INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Prosody Control of Utterance Sequence for Information Delivering

Ishin Fukuoka, Kazuhiko Iwata, Tetsunori Kobayashi

Faculty of Science and Engineering, Waseda University, Japan

fukuoka@pcl.cs.waseda.ac.jp

Abstract

We propose a conversational speech synthesis system in which the prosodic features of each utterance are controlled throughout the entire input text. We have developed a "news-telling system," which delivered news articles through spoken language. The speech synthesis system for the news-telling should be able to highlight utterances containing noteworthy information in the article with a particular way of speaking so as to impress them on the users. To achieve this, we introduced role and position features of the individual utterances in the article into the control parameters for prosody generation throughout the text. We defined three categories for the role feature: a nucleus (which is assigned to the utterance including the noteworthy information), a front satellite (which precedes the nucleus) and a rear satellite (which follows the nucleus). We investigated how the prosodic features differed depending on the role and position features through an analysis of news-telling speech data uttered by a voice actress. We designed the speech synthesis system on the basis of a deep neural network having the role and position features added to its input layer. Objective and subjective evaluation results showed that introducing those features was effective in the speech synthesis for the information delivering.
Index Terms: speech synthesis, conversational speech, prosody, discourse analysis, neural network

1. Introduction

Thanks to the progress in the speech technologies in recent years, various kinds of spoken dialog systems with a text-to-speech (TTS) function have emerged, such as information navigation systems [1, 2] and storytelling systems [3, 4]. We have developed a spoken dialog system that delivers news information through synthetic speech [5] (hereinafter, "news-telling system"). The system talks to the users about the news after collecting articles from news websites and converting them from written-style language into utterances in spoken-style language.

The news-telling system focuses on unfailingly delivering the noteworthy information in the news to the users, whereas the storytelling systems try to expressively and dramatically tell a story expressing characteristics and emotions of personages. When professional narrators, for example, deliver a news article, they are expected to change their way of speaking throughout their narration in order to highlight the noteworthy information, or as the article unfolds (i.e. an introduction, development, turn, and conclusion). A key word or phrase in a sentence should be uttered with a higher fundamental frequency (F_0), at a slower speech rate than the other words or phrases, and sometimes with a pause before and/or after it [6, 7]. We could expect these sentence-level phenomena to be observed in the text-level. In other words, a key utterance in a sequence of utterances should be uttered with some particular prosody. The prosodic features such as an F_0 average, an F_0 range, and a speech rate also differ depending on utterance positions (initial, medial, and final) in a discourse segment [8]. Nevertheless, most conventional TTS systems control the sentence-level prosody, that is to say, generate the prosodic features sentence by sentence. Consequently, every sentence utterance produced by the TTS systems is spoken in almost the same way. This makes the synthetic speech monotonous and boring and makes the noteworthy information obscure. The users are forced to concentrate while listening and in some cases miss the noteworthy information.

In consideration of this, we propose a new deep neural network (DNN) based speech synthesis system especially for an information delivering system such as the news-telling system. The prosodic features are controlled throughout the entire news article considering the role and position of individual sentences in order to highlight the noteworthy information. First, we created the scripts for news-telling speech data collection. A news article generally consists of several sentences that have different roles. We defined three categories of roles as the control parameters for the prosody generation: a nucleus, a front, and rear satellite. The terms "nucleus" and "satellite" are based on the rhetorical structure theory [9]. The nucleus is assigned to the utterances containing key information of the news. The front satellite is assigned to the utterances preceding the nucleus, which mainly provide introductory information. The rear satellite is assigned to the utterances following the nucleus, which mainly provide additional information. The news articles collected from the news websites were transformed from written-style into spoken-style language and were uttered by a voice actress. Section 2 describes the data collection. Next, we investigated how the prosodic features differed depending on the role and position of each utterance in the articles. Section 3 details the results for prosodic analysis. We designed the speech synthesis system on the basis of a DNN having the role and position features added to its input layer. Objective and subjective evaluations were conducted comparing the proposed system and a conventional system. The results of the evaluations, which indicated the effectiveness of introducing the role and position features, are described in detail in Section 4. Finally, we conclude the paper in Section 5.

2. Data Collection

In conversational speech synthesis studies, researchers have argued for the necessity of a suitable corpus for each task [10, 11]. We prepared the scripts for recording the speech data of the news-telling task.

The text of the news articles collected from the websites was converted into spoken-style language manually. Each article consists of several paragraphs, and each paragraph consists of three to six utterances. Figure 1 exemplifies a script, which is composed of three paragraphs. The utterance position (UP) indicates an utterance number counted from the beginning of the individual paragraph. The UPs are determined automatically. The utterance role (UR) indicates one of the three roles for the individual utterance. The URs should be determined manually.

http://dx.doi.org/10.21437/Interspeech.2017-708

Paragraph	UP	UR	Utterance	
1	1	FS	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	(a)
			(A Russian radio telescope that observes the universe is drawing attention.)	
	2	N	それが正体不明の「強い信号」を検知したっていって注目を集めてるんだって	(b)
			(Because it observed a "strong" signal from a Sun-like star.)	
	3	RS	世界中の天文学者に衝撃が走っているらしいよ	(c)
			(Astronomers all over the world are very surprised.)	
2	1	N	信号はヘラクレス座の近くの恒星から届いたんだって	(d)
			(The signal was delivered from the star near the Hercules.)	
	2	RS	その星は地球から約 95 光年離れていて	(e)
			(This is about 95 light years from Earth.)	
	3	RS	惑星も持っているみたい	(f)
			(In addition, it has some planets.)	
3	1	FS	地球外文明の進歩度を表す基準ていうのがあって	(g)
			(The signal was checked with a standard)	
	2	FS	それと照らしあわせたんだけど	(h)
			(that represents a degree of progress about extraterrestrial civilization.)	
	3	N	「地球よりもはるかに進歩した文明の可能性があるんだって」	(i)
			(It might be a civilization going ahead of the earth.)	
	4	RS	まだ確実なことはわからないんだけどね	(j)
			(though it is just possibility.)	

Figure 1: *Example of script created from news article. Nucleus utterances in bold font. UP: Utterance Position, UR: Utterance Role, FS: Front Satellite, N: Nucleus, RS: Rear Satellite.*

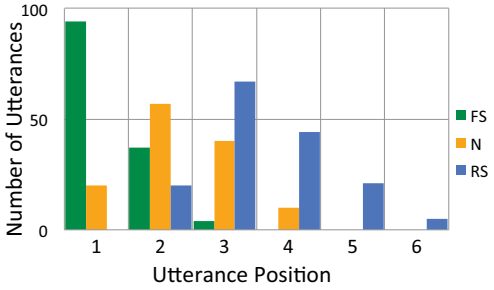


Figure 2: *Distribution of utterances for each UR.*

The nucleus is assigned to the utterance containing noteworthy information of the article. We defined at least one utterance including key information in the individual paragraph as a nucleus. Utterances (b), (d), and (i) are assigned the nuclei and indicated by the bold font in the script. Utterances (a), (g), and (h) are the front satellites, and utterances (c), (e), (f), and (j) are the rear satellites. Eventually, 39 articles, 114 paragraphs, and 419 utterances were created. Figure 2 illustrates the distribution of the utterances with each UR. The X-axis represents the UPs. The front-satellite utterances were mainly found in the first UP, the nuclei were frequently in the second and third UPs, and the rear satellites were in the third and fourth.

The scripts were uttered by the voice actress paragraph by paragraph. We presented the following instructions to her in advance. 1. *Speak like in a daily conversation.* 2. *Speak fluently.* 3. *Emphasize "nucleus" utterances.*

3. Prosodic Analysis

This section describes prosodic features affected by the UR and the UP in the news-telling task. We investigated how the prosodic features differed depending on the UR and the UP in the news-telling speech data.

3.1. Analysis Setup

We analyzed prosodic features such as an F_0 average in the whole of an utterance, an F_0 dynamic range in the whole of an utterance, an F_0 average in the first three moras[1] of an utterance, a speech rate of an utterance, and a pause between utterances. The F_0 dynamic range was calculated as a ratio of max F_0 value and minimum F_0 value in the utterance. The frame width and frame shift were 5 ms and 1 ms in an F_0 analysis with STRAIGHT [12]. A phoneme segmentation was performed with Julius[2] and HTS[3]. The speech rate in moras per second was measured by excluding the duration of pauses within the utterance.

[1]The mora is the smallest temporal unit in Japanese.

3.2. Prosodic Features Affected by Role and Position

Figure 3 illustrates the relationships between the UR and the prosodic features. Figure 4 depicts the relationships between the UP and the prosodic features.

Figure 3(a) shows that the UR affects the F_0 average of the utterances. The F_0 average distribution of the nuclei is higher than those of the other URs. The F_0 dynamic range distribution of the nuclei is also higher (Figure 3(b)). On the other hand, the F_0 average in the first three moras of the utterances are affected by the UPs (Figure 4(c)). The first utterance in the paragraph starts with higher F_0 than the succeeding ones. Figure 3(d) shows that the speech rate of the nucleus utterances become slower than those of utterances with the other URs.

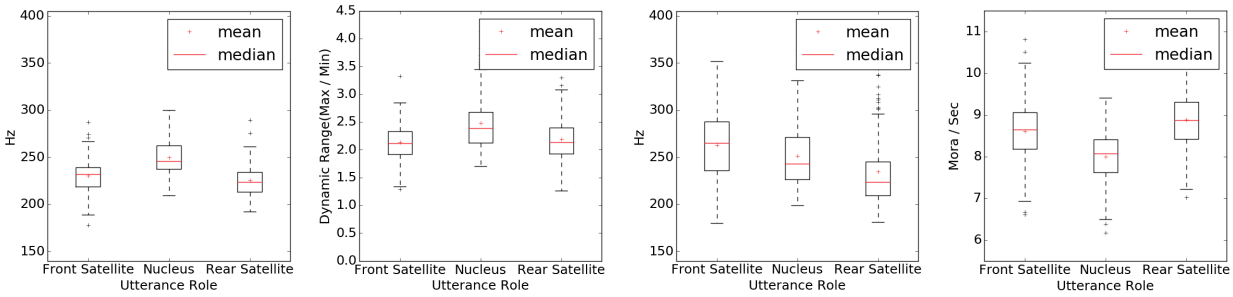
Figure 5 shows the pause distributions between the utterances. The X-axis represents transitions of the URs. There are five types of transitions in the three URs. The pauses between the nucleus and the rear satellite are significantly longer than the other pauses. This result shows that the voice actress took long pauses after the utterances that contained noteworthy information in the article.

4. System Design and Experiment

The analysis results in Section 3 indicated that the prosodic features of the sequence of utterances were affected by the UR and the UP. We constructed a DNN-based speech synthesis system by adding the UR and UP as some of the input features into conventional DNN models. The effectiveness of adding the UR and UP features was examined in objective and subjective experiments.

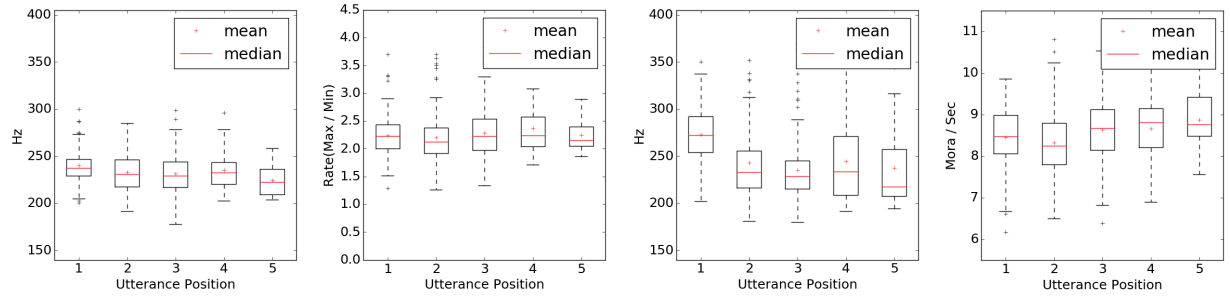
[2]http://julius.osdn.jp/

[3]http://hts.sp.nitech.ac.jp/



(a) F_0 *averages in whole of utterance.* (b) F_0 *dynamic ranges in whole of utterance.* (c) F_0 *averages at front of utterance.* (d) *Speech rates.*

Figure 3: *Relationship between UR and prosodic features.*



(a) F_0 *averages in whole of utterance.* (b) F_0 *dynamic ranges in whole of utterance.* (c) F_0 *averages at front of utterance.* (d) *Speech rates.*

Figure 4: *Relationship between UP and prosodic features.*

4.1. DNN-based system

Figure 6 depicts the proposed system, which includes two DNN models: duration DNN and acoustic DNN. Both DNN models have an input layer that includes individual UR and UP features. These features are expected to function as control parameters for generating the speech parameters that are suitable for the news-telling speech. The conventional DNN-based speech synthesis system [13] is the same as the proposed system shown in Figure 6 except for the elimination of the UR and UP features.

In the proposed system, the input layer of the duration DNN consists of linguistic, UR and UP features, which are constructed from the input text. The duration DNN predicts phoneme durations from these features. In addition to the linguistic, UR and UP features, the input of the acoustic DNN includes frame features that are constructed from the output of the duration DNN. The acoustic DNN maps these features to the speech parameters at every frame. The dimension number of the linguistic features is 596. The frame features are the same nine numeric values used in Merlin [14]. Both DNN models have 6 hidden layers, and each layer has 1024 units. The outputs of the acoustic DNN consist of 40 mel-cepstral coefficients, logarithmic F_0 (log F_0) values, 5-band aperiodicity coefficients, Δ, Δ^2 features, and a voiced-unvoiced feature. A STRAIGHT vocoder was used for generating speech waveforms from the speech parameters.

4.2. Experiment

4.2.1. Objective Evaluation

We compared the prosodic errors of the DNN models in the proposed and conventional systems. The difference between these systems was whether or not the DNN models include the UR and UP as the input features. The UR features were expressed as a one-hot vector that represents the UR category of the input utterance. The UP features consisted of three numeric values: an utterance number from the head in a paragraph, an utterance number from the tail in a paragraph, and the number of utterances in a paragraph. The collected news-telling speech data (about 400 utterances) were used for building the DNN models. Twenty utterances were assigned for validation data and another 20 for test data. The DNN models were trained using the rest of the data.

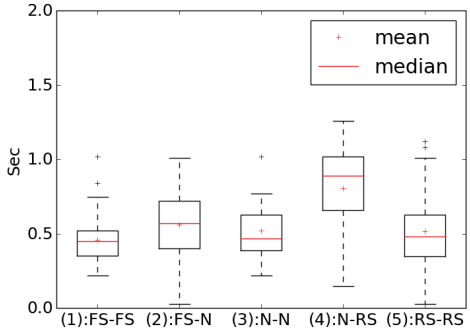


Figure 5: *Pause distributions between utterances. "FS": Front Satellite, "N": Nucleus, "RS": Rear Satellite.*

Table 1 shows the RMS errors (RMSE) of the DNN models in the conventional and proposed systems. Every RMSE was calculated by comparing the prosodic value of the test data with that output by the DNN model. "dur-RMSE" means a duration error. "F_0-RMSE" means an F_0 error for the whole of the utterance. "F_0-beg-RMSE" means an F_0 error for the beginning of the utterance with 350 ms duration (approximately 3-mora length). As for the calculation of the RMSEs related to the F_0,

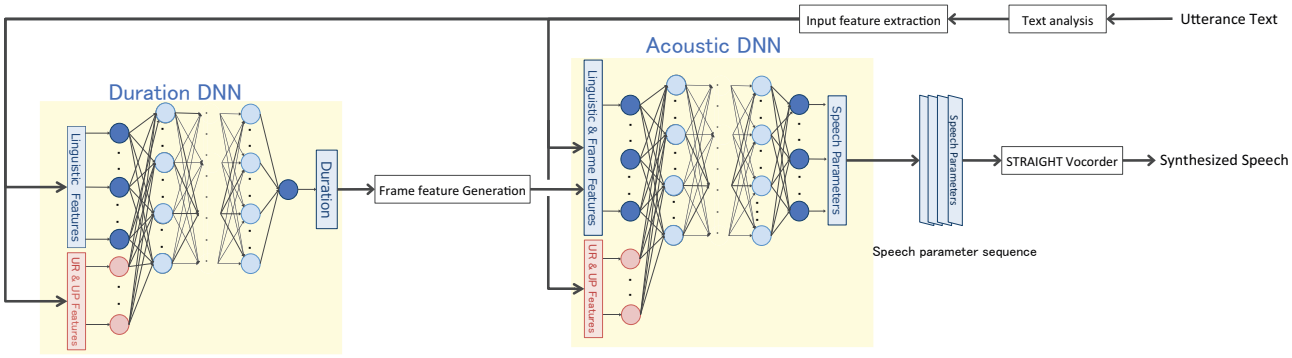


Figure 6: *Proposed System.*

Table 1: *Errors of DNN models in conventional and proposed systems.*

Model	dur-RMSE	F_0-RMSE	F_0-beg-RMSE
Conventional	4.85	37.78	41.72
Proposed	4.82	36.26	40.57

Table 2: *Differences among three models in subjective evaluation. "Single": Single Utterance Model, "Sequential (w/o features)": Sequential Utterances Model without UR and UP features, "Sequential (w/ features)": Sequential Utterances Model with UR and UP features. Training data "PB": phonetically balanced speech data, "NT": news-telling speech data.*

Model	Training data	System type
Single	PB	Conventional
Sequential (w/o features)	PB + NT	Conventional
Sequential (w/ features)	PB + NT	Proposed

the acoustic DNN generated F_0 by using the phoneme durations extracted from the test data as the frame features. All prosodic errors of the proposed system outperformed respective errors of the conventional system.

4.2.2. *Subjective Evaluation*

We conducted a perceptual experiment using speech samples synthesized by three models: the Single Utterance Model, the Sequential Utterances Model without UR and UP features, and the Sequential Utterances Model with UR and UP features. The speech samples consisted of four news articles being composed of 9 to 12 utterances. The articles were not included in the training data of the three models. For each article, 12 evaluators were presented 6 pairs of samples selected from 3 samples and asked which sample was appropriate for the news-telling task.

Table 2 lists the differences among three models. The difference between the Single Utterance Model and the Sequential Utterances Models (the Sequential Utterances Model w/o and w/ features) is the training data. The Single Utterance Model is a typical conventional model, in which the phonetically balanced (PB) speech data designed in our previous study [15] were used for building the DNN models. The PB speech data consists of 700 utterances. The content of each utterance was basically independent of the others. We recorded the PB speech data uttered by the same voice actress described in Section 2.

The difference between the Sequential Utterances Models is the system type. The proposed system, which is applied to the Sequential Utterances Model with UR and UP features, is the same as shown in Figure 6. In this perceptual experiment, the

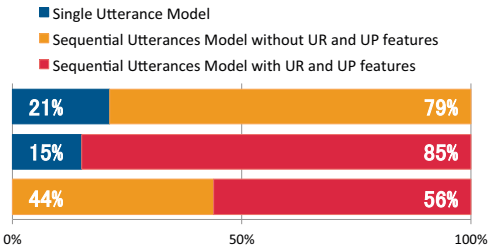


Figure 7: *Preference scores for perceptual experiment.*

UR and UP features consist of three role features (front satellite, nucleus, and rear satellite) represented by the one-hot vector and one position feature (an utterance number from the head in the paragraph). The PB speech data was used for initializing the model parameters of the Sequential Utterances Models before training using the news-telling speech data.

Two types of pauses – long and short – were inserted between the utterances. This is based on the analysis result on the pause distributions between utterances in Figure 5. The duration for the long pause (808 ms) is the average duration of the nucleus-rear satellite pauses. The duration for the short pause (527 ms) is the average of the other four types of pauses. The long pause was applied between the nucleus and rear satellite utterances. The short pause was applied to the other pauses.

Figure 7 shows the preference scores of the perceptual experiment. Each bar represents the preference score for two of the three models. A sign test (at the 95% confidence level) was performed and proved that the Sequential Utterances Models were preferred significantly more than the Single Utterance Model. The model with the UR and UP features was preferred to the model without the UR and UP features, though there was no significant difference in the sign test.

5. Conclusion

A deep neural network (DNN) based speech synthesis system was developed that can control prosodic features of utterances throughout an entire input text. The role and position of each utterance in the text were introduced into the input layer of the DNN as the control parameters for the prosody generation. The results of the analysis on the news-telling speech data indicated that the prosodic features differed depending on the role and position features. Objective and subjective evaluation results showed that the consideration of those features was effective in the speech synthesis for information delivering such as news-telling.

6. References

[1] H. Kashioka, T. Misu, E. Mizukami, Y. Shiga, K. Kayama, C. Hori, and H. Kawai, "Multimodal dialog system for Kyoto sightseeing guide," in *Proc. APSIPA Annual Summit and Conference (ASC 2011)*, Xi'an, China, Oct. 2011.

[2] K. Yoshino and T. Kawahara, "Conversational system for information navigation based on POMDP with user focus tracking," *Computer Speech & Language*, vol. 34, no. 1, pp. 275–291, Nov. 2015.

[3] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating expressive speech for storytelling applications," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1137–1144, Jul. 2006.

[4] R. Montaño, F. Alías, and J. Ferrer, "Prosodic analysis of storytelling discourse modes and narrative situations oriented to text-to-speech synthesis," in *Proc. 8th ISCA Speech Synthesis Workshop*, Barcelona, Spain, Aug. 2013, pp. 171–176.

[5] S. Fujie, I. Fukuoka, A. Mugita, H. Takatsu, Y. Hayashi, and T. Kobayashi, "A spoken dialog system for coordinating information consumption and exploration," in *Proc. ACM Conf. Hum. Inf. Interaction and Retrieval (CHIIR '16)*, Carrboro, USA, Mar. 2016, pp. 253–256.

[6] S. Takeda and A. Ichikawa, "Analysis of prominence in spoken Japanese sentences and application to text-to-speech synthesis," *Speech Commun.*, vol. 14, no. 2, pp. 171–196, Apr. 1994.

[7] K. Morizane, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Emphasized speech synthesis based on hidden Markov models," in *Proc. 2009 Oriental COCOSDA Int. Conf. Speech Database and Assessments*, Beijing, China, Aug. 2009, pp. 76–81.

[8] J. Hirschberg and C. H. Nakatani, "A prosodic analysis of discourse segments in direction-giving monologues," in *Proc. 34th Annual Meeting on Association for Comput. Linguistics (ACL '96)*, Santa Cruz, USA, Jun. 1996, pp. 286–293.

[9] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text-Interdisciplinary J. the Study of Discourse*, vol. 8, no. 3, pp. 243–281, Jan. 1988.

[10] S. Andersson, J. Yamagishi, and R. A. J. Clark, "Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis," *Speech Commun.*, vol. 54, no. 2, pp. 175–188, Feb. 2012.

[11] K. Sugiura, Y. Shiga, H. Kawai, T. Misu, and C. Hori, "Non-monologue HMM-based speech synthesis for service robots: A cloud robotics approach," in *Proc. IEEE Int. Conf. Robotics & Automation (ICRA)*, Hong Kong, China, May 2014, pp. 2237–2242.

[12] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci. & Tech.*, vol. 27, no. 6, pp. 349–353, Nov. 2006.

[13] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2013)*, Vancouver, Canada, May 2013, pp. 7962–7966.

[14] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proc. 9th ISCA Workshop on Speech Synthesis*, Sunnyvale, USA, Sep. 2016, pp. 202–207.

[15] T. Kobayashi and K. Iwata, "Speech synthesis for conversation system," in *IEICE Technical Report*, vol. 114, no. 303, Nov. 2014, pp. 19–24, (in Japanese).