dependent on the context and they cannot be properly evaluated by context-free speech samples. Fig. 1 illustrates the GUI for the baseline. The participants are first instructed to assume that they are talking with a robot, and it is the moment that they have just finished the dialogue displayed on a screen. Then, they listen to an audio file, assuming that it is uttered by the robot following the displayed dialogue. Last, they answer the questions.

2.2. Questionnaire with explicit definitions of intentions

In an informal and preliminary test conducted using the baseline protocol, participants commented that they were not confident with their ratings because they could not understand the definitions of intentions well. This is problematic because the obtained ratings will be unreliable. This is why we modified the questionnaire.

The proposed protocol explicitly defines each intention following theoretical definitions, “felicity conditions”, in speech-act theory. Speech-act theory defines an intentional action performed by saying something as an illocutionary act [19]. Necessary and sufficient conditions for performing an illocutionary act are called “felicity conditions” [3]. For example, the felicity conditions for performing PROMISE are given as follows;

- A is a future action by S.
- S believes H wants A to be done/S is able to do A/S has not already done A/H will benefit from A.
- S is willing to do A.
- The utterance counts as an undertaking to do A.

where S, H and A denote a speaker, a hearer and an action to be promised, respectively. We consider an intention to be conveyed and an action to be performed correctly when a hearer judges that all of the conditions are satisfied.

The proposed protocol first defines felicity conditions for each intention to be evaluated (examples in our experiments are illustrated in Tab. 1.). Then, participants are asked to rate on a 5-point MOS scale whether they think each of the conditions of a reference intention is satisfied. A rating for a sample is determined as a minimum of ratings given for all felicity conditions of the reference intention. Since some felicity conditions are trivial given a certain conversational context, we exclude them from the questionnaire entries. To make questions comprehensible, direct speech is used for the questions about the speaker’s beliefs and emotions, e.g., “She is thinking ’I will continue to speak.”’ The variables in the felicity conditions, e.g., “S”, “H” and “P”, are substituted with a concrete propositional content, considering the conversational context, e.g., “S” = “She”, “H” = “you” and “P” = “that tomorrow is rainy”. Fig. 2 illustrates the GUI of the proposed questionnaire for evaluating FILLER. We use the same GUI elements as the baseline other than questionnaires.

3. Experimental setup

3.1. The intention set

As the intention set for evaluation, we used dialogue acts (DAs) proposed in [20]. A dialogue act is an abstract expression of a speaker’s intention [21]. The set we used was designed to cover a wide range of utterances of non-task oriented open-domain conversation. The set consists of 33 DAs. We selected 19 DAs for the test set. We excluded the remaining 14 DAs from the test set because their definitions were similar to any one of those in the test set. For example, SELF-DISCLOSURE_FACTS, EXPERIENCE and HABIT were excluded because their definitions were the same as SELF-DISCLOSURE_PLAN except for the propositional contents.

The felicity conditions of these 19 DAs were determined by the authors (Tab. 1). These were determined considering the meaning of the corresponding utterances in the text-chat database. Differences between several pairs of DAs were clarified using these conditions. For example, we determined that the difference between QUESTION_INFORMATION and QUESTION_SELF was whether the speaker wants the hearer to reply. We also determined that the difference between CONFIRMATION and QUESTION_INFORMATION was whether the answer was implied by the context. The conditions presented with ’*’ in Tab. 1 are considered trivial given a context and excluded from questionnaires of our experiments.

3.2. Speech materials

Four different voice conditions, NEUTRAL, CONV, CONV-DA and NATURAL, were tested in these experiments (Table 2). We expected these voice conditions had different performance for intention communication, and investigated whether the performance difference was reflected in the results. The NEUTRAL voices were TTS read aloud in a neutral speech style. They were generated considering only the linguistic contexts of the input text such as phoneme and accent-type information. We also used synthetic speech in a conversational style, CONV and CONV-DA. While the CONV voices were generated considering only linguistic contexts, the CONV-DA voices were generated considering DA information as well. This allowed the model to predict the prosodic properties of each DA. The NATURAL voices were speech recorded by a voice actor that was uttered considering the text and DA information. These experiments were conducted using voices, texts and instructions in Japanese.

3.2.1. Conversational speech database

We built a speech database whose utterances were tagged with DAs. The database was used in two ways in our experiments: as the training data for TTS models (CONV and CONV-DA) and

Table 1: *The dialogue act set and felicity conditions used in our experiments. Felicity conditions presented with '*' are considered trivial and excluded from questionnaires of our experiments. The variables S, H, P, A and E denotes the speaker, the hearer, a proposition, an action and a psychological state, respectively.*

ID	Name	Felicity Conditions	
1	GREETING	• S is greeting.	* S and H are just encountered.
2	SELF-DISCLOSURE_PLAN	• S wants to inform H that S believes P. * S believes P.	* Perhaps H doesn't know S believes P. * P is a future plan of S.
3	SELF-DISCLOSURE_PREFERENCE+	• S wants to inform H that S feels E. • S feels E.	* Perhaps H doesn't know S feels E. * E is a positive, negative, neutral preference or a desire of S, respectively.
4	SELF-DISCLOSURE_PREFERENCE-		
5	SELF-DISCLOSURE_PREFERENCE0		
6	SELF-DISCLOSURE_DESIRE		
7	QUESTION_INFORMATION	• S wants H to inform P. • S wants to know P.	• S doesn't know P. * H may not inform P without being asked.
8	QUESTION_SELF	• S may not want H to inform P. • S wants to know P.	• S does not know P.
9	ACKNOWLEDGEMENT	• S understood what H said.	• S thinks H will continue to speak.
10	SYMPATHY	• S is in agreement with H about P.	• Perhaps H doesn't know S is in agreement.
11	NON-SYMPATHY	• S is in disagreement with H about P.	• Perhaps H doesn't know S is in disagreement.
12	CONFIRMATION	• S wants H to inform P. • S wants to know P.	• S doesn't know P. * H may not inform P without being asked. * P is implied by the context.
13	PROPOSAL	• S intends to do A or S wants H to do A.	* A is a future act by S or H. * S or H is able to perform A.
14	REPEAT	• S understood what H said. • S is interested in what H said.	* The utterance has similar contents to what H said.
15	APPROVAL	• S wants to inform H that S feels E. • S feels E.	* Perhaps H doesn't know S feels E. * E is a positive evaluation about H by S.
16	THANKS	• S feels grateful for A.	* A is a past act done by H. * A benefits S.
17	APOLOGY	• S feels regret for A.	* A is a past act done by S. * A is an offence against H.
18	FILLER	• S is thinking what to say.	• S thinks S will continue to speak.
19	ADMIRATION	• S feels E.	* E is an affection of S.

Speaker	Dialogue Act	Utterance
A		Is there a convenience store in your neighborhood?
B		There are 3 7-Elevens.
A	ADMIRATION	So many!

Figure 3: *Example of recording manuscript for conversational speech database. Voice for gray colored utterances was recorded. Preceding utterances (by Speaker A and B) were displayed to show conversational context. Dialogue was translated from Japanese by authors.*

as recorded speech samples for evaluation (NATURAL). The sentence set for speech recording was extracted from a text chat database. This database was originally gathered by Higashinaka et al. [22], and contains 3680 conversations (with 134K sentences). The utterances had been manually tagged with DAs by two experts. The sentence set for recording was extracted concerning the balance in the frequency of phonemes and DAs on an entropy basis [23].

For speech recording, we used the manuscripts illustrated in Fig. 3. The manuscript shows not only the pieces of text of the recorded utterances but also their DAs and several preceding utterances. We recorded speech from a Japanese female professional voice actor. We instructed her to read the manuscript to understand the conversational context before speaking every utterance. She spoke the utterance in natural conversational speaking style concerning the context and corresponding DAs. We derived 5177 sentences, about 180 minutes in total.

3.2.2. Neutral speech database

We also used a speech database of neutral read-aloud sentences as the training data for a TTS (NEUTRAL). This database consists of sentences from several genres including news and weather reports. The training data for NEUTRAL contained 7338 sentences, about 560 minutes from this database. The voice was from a Japanese female professional voice actor who was different from that of the conversational speech database.

Table 2: *The four voice conditions.*

Condition	Description
NEUTRAL	TTS, Neutral-style, w/o DA
CONV	TTS, Conversational-style, w/o DA
CONV-DA	TTS, Conversational-style, w/ DA
NATURAL	Recorded, Conversational-style, w/ DA

3.2.3. TTS conditions

For all of the three TTS voice conditions, NEUTRAL, CONV and CONV-DA, deep neutral network (DNN)-based speech synthesis systems [24] were built. The input linguistic feature vector for NEUTRAL and CONV contained 506 dimensional linguistic features. For CONV-DA, the input feature vector was concatenated with a 33 dimensional 1-hot vector that indicated the DA of an utterance. The WORLD vocoder [25] was employed to extract 40 dimensional mel-cepstral coefficients, five band aperiodicities, and F0 in log scale at 5-msec steps. When synthesizing speech parameters, the output parameters were modified by using a global variance-based post filter [26]. The speech samples used as the test set were excluded from the training data.

Fig. 4 illustrates F0 contours by the four voice conditions for an utterance "Sodesune", which means either "Let's see (FILLER)" or "Yes (SYMPATHY)" depending on the prosody. The CONV-DA and NATURAL samples were generated considering that the DA was FILLER. We can see that the NEUTRAL had less F0 variation compared with other voice conditions. The CONV had a different type of sentence-final intonation compared with NATURAL and CONV-DA. We expected this difference in sentence-final intonation would contribute to difference in intention communication performance [27].

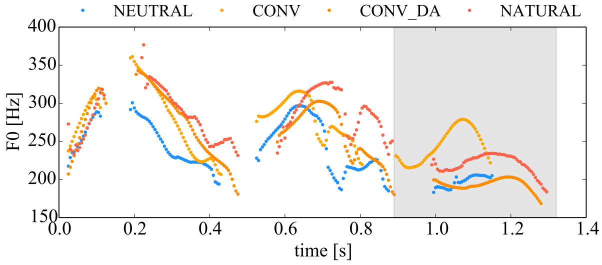


Figure 4: *Example of F0 of the sentence 18 "Sodesune. ('Let's see.' or 'Yes.' in English)" by the four voice conditions. Gray shaded area corresponds to sentence-final intonation.*

Table 3: *Experimental results. Bold values were confirmed significant by F-tests ($\alpha = 0.05$).*

Conventional			Proposed		
F-ratio	V_A	V_R	F-ratio	V_A	V_R
7.54	8.70	1.15	**11.35**	10.20	0.90

3.3. Experiment setting

We conducted two experiments, with the baseline and the proposed protocol. We recruited 10 native Japanese-speaking participants in their 20's and 30's (2 males and 8 females) from outside of the authors' organization. All participants reported normal hearing ability. The experiments were conducted by using headphones and the GUIs illustrated in Fig. 1 and Fig. 2. No time constraint was given for the answers in both experiments. The test sentences were selected from the conversational speech database. We intentionally selected these sentences so that the DA perceived from at least 1 of 3 TTS voices would be different from the reference due to the prosodic variety. As a context to display on GUIs, we used the original conversational context of the text chat database. We had 76 samples (19 sentences × 4 systems) for each of the 10 participants and 2 protocols. We also added 10 samples with fake reference DAs to prevent participants from careless ratings, while these ratings were not considered in statistical analysis. To minimize ordering effects, the order of all samples was randomized for each listener.

In order to investigate inter-rater reliability and sensitivity of the protocols, we conducted one-way analysis of variance (ANOVA). For each protocols, unbiased variances of between-group variation (V_A) and those of within-group variation (V_R) were calculated. Here, a group was defined as ratings for a speech sample by different participants. We conducted F-test with 95% confidence to determine that the change of the variances between two protocols was significant or not. We also calculated F-ratio (=V_A/V_R) for each sentence and protocol. Since a larger V_A and a smaller V_R indicates a higher sensitivity and a higher inter-rater reliability, respectively, F-ratio values can be considered as a measure of quality for an evaluation protocol.

4. Results

Tab. 3 shows the experimental results. We can see that the proposed protocol had larger V_A and smaller V_R. The calculation of the F-test confirmed that both of these differences by protocols were significant. These indicate that the proposed protocol improved both sensitivity and inter-rater reliability. As a result of these changes of variances, the F-ratio was also improved by the proposed protocol.

Fig. 5 shows the scatter plot of F-ratio calculated for each sentence and protocol. The indexes correspond to DA and sentence IDs in Tab. 1. We can see that F-ratios for mots sentences were improved by the proposed protocol. The contribution of the proposed protocol was especially large for sentence 2 (SELF-DISCLOSURE_PLAN), 7 (QUESTION_INFORMATION), 8 (QUESTION_SELF), 17 (APOLOGY) and 18 (FILLER). It is considered that the definition of these intentions by the baseline was difficult for participants to understand and that the proposed protocol alleviated this problem. F-ratio was deteriorated for some sentences by the proposed protocol, while the difference was relatively small. In order to investigate the cause of the deterioration, it is considered that experiments with more participants are needed.

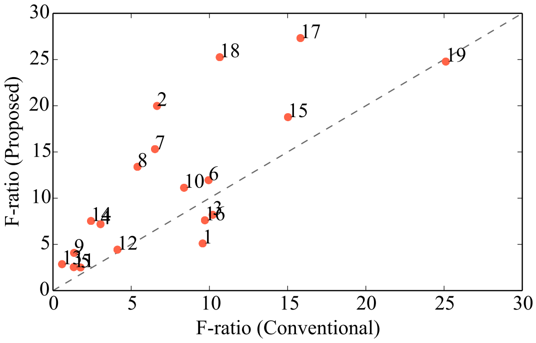


Figure 5: *Scatter plot of F-ratio for each sentence and protocol. Indexes correspond to DA and sentence IDs in Tab. 1*

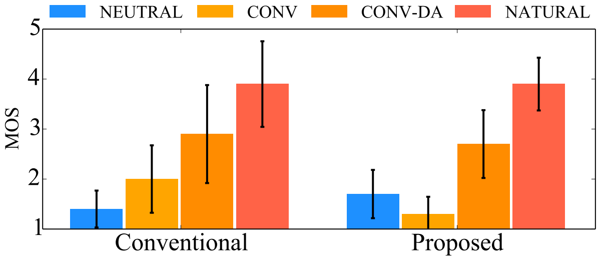


Figure 6: *MOS results for sentence 18 (FILLER) with 95% confidence interval*

Fig. 6 shows the MOS results for sentence 18 (FILLER), whose F-ratio was largely improved by the proposed protocol. We can see that confidence intervals with the proposed protocol were reduced. Although we can see a trend that CONV-DA outperforms CONV in MOS scores for both protocols, significant difference was confirmed only for the proposed protocol. This is an example of the benefit of improved sensitivity by the proposed protocol.

5. Conclusion

From the perspective of pragmatics, we consider it necessary to evaluate intention communication by TTS. For this purpose, in this paper, we proposed a novel evaluation protocol. We compared the proposed protocol with the conventional baseline using four different voice conditions. The results of evaluation experiments showed that the proposed protocol improved sensitivity and inter-rater reliability. Although these results were confirmed statistically significant in our experimental conditions, their robustness should be confirmed under different conditions, such as with more participants [28], using more diverse contexts, prompts and voice conditions, and in different languages. Our future work will also include investigating the validity of the felicity conditions and the trivial ones which were defined by authors in our experiments.

6. References

[1] H. P. Grice, *Studies in the Way of Words*. Harvard University Press, 1991.

[2] J. L. Austin, *How to do things with words*. Oxford university press, 1975.

[3] J. R. Searle, F. Kiefer, M. Bierwisch *et al.*, *Speech act theory and pragmatics*. Springer, 1980, vol. 10.

[4] K. Maekawa, "Production and perception of 'paralinguistic' information," in *Speech Prosody 2004, international conference*, 2004.

[5] N. Hellbernd and D. Sammler, "Prosody conveys speaker's intentions: Acoustic cues for speech act perception," *Journal of Memory and Language*, vol. 88, pp. 70–86, 2016.

[6] D. Wilson and D. Sperber, *Meaning and relevance*. Cambridge University Press, 2012.

[7] A. Schmidt-Nielsen, "Intelligibility and acceptability testing for speech technology," NAVAL RESEARCH LAB WASHINGTON DC, Tech. Rep., 1992.

[8] ITU-T, *A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices*, International Telecommunication Union Std., 1994.

[9] R. van Bezooijen and V. J. van Heuven, "Assessment of speech synthesis," *Handbook of standards and resources for spoken language systems*, pp. 481–653, 1997.

[10] T. Nagata, H. Mori, and T. Nose, "Dimensional paralinguistic information control based on multiple-regression hsmm for spontaneous dialogue speech synthesis with robust parameter estimation," *Speech Communication*, vol. 88, pp. 137–148, 2017.

[11] F. Hinterleitner, G. Neitzel, S. Möller, and C. Norrenbrock, "An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks," in *Proc. Blizzard Challenge Workshop. International Speech Communication Association (ISCA)*. Citeseer, 2011.

[12] R. E. Mayer, K. Sobko, and P. D. Mautone, "Social cues in multimedia learning: Role of speaker's voice." *Journal of educational Psychology*, vol. 95, no. 2, p. 419, 2003.

[13] C. I. Nass and S. Brave, *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge, MA, 2005.

[14] R. K. Atkinson, R. E. Mayer, and M. M. Merrill, "Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice," *Contemporary Educational Psychology*, vol. 30, no. 1, pp. 117–139, 2005.

[15] S. Hennig and R. Chellali, "Expressive synthetic voices: Considerations for human robot interaction," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 589–595.

[16] B.-H. Su, P.-W. Fu, P.-C. Lin, P.-Y. Shih, Y.-C. Lin, J.-F. Wang, and A.-C. Tsai, "A spoken dialogue system with situation and emotion detection based on anthropomorphic learning for warming healthcare d," in *2014 International Conference on Orange Technologies*. IEEE, 2014, pp. 133–136.

[17] J. Mendelson and M. P. Aylett, "Beyond the listening test: An interactive approach to tts evaluation." in *INTERSPEECH*, 2017, pp. 249–253.

[18] Y. Chiba, T. Nose, T. Kase, M. Yamanaka, and A. Ito, "An analysis of the effect of emotional speech synthesis on non-task-oriented dialogue system," in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, 2018, pp. 371–375.

[19] P. F. Strawson, "Intention and convention in speech acts," *The philosophical review*, vol. 73, no. 4, pp. 439–460, 1964.

[20] T. Meguro, R. Higashinaka, Y. Minami, and K. Dohsaka, "Controlling listening-oriented dialogue using partially observable markov decision processes," in *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 2010, pp. 761–769.

[21] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[22] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo, "Towards an open-domain conversational system fully based on natural language processing." in *COLING*, 2014, pp. 928–939.

[23] T. Nose, Y. Arao, T. Kobayashi, K. Sugiura, Y. Shiga, and A. Ito, "Entropy-based sentence selection for speech synthesis using phonetic and prosodic contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[24] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *In Proc. ICASSP 2013*, 2013, pp. 7962–7966.

[25] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[26] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Ways to implement global variance in statistical speech synthesis," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[27] K. Iwata and T. Kobayashi, "Expressing speaker's intentions through sentence-final intonations for japanese conversational speech synthesis," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[28] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No!—An empirically-supported critique of interspeech 2014 TTS evaluations," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.