

Experimental Environment for Verification of Speech-based News Delivery System

Hiroshi HONDA¹Hiroaki TAKATSU, Ph.D.²Yoichi MATSUYAMA, Ph.D.²Tetsunori KOBAYASHI, Ph.D.²

ABSTRACT

An original spoken dialogue system was developed that is capable of diverse responses to user utterances while communicating information using news articles distributed on the Internet. Most conventional spoken dialogue systems are question-and-answer types, and are not suited to communicating large amounts of information such as news articles. In addition, users are unable to interrupt news article read-aloud systems with questions. Therefore, new key technologies related to spoken dialogue systems were developed, including conversation summarization technology, a paralinguistic recognizer, pause adjustment, and a speech synthesizer, and an experimental system was constructed that is capable of spoken dialogue in Japanese using news articles distributed in real time on the Internet. This enabled easy system verification in various usage scenarios. Practical application of this system will enable spoken dialogue in the manner of a conversation partner for drivers while driving.

1. Introduction

The automotive industry business environment is changing precipitously on a global scale, and new technologies and services such as connected, autonomous, shared, and electric (CASE) are being developed at a rapid pace. Advances in connected technology in particular are letting drivers increasingly enjoy a variety of services while in the car, and in-car systems capable of spoken dialogue such as smart speakers are becoming more and more common. One example of this is the Honda Personal Assistant, which understands intentions from driver utterances and provides functions such as music playback, route search, and display of user manuals.

This type of in-car spoken dialogue system has evolved with the advancement of connected technology, and processing such as speech recognition and language comprehension that was formerly executed by onboard units is increasingly executed on the Cloud. This has dramatically increased system performance, and linkage with other Cloud services such as music distribution and geographic information is also expanding.

Furthermore, as automated driving becomes more widespread, the use of time in cars is also about to change. Drivers have thus far had to focus on driving, but if

driving is performed only when needed, more time can be spent freely while on the move, and needs for external communication will also increase. The present level of automated driving in this process has the constraint that the driver needs to constantly watch the road ahead. For this reason, a spoken dialogue system that enables communication only by voice is thought to have high value. In this manner, it is thought that in the CASE era, demand will grow for a spoken dialogue system that can be a conversation partner for drivers while driving.

Therefore, we developed an original spoken dialogue system⁽¹⁾ that is capable of diverse responses in accordance with user utterances while communicating information in news articles distributed in real time on the Internet.

This paper introduces an experimental environment created for verification of this system, which enables spoken dialogue with the user via a tablet terminal.

2. Spoken Dialogue Systems

2.1. Conventional Spoken Dialogue Systems

Most conventional spoken dialogue systems such as smart speakers are question-and-answer (Q&A) or read-aloud types.

Q&A-type systems return a single response to a single

¹ Honda Motor Co., Ltd.

² Waseda University

question. For example, if the user asks, “What is today’s weather?” the system returns a response such as, “Today’s weather in Tokyo is clear.” With this method, the system does not speak spontaneously, so the user is unable to freely engage in dialogue.

In a read-aloud-type system, for example, if the user asks, “What is today’s news?” the system reads aloud news articles in the manner of, “The 19th Round Mexican Grand Prix Final Race was held starting from 1:10 p.m. on October 29. The weather that morning was blue skies, but...” Therefore, with this method as well, even if the user asks a question, the system does not answer, so the user is unable to engage in dialogue.

These types of conventional spoken dialogue systems generally consist of a speech recognizer, dialogue manager, and speech synthesizer (text-to-speech) as shown in Fig. 1. When the user voice is input to the system, the speech recognizer first converts it to text and then passes the converted text to the dialogue manager. The dialogue manager references the scenario database and generates a response sentence in accordance with the user utterance intention. Finally, the generated response sentence is passed to the speech synthesizer and the system voice is output.

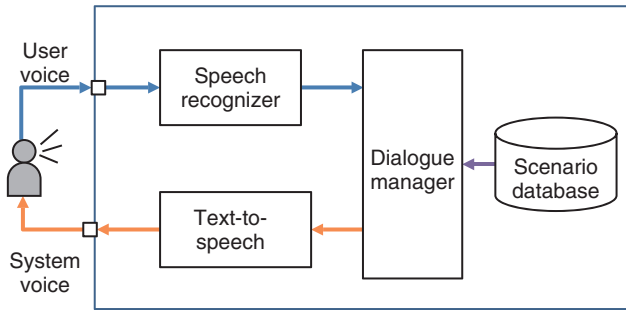


Fig. 1 Architecture of conventional system

2.2. Proposed Spoken Dialogue System

In contrast, the proposed spoken dialogue system realizes free dialogue between the user and the system using news articles distributed on the Internet. For example, if the system utters the sentence, “The 19th Round Mexican Grand Prix Final Race was held.” and the user asks, “When?,” the system can respond, “At 1:10 p.m. on October 29.”

With conventional system technology, the system was unable to respond even if the user asked a question about the information that the system was currently communicating. The system also performed response processing after receiving a question from the user, which made highly responsive dialogue a challenge. The proposed system consists of a dialogue pre-processing block and a dialogue processing block as shown in Fig. 2. The dialogue pre-processing block collects and analyzes news articles on the Internet in advance to create a scenario database of utterances to be communicated to the user and a QA database for responding to questions from the user.

These dialogue pre-processing and dialogue processing blocks realize Q&A during information communication and highly responsive dialogue, which are challenges for conventional systems.

2.2.1. Dialogue pre-processing block

The dialogue pre-processing block uses conversation summarization technology to generate scenarios and Q&A from news articles. This conversation summarization technology is realized by the scenario generator and QA generator shown in Fig. 2. The news articles input to this system are expressed using written language and are not suitable for voice transmission. Therefore, the dialogue pre-processing block summarizes the news articles for conversation and converts them into spoken language. The summarized sentences are stored in the scenario database, and sentences not used in these summaries are stored in the QA database. In addition, Wikipedia and other information

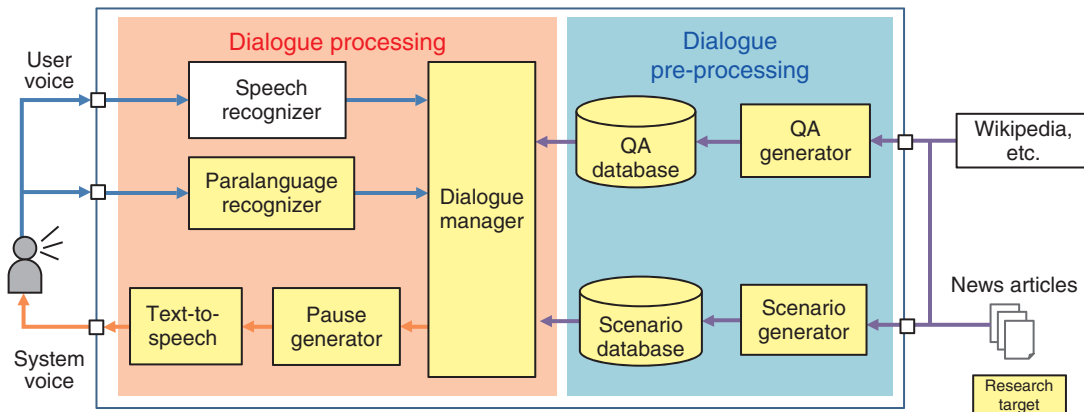


Fig. 2 Architecture of proposed system

on the Internet is used to generate definitions of words extracted from the news articles, which are used for Q&A.

2.2.2. Dialogue processing block

The dialogue processing block converts the scenario database information from text to speech and communicates it to the user. Furthermore, when the dialogue processing block recognizes a user utterance, it uses the QA database information to respond in accordance with the utterance contents.

The dialogue processing block has a paralinguage recognizer as shown in Fig. 2, which performs system actions in accordance with the utterance intention. Paralinguage refers to peripheral language information such as intonation, rhythm, and pauses.

For example, if the system says, “The 19th Round Mexican Grand Prix Final Race was held,” and the user asks, “The Mexican Grand Prix?,” conventional speech recognition alone can acquire only the text information “The Mexican Grand Prix,” so the user’s intention cannot be interpreted. However, the proposed system recognizes six types of paralinguage to tell whether the user’s intention is a question or a request for repetition. If the sentence ending seems to have rising intonation, the system determines that the user is asking a question and can perform Q&A in the manner of, “The Mexican Grand Prix is...” If the sentence ending has falling intonation, the system determines that the user wants agreement and repeats, “Yes, the Mexican Grand Prix.”

The dialogue processing block also has a pause generator as shown in Fig. 2, which adjusts the pauses between system utterances to realize a conversational rhythm that makes it easy for the user to understand and ask questions. The proposed system estimates the pauses between phrases within utterances and the pauses between utterances, and inserts pauses into system utterances. This realizes pauses that make it easier for the user to understand and to interrupt mid-utterance, even when communicating large amounts of information such as news articles.

The dialogue processing block communicates large amounts of information such as news articles, so it also has a speech synthesizer that is capable of distinctive speech in accordance with the roles of utterances within the discourse structure.

3. Key Technologies

3.1. Conversation Summarization Technology

A summarization model that uses deep learning⁽²⁾ was developed to generate conversational summaries from news articles.

3.1.1. Word importance estimation model

Important sentences are extracted from news articles to create summaries from those articles. The important sentence extraction model used a maximum coverage

model^{(3),(4)} which extracted the set of sentences that covers as many important words as possible within the limited number of sentences. Values estimated by random forest, which is a machine learning technique traditionally used for classification and regression, were used to extract the important words in articles. This study applied deep learning to extract important words.

Figure 3 shows the word importance estimation model, which uses deep learning. This model uses BERT⁽⁵⁾, and estimates the importance of words in context. BERT is a natural language processing model that enhances task performance by pre-training with large amounts of document data. The training data used 100 news articles that were summarized by three workers. The auxiliary information input to the model used the following twelve types of information.

- (1) Morphological information consisting of part of speech class, part of speech subclass, conjugated form, conjugation, category, and domain
- (2) Index value representing the word appearance frequency (Term Frequency: TF)
- (3) Index value representing the rarity of a word (Inverse Document Frequency: IDF)
- (4) Product of TF and IDF (TF-IDF)
- (5) Whether the word is included in the title
- (6) Whether the word is in parentheses
- (7) Positions of the sentence and paragraph in the article
- (8) Unique expression class
- (9) Modification type
- (10) Depth of the modification tree (tree representing the modification relation of clauses)
- (11) Number of clauses of the modification source
- (12) Position of the clause from the start of the sentence

The worker ID was used to mix and learn data created by multiple workers, and the worker information was distinguished and given by 1-of-k. 1-of-k is a vector expression with one component as 1 and the remaining components all 0.

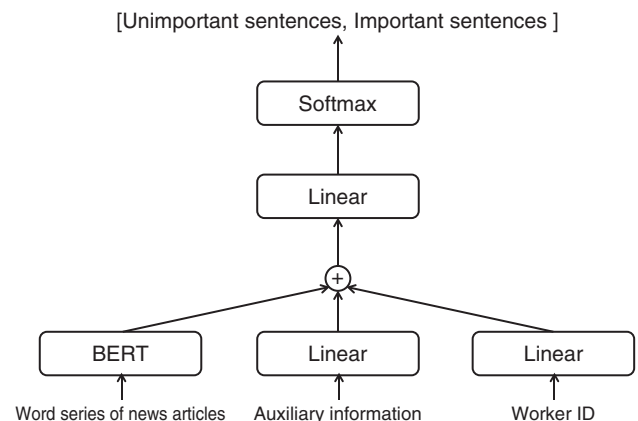


Fig. 3 Word importance estimation model

Figure 4 shows the model evaluation results. The rating scale used ROUGE-1, which is a text summary rating value that represents a better rating closer to 1.0. SEM1 is the model that learned using the data of a single worker without BERT pre-training or auxiliary information, SEM1_pre is the model with pre-training added to SEM1, SEM1+_pre is the model with auxiliary information further added to SEM1_pre, and SEM2+_pre is the model that learned using the data of two workers with BERT pre-training and auxiliary information. For comparison, RF is the random forest results, Human is the worker results, Random is the results when random sentences are extracted, and LEAD is the results when the first K sentences of the article are extracted. The experiment results showed that the proposed method, SEM2+_pre, provided the best results and achieved higher performance compared to the conventional method, random forest.

3.1.2. Sentence compression model

News articles are described in written language, so a single sentence contains a large amount of information. Therefore, articles are converted into a form suitable for communicating in spoken language by reducing the amount of information contained in a single sentence and performing sentence compression. Sentence compression was formulated as a sequence labeling problem at the clause level. A sequence labeling problem is a problem that assigns an appropriate label string to each element of a sequence. This study labeled whether or not to include clause units in utterances.

Figure 5 shows the sentence compression model using deep learning. This model uses BERT, and differs from the word importance estimation model in that the embedded expression of words obtained by BERT is constructed for each clause by self-attention⁽⁶⁾. Like the word importance estimation model, the training data used 100 news articles summarized by three workers. The auxiliary information input to the model was the same as that used by the word importance estimation model minus the sentence and paragraph positions.

Figure 6 shows the results evaluated by accuracy when test data was input to the model. SCM1 is the model that

learned using the data of a single worker without BERT pre-training or auxiliary information and SCM1_pre is the model with pre-training added to SCM1. SCM1+_pre is the model with auxiliary information further added to SCM1_pre, and SCM2+_pre is the model that learned using the data of two workers with BERT pre-training and auxiliary information. For comparison, BiLSTM shows the results with sentence compression using bidirectional long short-term memory (LSTM), which is a deep learning model that can handle time series data using the word appearance frequency vector (bag of words) that constitutes the auxiliary information and clause as the feature. In addition, Human shows the worker results, Random shows the results when random clauses are extracted, and NoCompression shows the results without sentence compression. The experiment results showed that the proposed method, SCM2+_pre, provided the best results and achieved higher performance compared to BiLSTM.

3.2. Paralanguage Recognizer

Deep learning was used to generate a user utterance intention recognition model, and a paralanguage recognizer⁽⁷⁾⁻⁽¹¹⁾ was created to realize system responses that

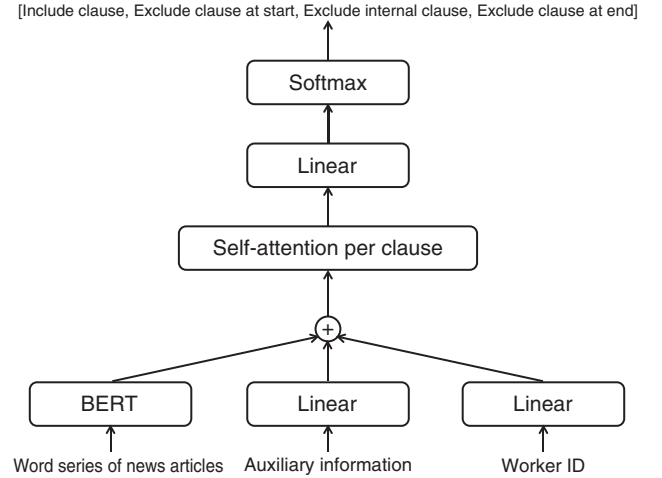


Fig. 5 Sentence compression model

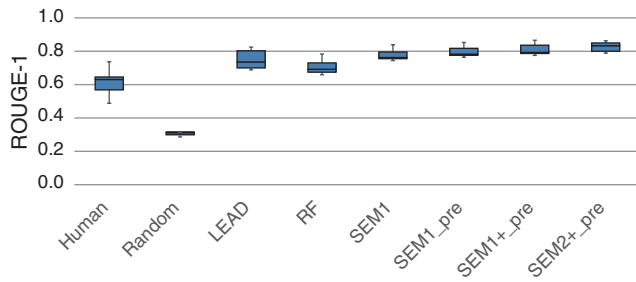


Fig. 4 Evaluation results of important sentence extraction

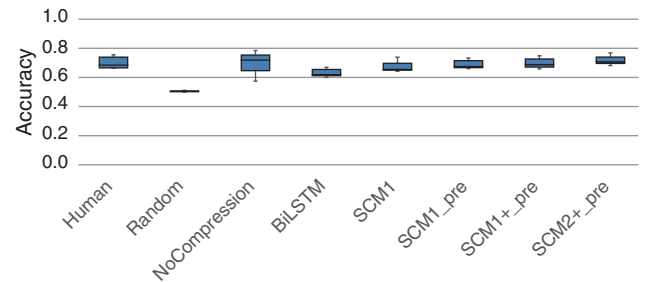


Fig. 6 Evaluation results of sentence compression

Table 1 User intention types, their effects, and corresponding system actions

Effects	User intention	System action
Increasing information	Question	Answer
	Request supplement	Provide explanations
	Request repeat	Repeat
Decreasing information	Disinterest	Change topic
	Already known	Omit details
Avoiding overlaps	Wait request	Listen

Table 2 Discrimination results of intention recognition

		With BERT pre-training With auxiliary information			With BERT pre-training Without auxiliary information		Without BERT pre-training Without auxiliary information	
		P	P+U	P+U+S	P+U	P+U+S	P+U	P+U+S
Question	Majority vote	0.920	0.958	0.962	0.957	0.961	0.940	0.951
	At least one vote	0.920	0.957	0.961	0.953	0.956	0.923	0.938
Request supplement	Majority vote	0.836	0.911	0.913	0.901	0.906	0.895	0.896
	At least one vote	0.827	0.882	0.887	0.881	0.887	0.847	0.865
Request repeat	Majority vote	-	-	-	-	-	-	-
	At least one vote	0.700	0.826	0.832	0.795	0.805	0.784	0.805
Disinterest	Majority vote	0.893	0.913	0.908	0.903	0.898	0.898	0.897
	At least one vote	0.797	0.811	0.805	0.810	0.804	0.804	0.803
Already known	Majority vote	0.818	0.939	0.939	0.939	0.939	0.939	0.939
	At least one vote	0.702	0.825	0.831	0.817	0.828	0.806	0.817
Wait request	Majority vote	0.716	0.885	0.865	0.885	0.858	0.845	0.838
	At least one vote	0.614	0.688	0.692	0.688	0.689	0.634	0.672

correspond to user utterance intentions.

Figure 7 shows the model for recognizing user utterance intentions. When time series voice data is input to this model, the currently anticipated utterance intent label is output. Training used an utterance intention data set created from approximately 2000 dialogues with this system. This data set was created by labeling user utterances with the user utterance intentions shown in Table 1.

First, a spectrogram generated by cutting out a short voice segment is input to CNN Autoencoder. CNN

Autoencoder is a deep learning model capable of extracting feature values without a teacher. Utterance intention labels can be estimated by inputting rhythm feature values obtained from the middle layer of CNN Autoencoder to discrimination part (P). The final labels are estimated by inputting the user utterance speech recognition results and the immediately preceding system utterance to discrimination part (L). BERT pre-training is applied to discrimination part (L).

Table 2 shows the utterance intention recognition

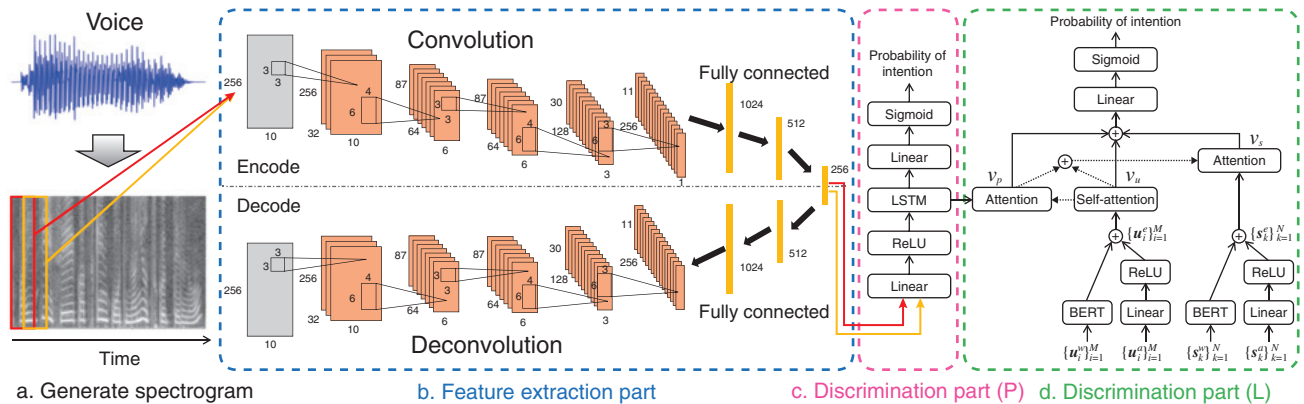


Fig. 7 Intention recognition model

results using this model. P is the results estimated using only the user utterance rhythm, P+U is the results estimated using the user utterance rhythm and speech recognition results, and P+U+S is the results estimated using the user utterance rhythm and speech recognition results and the system utterance. Input of both the user utterance rhythm and the user utterance speech recognition results resulted in higher accuracy for all intentions. In addition, BERT pre-training also increased the accuracy.

3.3. Pause Adjustment

A function was developed that uses deep learning to estimate the pauses of system utterances⁽¹²⁾. This realizes pauses that make it easier for the user to understand and ask questions.

3.3.1. Pause estimation model between phrases within utterances

Figure 8 shows the model for estimating pauses between phrases within utterances. This model uses bidirectional LSTM, and determines the pause between phrases based on the preceding and following LSTM states. The phrase text was input to the model using LSTM.

Model training used a corpus with pause length tags created by having a voice actor read aloud 100 summarized, colloquialized news articles. The voice actor was instructed to observe the following three points when recording the articles.

- (1) Speak with the intention of having the listener understand.
- (2) Speak in a manner that distinguishes important and unimportant parts so that the listener can understand.
- (3) Be aware of the potential for question interruptions while speaking, and insert natural pauses in locations where questions are anticipated.

Figure 9 shows a histogram of the deviation between the pause values estimated using this model and the correct values. Deviation of less than 0.2 s accounted for 77% of the total.

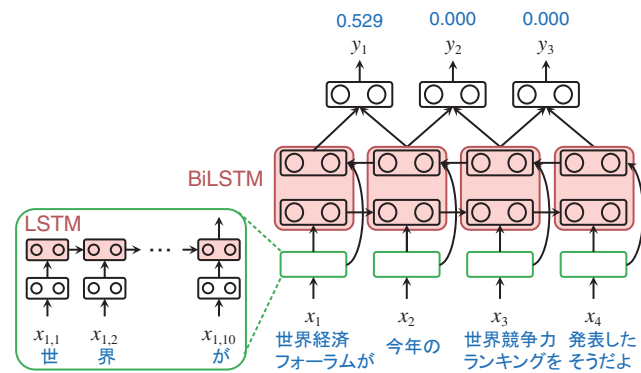


Fig. 8 Pause estimation model between phrases in utterance

3.3.2. Pause estimation model between utterances

Figure 10 shows the model for estimating pauses between utterances. Pauses between utterances are estimated using the information of the current and next utterances. Pauses between utterances were estimated using a Bayesian ridge model, which is a linear regression model, and used the values of the last hidden layer of LSTM in the pause estimation model between phrases within utterances as the feature. Like the pause estimation model between phrases within utterances, training used a corpus with pause length tags. Figure 11 shows a histogram of the deviation between the pause values estimated using this model and the correct values. Deviation of less than 0.5 s accounted for 75% of the total.

3.4. Speech Synthesizer

In order to communicate large amounts of information such as news articles, a speech synthesizer was developed that is capable of distinctive speech using the roles of utterances within the discourse structure⁽¹³⁾. Utterance sentences are divided into “spoken phrase” units that

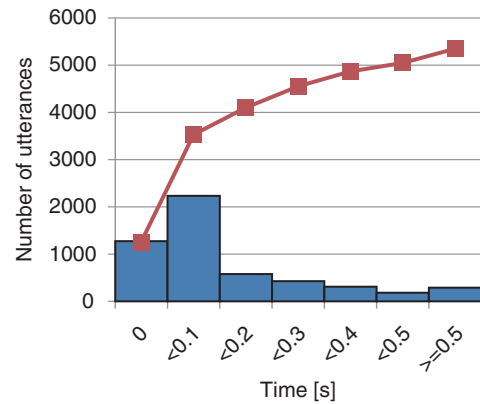


Fig. 9 Pause error between phrases in utterance

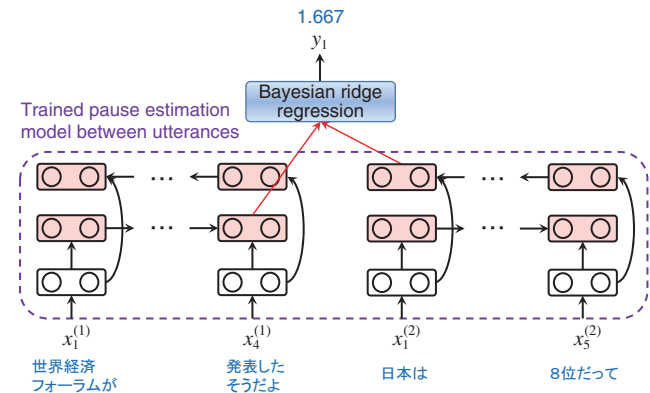


Fig. 10 Pause estimation model between utterances

do not contain more than one new piece of information and can be spoken in a single breath. Spoken phrases are further divided into three roles: sentences containing the information that is the most important to communicate (nucleus), sentences leading into the nucleus (front satellite), and sentences providing supplemental information (rear satellite). These roles are used as auxiliary information for voice synthesis. This enabled us to control the rhythm and pauses of the entire utterance sequence. Figure 12 shows examples of the roles of spoken phrases. Merlin⁽¹⁴⁾, a speech synthesis system based on deep learning, was used as the base speech synthesis system.

3.4.1. Speech synthesis model

Figure 13 shows the voice synthesis model that uses deep learning. As inputs, in addition to linguistic and frame features, this model also applies features (utterance sequence features) that express the positions and roles of

spoken phrases. The spoken phrase position is information that indicates the number of the current utterance within the paragraph.

Speech synthesis model training used a data set created by having a voice actor read aloud 39 news articles. The manuscript was pre-marked with the roles of the spoken phrases, and the voice actor was instructed to observe the following three points.

- (1) Speak in a manner that clearly conveys the contents of nucleus roles to the listener.
- (2) Suppose that the listener is listening while nodding, and speak without leaving large pauses.
- (3) Speak with the intention of communicating news to the listener in a tone similar to everyday conversation.

Table 3 shows the results of calculating the deviation between the speech synthesis model output and the acoustic feature values of the original data by 10-fold cross validation. F0-RMSE and dur-RMSE are respectively the

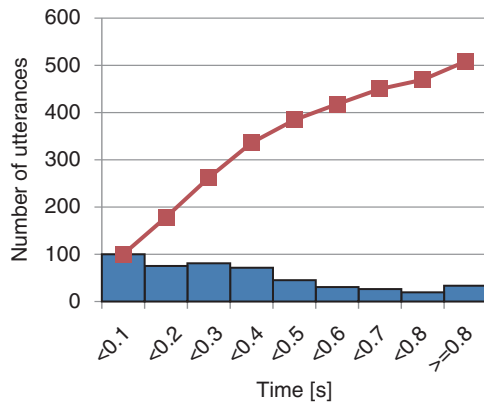


Fig. 11 Pause error between utterances

Paragraph	Position	Role	Content of spoken phrase
1	1	Front satellite	宇宙を観測しているロシアの電波望遠鏡があるんだけど
	2	Nucleus	それが正体不明の「強い信号」を検知したっていつて注目を集めてるんだって
	3	Rear satellite	世界中の天文学者に衝撃が走っているらしいよ
2	1	Nucleus	信号はヘラクレス座の近くの恒星から届いたんだって
	2	Nucleus	その星は地球から約95光年離れていて
	3	Rear satellite	惑星も持っているみたい
3	1	Front satellite	地球外文明の進歩度を表す基準っていうのがあって
	2	Front satellite	それと照らし合わせたんだけど
	3	Nucleus	地球よりはるかに進歩した文明の可能性があるんだって
	4	Rear satellite	まだ確実なことはわからないんだけどね

Fig. 12 Example of discourse structure

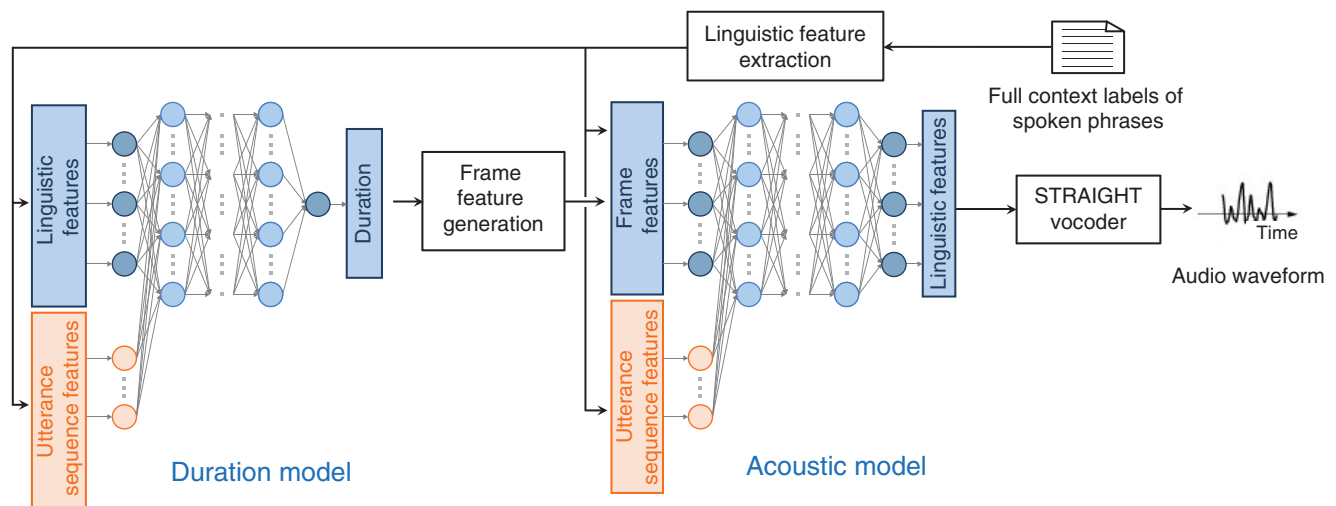


Fig. 13 DNN speech synthesis model considering utterance sequence features

Table 3 Comparison of performances with and without utterance sequence features

	F0-RMSE	dur-RMSE	MCD	BAD
Base line	36.420	4.779	5.064	0.660
Proposed model	35.435	4.768	5.047	0.659

root mean square error of the fundamental frequency and the phoneme duration. MCD and BAD are respectively the mel-cepstral distortion and the band aperiodicity distortion. Mel-cepstral distortion is an objective index for measuring the deviation between the original data and the speech synthesis model output. The addition of utterance sequence features reduced the deviations of the fundamental frequency and the phoneme duration compared to the base line, and also showed slight enhancement in terms of mel-cepstrum and aperiodicity.

4. Experimental System

An experimental system was constructed based on the proposed system described in Section 2.2. This experimental system enables spoken dialogue with the user via a tablet terminal for evaluation of the system by many users in a variety of scenarios inside a traveling vehicle.

Figure 14 shows the configuration of the experimental system. The dialogue pre-processing server collects and analyzes news articles on the Internet. By periodically collecting news articles from the Internet, the dialogue pre-processing server can provide information to the user in real-time. The dialogue processing server performs scenario speech synthesis and other tasks, and the user interacts with

the system via a tablet terminal. Speech recognition used Google Speech API, a speech recognition cloud service provided by Google LLC.

Figure 15 shows an example of using the experimental system. This experimental system uses Japanese language news articles and is capable of spoken dialogue in Japanese. The user performs spoken dialogue using the speaker and microphone of the tablet terminal. This system lets the user hear timely news articles distributed on the Internet and ask the system things they wonder about.

This experimental system can be used to subjectively evaluate willingness to use and to measure items such as information communication efficiency. Verification of the effectiveness of spoken dialog systems will continue in this manner.



Fig. 15 Example of using experimental system

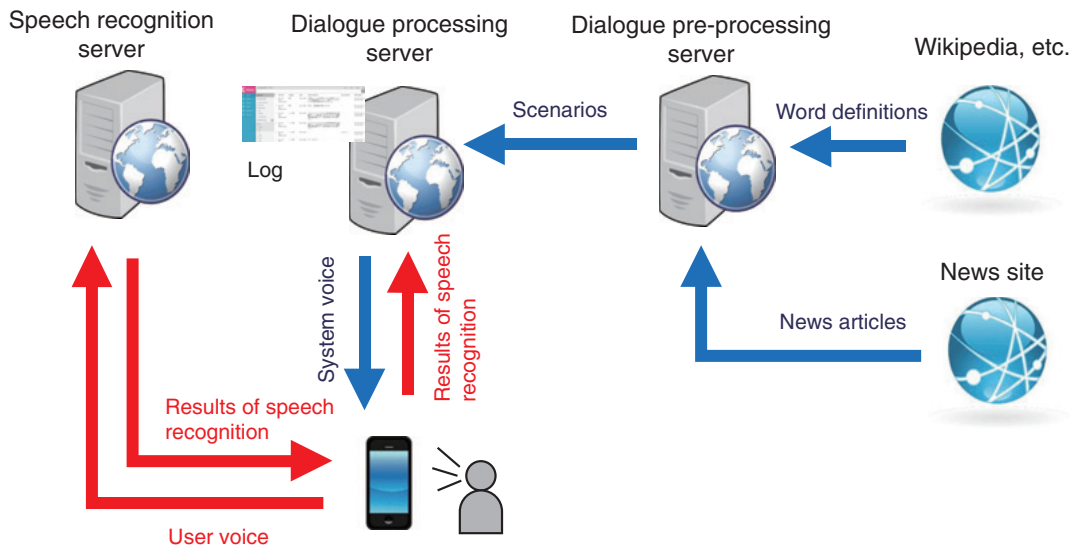


Fig. 14 Experimental system

5. Conclusion

Key spoken dialogue system technologies such as conversation summarization, paralinguage recognition, pause adjustment, and speech synthesis were developed. In addition, a system was constructed that enables dialogue via a tablet terminal equipped with these technologies.

Compared to conventional Q&A systems and news article read-aloud systems, the constructed system enables Q&A during information communication and highly responsive dialogue.

Future plans include evaluation of this system during actual use in a car and support of languages other than Japanese.

References

- (1) Takatsu, H., Fukuoka, I., Fujie, S., Hayashi, Y., Kobayashi, T.: A Spoken Dialogue System for Enabling Information Behavior of Various Intention Levels, Transactions of the Japanese Society for Artificial Intelligence, Vol. 33, No. 1, p. 1-24, (2018) (in Japanese)
- (2) Takatsu, H., Honda, H., Fujie, S., Hayashi, Y., Kobayashi, T.: Kaiwa niyoru Nyusukiji Dentatsu notameno Yoyaku, NLP2019 proceedings, p. 518-521, (2019) (in Japanese)
- (3) Gillick, D., Favre, B., Hakkani-Tür, D.: The ICSI Summarization System at TAC 2008, Text Analysis Conference (TAC) 2008, (2008)
- (4) Takamura, H., Okumura, M.: Text Summarization Model based on Maximum Coverage Problem and its Variant, Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, p. 781-789, (2009)
- (5) Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of NAACL-HLT 2019, p. 4171-4186, (2019)
- (6) Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A Structured Self-attentive Sentence Embedding, Proceedings of the 5th International Conference on Learning Representations, p. 1-15, (2017)
- (7) Takatsu, H., Yokoyama, K., Matsuyama, Y., Honda, H., Fujie, S., Kobayashi, T.: Recognition of Intentions of Users' Short Responses for Conversational News Delivery System, Proceedings of the 20th Annual Conference of the International Speech Communication Association, p. 1193-1197, (2019)
- (8) Yokoyama, K., Takatsu, H., Honda, H., Fujie, S., Kobayashi, T.: Investigation of Users' Short Responses in Actual Conversation System and Automatic Recognition of their Intentions, Proceedings of the 2018 IEEE Workshop on Spoken Language Technology, p. 934-940, (2018)
- (9) Yokoyama, K., Takatsu, H., Honda, H., Fujie, S., Kobayashi, T.: Utterance Intention Classification and Recognition for Conversational News Delivery System, Information Processing Society of Japan SIG-SLP, Vol. 2018-SLP-123, No. 5, p. 1-8, (2018) (in Japanese)
- (10) Takatsu, H., Yokoyama, K., Honda, H., Fujie, S., Hayashi, Y., Kobayashi, T.: Utterance Intention Understanding for News Articles Transfer by Conversation, JSAI2018 proceedings, 4Pin1-29, p. 1-4, (2018) (in Japanese)
- (11) Yokoyama, K., Takatsu, H., Honda, H., Fujie, S., Hayashi, Y., Kobayashi, T.: Utterance Intention Classification and Database Construction for Conversational News Delivery System, JSAI2018 proceedings, 3Pin1-23, p. 1-4, (2018) (in Japanese)
- (12) Takatsu, H., Yokoyama, K., Honda, H., Fujie, S., Hayashi, Y., Kobayashi, T.: Kaiwa niyoru Nyusukiji Dentatsu notameno Ma no Chosei, NLP2018 proceedings, p. 212-215, (2018) (in Japanese)
- (13) Takatsu, H., Fukuoka, I., Fujie, S., Iwata, K., Kobayashi, T.: Speech Synthesis for Conversational News Contents Delivery, Transactions of the Japanese Society for Artificial Intelligence, Vol. 34, No. 2, p. 1-15, (2019) (in Japanese)
- (14) Wu, Z., Swietojanski, P., Veaux, C., Renals, S., King, S.: A study of speaker adaptation for DNN-based speech synthesis, Proceedings of the 16th Annual Conference of the International Speech Communication Association, p. 879-883, (2015)

■ Author ■



Hiroshi HONDA



Hiroaki TAKATSU



Yoichi MATSUYAMA



Tetsunori KOBAYASHI