Are you professional?: Analysis of prosodic features between a newscaster and amateur speakers through partial substitution by DNN-TTS

Takuya Ozuru[1], *Yusuke Ijima*[2], *Daisuke Saito*[1], *Nobuaki Minematsu*[1]

[1]Graduate School of Engineering, The University of Tokyo, Japan
[2]NTT Media Intelligence Laboratories, NTT Corporation, Japan

{ozuru,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp, ijima@m.ieice.org

Abstract

This paper analyzes prosodic differences between a professional newscaster and amateur speakers which affect listeners' perceptual impression in Japanese. Speech of professional newscasters easily convey his/her occupation, which is that of a newscaster. Although people perceive many factors from human's speech, it is not revealed what factors are dominant for him/her to be professional. To this end, we conduct a large scale perceptual experiment using synthesized speech by deep neural networks (DNN) based speech synthesis. Speech stimuli are synthesized, in which prosodic features such as phoneme duration or F_0 are partially substituted with those of target speakers by changing a DNN trained from professional and amateur speakers. To exclude the influence of the voice quality, spectral features with the same speaker characteristics were used for the experiment. Listeners are asked to choose one speech which he/she thought that it is more acceptable as speech of a newscaster. The results of the perceptual experiment indicate that listeners' impressions are affected by F_0 rather than phoneme duration, although both features affect the listeners' impressions. We further analyze the relation between the obtained perceptual scores and some prosodic related features. It suggests that the larger the SD of F_0 pattern, the more listeners perceive the speech as professional.

Index Terms: prosodic features, professional newscaster, amateur speakers, phoneme duration, fundamental frequency, speech synthesis

1. Introduction

News read out in daily life should be comprehensible and understandable, and it is one of the goals of speech synthesis in that domain. The profession of newscasters plays an important role to achieve this comprehensibility and understandability. Understanding the speech factors of newscasters are very useful to achieve the goal of speech synthesis in the domain. Although people perceive so many kinds of factors from human's speech, it is not clear what factors are dominant for him/her to be professional. Out of two main factors included in speech, i.e., voice quality and prosody, this study focuses on prosodic features which would be regarded as an acquired speaking ability. Finding such prosodic features would lead to construct a voice training system which can take the comprehensibility of speech into account. In addition, taking the obtained prosodic features into account more effective for text-to-speech synthesis system would yield synthesized speech with better comprehensibility.

One promising approach for finding such perceptual speech factors is to analyze the relation between speech-related features and subjective scores obtained from perceptual experiments. There are many analyses between speech-related features and perceptual scores for instance speaker identification [1], voice quality similarity [2], emotional speech [3], speech perception [4], and paralinguistic information [5]. As for analyzing the speech characteristics of professional newscasters, a variety of approaches have also been proposed [6–8]. According to [6], both voice quality and prosody play important roles to characterize utterances of professinal newscasters. However, it would be difficult to conduct a perceptual experiment for evaluating what factors are important for professional speech since voice quality and prosody are evaluated simultaneously in perceptual experiments. Furthermore, these analyses utilized only a small amount of words or sentences. Therefore the various prosody have not been covered in these analyses although prosodic features are generally influenced by semantic information, i.e., sentences.

In this study, our aim is to analyze prosodic differences between a professional newscaster and amateur speakers which affect listeners' perceptual impression in Japanese. The key to finding such prosodic differences is to analyze a large amount of sentences uttered by professional and amateur speakers. However, it is not generally easy to construct a large scale parallel speech corpus. Furthermore, to analyze the relation between perceptual scores and prosodic features, some prosodic features of speech, i.e., phoneme duration and F_0, should be separated in the evaluation. To address these problems, we conduct a large scale perceptual experiment using the partially substituted synthesized speech by DNN-based speech synthesis [9]. Speech stimuli are synthesized, in which prosodic features such as phoneme duration or F_0 are partially substituted with those of target speakers by changing a DNN trained from professional and amateur speakers. To exclude the influence of the voice quality, the same spectral feature were used for the perceptual experiment. Several prosodic features highly correlated to perceptual scores are found by analysis of the results of the perceptual experiment.

2. Perceptual experiment

We first conducted a large scale perceptual experiment to evaluate listeners' impression between speech of a professional newscaster and that of amateur speakers. Speech stimuli and details of the perceptual evaluation are described below.

2.1. Speech data

For the following experiments, we used speech database uttered by one male professional newscaster and 11 amateur male speakers. The sampling frequency of speech was 22.05 kHz and the quantization bit was 16 bits. To reduce the cost for the perceptual experiment, five amateur speakers were chosen from 11

10.21437/SpeechProsody.2020-188

Table 1: *Four types of speech stimuli used for the perceptual experiment.*

Speech stimuli	**Phoneme duration**	$\boldsymbol{F_0}$
PrPr	Professional	Professional
PrAm	Professional	Amateur
AmPr	Amateur	Professional
AmAm	Amateur	Amateur

amateur speakers on the basis of the difference of mean of F_0 between a professional and amateur speakers. Speech data of a professional newscaster (Pr) are about 300 minutes of news speech. Those of selected amateur speakers (Am1–5) are about 60 minutes of speech reading phonetically balanced sentences.

2.2. Speech stimuli

2.2.1. An overview

To analyze the relation between the perceptual listeners' impression and prosodic features, each prosodic feature of speech, i.e., phoneme duration and F_0, should be separated in the evaluation. In addition, it would be desirable to use speech with the same spectral characteristics to exclude the influence of voice quality. In this experiment, the synthesized speech with the partially substituted prosody (phoneme duration and F_0) generated from DNN-based speech synthesis for each speaker, was employed as speech stimuli. The spectral features with the same speaker characteristics between all stimuli were also generated by DNN. Details of the DNN training and speech stimuli generation are given below.

2.2.2. Setups for DNN-based speech synthesis

We used Japanese speech data uttered from six male speakers described in Sect. 2.1. The training corpus included 2,756 utterances (about 565 minutes) from six speakers. All speech samples were manually labeled with the phoneme segmentations and the accentual information. The WORLD vocoder [10] (D4C edition [11]) was employed, and frame shift was 5 ms.

For generating prosodic features of each speaker, it is desirable to train speaker-dependent DNNs for each speaker. However, it would be difficult to train such DNNs because the amount of training data uttered by amateur speakers is not enough to generate high quality prosody. To avoid this problem, multi-speaker modeling based on speaker codes [9] was adopted. In the speaker code-based DNN, speaker code $\boldsymbol{S}$ is fed to each layer of DNN.

We trained two types of DNNs, a duration model and an acoustic model. As the duration DNN, we used three hidden layers with 64 units. The input was 303 dimensional linguistic features, and the output was 1 dimensional phoneme duration which is normalized by global mean and variance. As the acoustic DNN, we used six hidden layers with 512 units. The input was 303 dimensional linguistic and 2 dimensional phoneme duration related features. Each observation vector consisted of 61 mel-cepstral coefficients, log F_0, five band aperiodicities, their delta and delta-delta features and a voiced/unvoiced binary value. In both DNNs, a ReLU activation function was used. Each DNN was optimized by minimum mean squared error criterion and Adam-based back-propagation algorithm [12]. We applied MLPG [13] to the output acoustic features. We used pre-computed variances from the training data for MLPG.

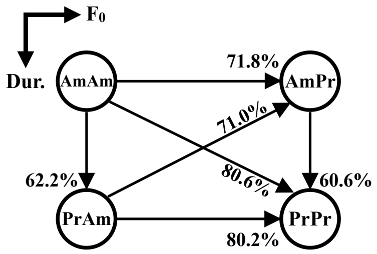

Figure 1: *The results of the perceptual experiment. The horizontal axis indicates substitution of F_0, and the vertical one indicates substitution of phoneme duration. The direction of arrows indicates that which speech stimuli is acceptable as speech of a newscaster in each paired comparison. Scores are mean values obtained from 5 amateur speakers.*

2.2.3. Speech generation for the perceptual experiment

To separately analyze phoneme duration and F_0, we generated partially substituted speech by the trained DNNs. For synthesizing speech stimuli, phoneme duration and/or F_0 were partially substituted with each target speaker by changing the speaker code of DNNs. All combinations between phoneme duration and F_0 are summarized in Table 1.

"PrPr" was synthesized as totally professional utterance. The remaining three utterance types of the phoneme duration or/and F_0 was substituted with those of one amateur speaker. Since we used 5 amateur speakers for the experiment, 16 types of synthesized speech (1 professional speech + (3 substituted speech × 5 amateur speakers)) were generated in total.

To exclude the influence of voice quality, the average-voice-like spectral features were used. Specifically, we set the average value, i.e., $\frac{1}{K}$, K is the number of training speakers, to all dimension of the speaker code $\boldsymbol{S}$, and fed them to the acoustic DNN. It would be expected to generate mel-cepstrum and aperiodicities with average characteristics of the training speakers.

For each utterance type, we synthesized 500 utterances by switching the speaker codes according to the utterance type. These 500 utterances are not included in the training data for DNNs. To avoid the influence of text analysis errors in the experiments, linguistic information, i.e., phoneme and accentual information, was obtained from speech uttered by one female professional newscaster. The duration of each utterance ranged from 3 seconds to 7 seconds.

2.3. Experimental conditions

We conducted large scale AB tests via crowd-sourcing to obtain perceptual scores. Each subject was presented synthesized speech samples and then asked which one was more acceptable as speech of a newscaster. Speech pairs of 10 utterances were randomly chosen from 500 utterances for each subject. All permutations of synthetic speech pairs were presented in both orders (AB and BA) to eliminate bias in the order of stimuli.

To analyze prosodic features which affect listeners' impression, all combinations of four types of partially substituted synthesized speech were evaluated. Since we also conducted evaluations between a professional and 5 amateur speakers, 30 evaluations (6 combinations of speech × 5 amateur speakers) were carried out in total. For each evaluation, subjects were 25 native Japanese speakers. Although we allowed that subjects take part in multiple experiments, duplications were not allowed in the same experiment. Total number of subjects was 750 (30 experiments × 25 subjects) for this perceptual experiment.

2.4. Results

The results of the experiment are shown in Fig. 1. The indicated preference scores are mean values obtained from those of 5 amateur speakers. Although obtained scores for each amateur speaker were varied, the tendency of subjects' selection was not changed by each amateur speaker.

We first compared the difference between professional ("PrPr") and amateur ("AmAm") prosody. It can be seen that more than 80% of subjects selected "PrPr" as acceptable for professional newscasters' speech. It indicates that there is a large gap between prosody of professional and amateur speakers. It suggests that prosody plays an important role in characterization of newscasters' speech.

To analyze the effect of changing phoneme duration, we next compared the score between "PrPr" and "AmPr" and that between "PrAm" and "AmAm". We can see that using professional phoneme duration ("Pr*") have slight better scores than using amateur one. It indicates that phoneme duration affects listeners' impression, but its influence would be small compared with the simultaneous use of phoneme duration and F_0.

As for F_0 differences, we also compared the score between "PrPr" and "PrAm" and that between "AmPr" and "AmAm". It can be seen that using professional F_0 ("*Pr") have better perceptual scores than using amateur one. It indicates that F_0 contour also affects listeners' impression.

Finally, we focused on the score between "PrAm" and "AmPr" to directly compare the effect of changing phoneme duration with that of changing F_0. The result confirms that "AmPr" has better score that "PrAm". This indicates that F_0 contour would be dominant factor to be professional newscasters than phoneme duration.

3. Objective analysis of synthesized speech

In the previous section, we confirmed that F_0 is dominant factor than phoneme duration. However, since the previous section analyzed only global F_0 contour and phoneme duration, we cannot find what factors are dominant in F_0 and phoneme duration. We also analyze the relation between the perceptual scores obtained from each amateur speaker and prosodic differences with respect to a professional newscaster. For the analysis, we used the same 500 utterances used for the perceptual experiment.

3.1. Analysis of F_0

We first analyzed the F_0 of synthesized utterances. Table 2 is the result of six utterance types whose F_0 was synthesized with each speaker's speaker code. RMSEs and correlations are calculated with "PrPr". To exclude the influence of phoneme duration, phoneme durations were produced with the speaker code corresponding with a professional speaker.

There are no large differences in the mean of log F_0 among these six types. This is because we selected the speakers for experiments on the basis of mean of log F_0. Unlike mean of log F_0, the standard deviations (SD) of log F_0 varies for each speaker. Synthesized F_0 contour generated by a professional newscaster ("PrPr") had the largest SD of log F_0 among the six types. The SD and the selected rate of "PrAm1" were respectively 0.214 and 30.8%, which were the largest values among the five amateur speakers. It seems that the selected rate tends to be higher for utterances with larger SD of log F_0. The trend of relation between SD and selected rate was also applied to those of "PrAm2". The RMSE of "PrAm2" was the smallest among five amateur speakers (250.9 cent). In addition, the correlation of "PrAm2" and "PrPr" was the highest among five amateur speakers. However, the selected rate of "PrAm2" was lower than "PrAm1", whose SD was the largest.

3.2. Analysis of phoneme duration

We also analyzed the phoneme duration of synthesized utterances. Each utterance was manually divided into some accentual phrases and phonemes. Table 3 shows the analysis statistics of six types of utterances. In essentially the same manner as F_0 analysis, F_0 were synthesized as a professional newscaster. Phoneme duration was caluculated as an absolute value of time length of each utterance, accentual phrase, or phoneme. The number of utterances, accentual phrases, and phonemes for each utterance type are respectively 500, 4,882, and 37,030.

From the results of utterances and accentual phrases, the mean duration of "PrPr" was faster than those of 4 of 5 amateur speakers. However, speaking rates are not necessarily associated with the selected rate. For example, even though the mean duration of "Am1Pr" and "Am3Pr" are quite different (4.660 s and 5.044 s), the selected rates of them were very close (40.0% and 39.2%). Conversely, the mean duration of "Am3Pr" and "Am4Pr" were very close (5.044 s and 4.995 s), however, the selected rates of them were quite different (39.2% and 25.2%). In contrast, as for phonemes, there are significant differences between RMSE of "Am4Pr", which have the lowest selected rate (25.2%), and those of other 4 speakers.

4. Discussions

From the results of the perceptual experiments and the objective analysis, it is revealed that both of phoneme duration and F_0 contribute to be professional newscasters. Interestingly, the results indicate that F_0 would be much more important factor to be professional than phoneme duration.

The results of the phoneme duration analysis indicate that there are weak association between perceptual scores and speaking rates of utterances and accentual phrases, however, phoneme durations have associations to some extent. In addition, the results of F_0 analysis imply that the SD of F_0 contour is more important than the correlation of F_0, which indicates the similarity of F_0 contours. It means that even if the F_0 contour is similar to that of professional, the listeners' impression would not be affected when the SD of F_0 is small. These results are consistent with the previous analysis [6]. To investigate this hypothesis, we further conducted additional perceptual experiments by changing SDs or patterns of F_0.

4.1. Additional exp. 1 : Effects of changing SD of log F_0

To investigate the effects of changing the SD of log F_0, two types of speech stimuli which differ only in SD of log F_0 were synthesized and evaluated in perceptual experiments.

From the statistics of F_0 of synthesized utterances shown in Table 2, the SD of log F_0 of "PrPr" is the highest among the six utterance types, and that of "PrAm3" is the lowest. Then for each utterance type, we modified the SD of log F_0 into that of "PrPr" or "PrAm3" by applying affine transformation. We describe the former one as "pro" (professional) and the latter one as "ama" (amateur). In order to exclude the effects of differences in mean of log F_0, for all utterance types we modified the mean of log F_0 into that of "PrPr" when applying affine transformation. In this way, we prepared 12 types of utterances (Pr-pro, Pr-ama, Am1-pro, ..., Am5-ama) in total. For example, "Am1-ama" was re-synthesized by applying affine trans-

Table 2: *Statistics of F_0 of synthesized utterances. RMSEs and correlation coefficients are calculated with "PrPr". "Selected rate" means the rate of being selected when the utterance type was compared with "PrPr" in the perceptual experiment.*

Utterance type	Mean of Log F_0	SD of Log F_0	RMSE [cent]	Correlation coefficients	Selected rate [%]
PrPr	4.902	0.272	—	—	—
PrAm1	4.955	0.214	282.7	0.849	30.8
PrAm2	4.907	0.189	250.9	0.875	23.6
PrAm3	4.989	0.136	359.7	0.790	13.2
PrAm4	4.995	0.138	353.1	0.812	19.2
PrAm5	4.933	0.144	330.3	0.798	12.0

Table 3: *Statistics of duration of synthesized utterances. RMSEs are calculated with "PrPr".*

Utterance type	Utterances			Accentual phrases			Phonemes			Selected rate [%]
	Mean [s]	SD [s]	RMSE [s]	Mean [s]	SD [s]	RMSE [s]	Mean [ms]	SD [ms]	RMSE [ms]	
PrPr	4.624	0.918	—	0.730	0.278	—	62.420	29.181	—	—
Am1Pr	4.660	0.926	0.083	0.736	0.294	0.039	62.906	28.874	13.345	40.0
Am2Pr	4.491	0.901	0.154	0.708	0.269	0.042	60.633	29.915	12.678	47.2
Am3Pr	5.044	0.994	0.439	0.799	0.328	0.093	68.099	33.667	14.122	39.2
Am4Pr	4.995	1.004	0.398	0.791	0.323	0.092	67.434	36.560	19.798	25.2
Am5Pr	4.848	0.967	0.242	0.767	0.302	0.054	65.446	29.805	12.802	45.2

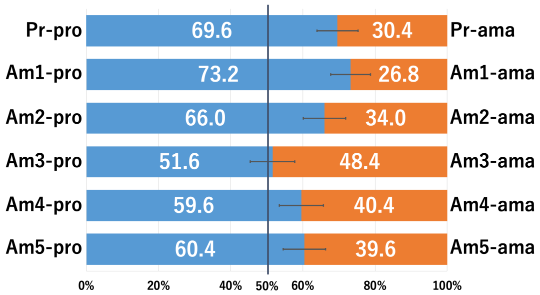

Figure 2: *The results by changing SD of log F_0. Each error bar shows its 95% confidence interval.*

formation to "PrAm1" in order to modify the SD of log F_0 of "PrAm1" into that of "PrAm3".

We conducted six perceptual experiments to compare "pro" and "ama" of all F_0 patterns. They were conducted under the same condition as the experiment described in Sect. 2, however, the data set was reduced from 500 utterances to 125 utterances at random in order to make the comparison condition better.

The results are shown in Fig. 2. Although the comparison between "Am3-pro" and "Am3-ama" did not show a significant difference, all other comparisons showed significant differences.

4.2. Additional exp. 2 : Effects of changing pattern of F_0

To investigate the effects of changing the pattern of F_0, five types of speech stimuli (Am1-pro – Am5-pro) were compared to the professional one (Pr-pro). To exclude the effects of differences in SD of log F_0, the SD of log F_0 of all stimuli were unified as that of "PrPr" (pro). Five perceptual experiments were conducted under the same condition as additional experiment 1.

The results are shown in Fig. 3. Three comparisons showed significant differences.

4.3. Discussion and future works

The results of the additional experiment 1 suggest that enlarging the SD of log F_0 has high correlations with the score of the perceptual experiment. This is consistent with the analysis in Sects. 2 and 3. The largest RMSE and the smallest correlation of "PrAm3" in the result of Table 2 may suggest that there are some accentual errors in the F_0 pattern of Am3. In fact, "Am3-pro" got the worst score in the additional experiment 2 shown as Fig. 3. Enlarging the SD of log F_0 of Am3 may make these errors stand out.

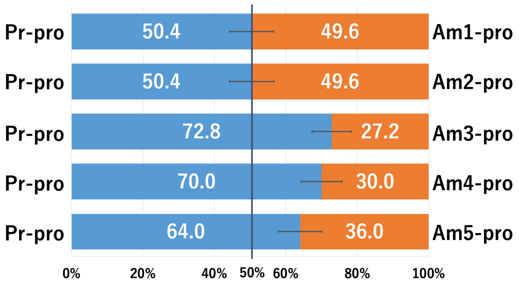

Figure 3: *The results by changing pattern of F_0. Each error bar shows its 95% confidence interval.*

So far, we have classified the speakers as "professional" or "amateur". However, the results of perceptual experiments suggest that there is a great individual difference in speaking skill among the amateur speakers. We are going to conduct large scale perceptual experiments to evaluate natural speech of many amateur speakers as future works.

5. Conclusion

In this paper, we have conducted a large scale subjective perceptual experiment to evaluate listeners' impression between speech of a professional newscaster and that of amateur speakers. The results of the subjective perceptual experiment have shown that using F_0 pattern of the professional newscaster had the biggest impact on listeners' impression. In order to clarify the relation between acoustic features and listeners' impression, we have analyzed speech stimuli which were used in the perceptual experiment. The results of the objective analysis and the additional experiments have shown that the SD of log F_0 had high correlations with the results of the perceptual experiment.

6. References

[1] N. Higuchi and M. Hashimoto, “Analysis of acoustic features affecting speaker identification,” *Journal of the Acoustical Society of Japan (E)*, vol. 17, no. 1, pp. 33–35, 1996.

[2] Y. Ijima, M. Isogai, and H. Mizuno, “Correlation Analysis of Acoustic Features with Perceptual Voice Quality Similarity for Similar Speaker Selection,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[3] S. Takeda, Y. Yasuda, R. Isobe, S. Kiryu, and M. Tsuru, “Analysis of Voice-Quality Features of Speech that Expresses “Anger”, “Joy”, and “Sadness” Uttered by Radio Actors and Actresses,” in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[4] I. Grichkovtsova, M. Morel, and A. Lacheret, “The role of voice quality and prosodic contour in affective speech perception,” *Speech Communication*, vol. 54, no. 3, pp. 414–429, 2012.

[5] Y. Greenberg, N. Shibuya, M. Tsuzaki, H. Kato, and Y. Sagisaka, “Analysis on paralinguistic prosody control in perceptual impression space using multiple dimensional scaling,” *Speech Communication*, vol. 51, no. 7, pp. 585–593, 2009.

[6] H. Kuwabara and K. Ohgushi, “Acoustic Characteristics of Professional Male Announcers’ Speech Sounds,” *Acta Acustica united with Acustica*, vol. 55, no. 4, pp. 233–240, 1984.

[7] M. Finkelstein and O. Amir, “Speaking Rate among Professional Radio Newscasters: Hebrew Speakers,” *Studies in Media and Communication*, vol. 1, no. 1, pp. 131–139, 2013.

[8] E. Strangert, “Prosody in Public Speech: Analyses of a News Announcement and a Political Interview,” in *Ninth European Conference on Speech Communication and Technology*, 2005.

[9] N. Hojo, Y. Ijima, and H. Mizuno, “DNN-Based Speech Synthesis Using Speaker Codes,” *IEICE TRANSACTIONS on Information and Systems*, vol. 101, no. 2, pp. 462–472, 2018.

[10] 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.

[11] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, vol. 84, pp. 57–65, 2016.

[12] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[13] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, vol. 3, pp. 1315–1318, 2000.