Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

A Study of the Influence of the Reverberation Time in Speech Synthesis

Takashi MATSUBARA (1), Masanori MORISE (2) and Takanobu NISHIURA (2)

(1) Graduate School of Science and Engineering, Ritsumeikan University 1-1-1 Noji Higashi, Kusatsu, Shiga 525-8577, Japan
(2) College of Information Science and Engineering, Ritsumeikan University 1-1-1 Noji Higashi, Kusatsu, Shiga 525-8577, Japan

PACS: 43.72.Ja

ABSTRACT

Vocoder-based methods to extract speech parameters (fundamental frequency and spectral envelope) have been proposed to synthesize natural speech based on these parameters. Although these methods can manipulate the parameters flexibly, the sound quality of the resulting speech is not sufficient for practical use. The STRAIGHT and TANDEM-STRAIGHT methods have been proposed to manipulate speech parameters flexibly and to synthesize high-quality speech. These methods require high-SNR speech signal to synthesize speech with high quality. In conventional studies, the speech segments are recorded in an anechoic room, a sound proof room, and a recording studio. In our study, we focus on the influence of reverberation time in the recording environment on the sound quality of the synthesized speech. The relationship between the two is observed in a subjective experiment. Impulse responses with various reverberation times were applied to all segments, and these segments were then processed by TANDEM-STRAIGHT. The synthesized speech segments were used as the stimuli. The kind of employed impulse responses was four. We used the mean opinion score (MOS) to conduct the experiment. The results of quality indicate that the quality of the synthesized speech will be lower than the original regardless of the length of the reverberation time. As a result of reverberation showed TANDEM-STRAIGHT can be re-synthesised without losing a reverberation of the source of speech signal.

INTRODUCTION

In recent years, the demand for flexible, high-quality speech manipulation has been expanded. Conventional vocoder-based methods to extract speech parameters (Fundamental frequency and spectral envelope) have been proposed to synthesize natural speech based on these parameters. Although these methods can manipulate the parameters flexibly, the sound quality of the resulting speech is not sufficient for practical use. The STRAIGHT [1] and TANDEM-STRAIGHT [2] methods have been proposed to manipulate speech parameters flexibly and to synthesize high-quality speech signal. These methods require high-SNR speech signal to synthesize high quality speech signal. In conventional studies, the speech segments are recorded in an anechoic room, a sound proof room, and a recording studio. However, the speech manipulation system should develop if the speech synthesis based on the speech signal recorded by a general room is achieved.

In our study, we focus on the influence of reverberation time in the recording environment on the sound quality of the synthesized speech. The speech segments used for experiment were recorded in an anechoic room. Impulse responses with various reverberation times were applied to all segments, and these segments were then processed by TANDEM-STRAIGHT. The synthesized speech segments were used as the stimuli. All speech segments were comprised of the five Japanese vowels (/aiueo/) by a total of six speakers (three females and three males). These segments were sampled at 44.1 kHz with 16 bit resolution. The employed reverberation time were 100 msec (the sound proof room), 400 msec (the japanese-style room), 600 msec (the corridor), and 900 msec (standard stairs). The relationship between the quality and reverberation time was observed in the subjective experiment. We used the mean opinion score (MOS) [3] [4] to conduct the experiment. Subjects were asked to evaluate the sound quality and a reverberation-perception of the reproduced stimulus from grade 1 to 5. Grade 1 means bad, grade 5 means excellent for quality and grade 1 means reverberant, grade 5 means anechoic for reverberation.

The following sections begin with explanation of STRAIGHT and TANDEM-STRAIGHT as the speech synthesis methods in this paper. Next, the method and the stimuli of experiment were represented. Finally, it explained the result of each experiment, and we discussed them.

METHOD OF SPEECH SYNTHESIS

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [1] is high-quality speech analysis/synthesis method. The speech analysis/synthesis methods, such as Channel VOCODER [5] should separate fundamental frequency and spectral envelope from speech. STRAIGHT enabled to synthesis high-quality speech because these systems can separate fundamental frequency and spectral envelope completely. As STRAIGHT consisted of ad-hoc methods, improvement and expansion of STRAIGHT was difficult. However, TANDEM-STRAIGHT that applied TANDEM (Temporally Aligned, Non-Dispersive Envelope Measurement windows) [2] solved this problem with an algorithm on a simple theory. TANDEM-STRAIGHT enabled to synthesize speech as well as STRAIGHT. In this paper, we employed TANDEM-STRAIGHT as high-quality speech analysis/synthesis method. These methods require high-SNR speech to synthesize high quality speech.

REVERBERATION TIME

Reverberation time is a parameter that expresses the duration of sound after the original sound is removed. It is the time required for a sound to decay by 60 dB in a room and the time is defined as T_{60}. One of the methods to measure reverberation time is developed by M. R. Schroeder as integrating the square of the impulse response [6]. The reverberation curves are derived from Eq. 1 with impulse response $h(t)$.

$$\left\langle y_d^2(t) \right\rangle = N\int_t^{\infty} h^2(\lambda)d\lambda, \qquad (1)$$

where < > is the ensemble average, and N is the power of the unit frequency of random noise. $h^2(\lambda)$ is power spectrum. The reverberation time is the time it takes to drop 60 dB blow the original level.

As previously indicated, the experiments were employed speech signals that does not include reverberant and noise to improve the accuracy of the analysis [7]. As it was not clear that the relationship between synthesized speech quality and reverberation. Therefore, we focused on the length of reverberation. The speech signals for evaluation were generated by impulse responses that were recorded at real environment. Then these reverberant speech signals were synthesized by TANDEM-STRAIGHT. These two kinds of speech signals were employed to evaluation. These speech signals were evaluated based on quality and reverberation. Then we examined the relationship between quality and reverberation time of re-synthesized speech.

EXPERIMENT

Evaluation method

The evaluation experiment was conducted by Mean Opinion Score (MOS). This method is employed as the measurement of the quality of speech signal. The subjects have normal hearing ability in the experiment. Experimental conditions were shown in Tab. 1

Table 1. Experimental conditions

Listening environment	
Environment to listening	Sound proof room (19.0 dBA)
Headphone	SONY MDR-CD900ST
Number of subjects	7
Speech stimuli	
Number of stimuli	48
Number of impulse responses	4
Number of speakers	6
Sampling frequency	44.1 kHz
Resolution	16 bit

Evaluation of speech quality

They listened to a stimulus, then they evaluated stimulus within 2.0 seconds and answered by five-grade evaluation on check sheet shown in Fig. 1. Grade 1 means bad, and grade 5 means excellent. Table 2 shows the allocation of MOS value and opinion for quality.

Evaluation of reverberation

They listened to a stimulus, then they evaluated stimulus within 2.0 seconds and answered by five-grade evaluation on check sheet shown in Fig. 2. Grade 1 means reverberant, and grade 5 means anechoic. Table 3 shows the allocation of MOS value and opinion for reverberation.

Table 2. Allocation of MOS value and opinion for quality

Score	*Quality of the speech*
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

	Bad	Poor	Fair	Good	Excellent
Num	1	2	3	4	5
1	□	□	□	□	□
2	□	□	□	□	□

Figure 1. Check sheet for quality

Table 3. Allocation of MOS value and opinion for reverberation

Score	*Reverberation of the speech*
5	Anechoic
4	-
3	-
2	-
1	Reverberant

	Reverberant				Anechoic
Num	1	2	3	4	5
1	□	□	□	□	□
2	□	□	□	□	□

Figure 2. Check sheet for reverberation

Speech used for the experiment

All speech stimuli were generated based on five Japanese vowels (/aiueo/) by total six speakers (three females and three males). These speech signals were sampled at 44.1 kHz with 16 bits resolution. The speech signals employed for the experiment were made with four different impulse responses. The impulse responses were 100 msec, 400 msec, 600 msec, and 900 msec. Table 4 shows the kind of impulse responses to add reverberation for speech signals used for the experiment.

Table 4. The kind of impulse responses

Reverberation time	*Measured environment*
T_{60}=100 msec	Sound proof room
T_{60}=400 msec	Japanese-style room
T_{60}=600 msec	Corridor
T_{60}=900 msec	Standard stairs

Speech for evaluation

Eight kinds of conditions were used for the experiment.

- RT100_ORG
- RT100_TS
- RT400_ORG
- RT400_TS
- RT600_ORG
- RT600_TS
- RT900_ORG
- RT900_TS

RT stands for Reverberation Time. Three-digit numbers represents reverberation time (100 msec, 400 msec, 600 msec, and 900 msec). ORG represents original speech signal and TS represents synthesized speech signal by TANDEM-STRAIGHT. In 48 different speech signals for evaluation were generated by the six speaker's speech signals. All stimuli are evaluated in twice. The total number of stimuli is 96 for each evaluation (quality and reverberation).

EXPERIMENTAL RESULT

Results of quality

Figures 3, 4 and 5 illustrate the experimental results in the sound quality. The horizontal axis represents the value of the reverberation time and the vertical axis represents MOS about quality. The higher scores indicate that listener sensed a high-quality speech. Figure 3 shows the result of the average of all speakers, Fig. 4 shows the result of all female speakers and Fig. 5 shows the result of all male speakers. Tables 5, 6 and 7 show the numerical value corresponding to the experimental results in Figs. 3, 4 and 5. MOS of TS is lower than MOS of ORG in all results.

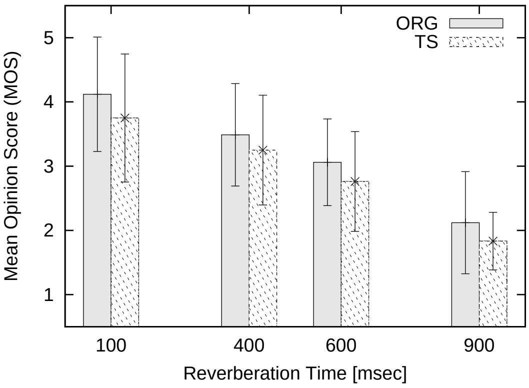

Figure 3. The results of quality (all speakers)

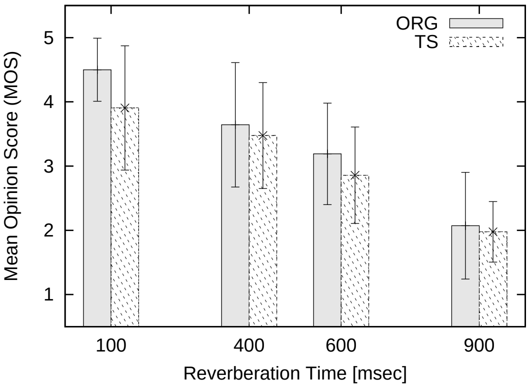

Figure 4. The results of quality (female speakers)

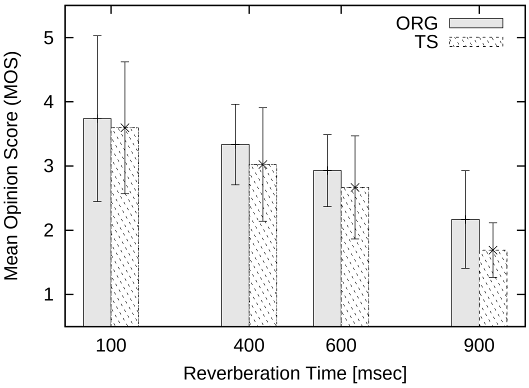

Figure 5. The results of quality (male speakers)

Table 5. Average and variance of quality (all speakers)

	Average		Variance	
T_{60}[msec]	ORG	TS	ORG	TS
100	4.12	3.75	0.89	1.00
400	3.49	3.25	0.80	0.85
600	3.06	2.76	0.68	0.78
900	2.12	1.83	0.80	0.45

Table 6. Average and variance of quality (female speakers)

	Average		Variance	
T_{60}[msec]	ORG	TS	ORG	TS
100	4.50	3.90	0.49	0.97
400	3.64	3.48	0.97	0.82
600	3.19	2.86	0.79	0.75
900	2.07	1.98	0.83	0.47

Table 7. Average and variance of quality (male speakers)

	Average		Variance	
T_{60}[msec]	ORG	TS	ORG	TS
100	3.74	3.60	1.29	1.03
400	3.33	3.02	0.63	0.88
600	2.93	2.67	0.56	0.80
900	2.17	1.69	0.76	0.42

Results of reverberation

Figures 6, 7 and 8 illustrate the experimental results for the reverberation. The horizontal axis represents the value of the reverberation time and the vertical axis represents MOS about reverberation. The lower scores indicate that the listeners sense higher reverberation. Figure 6 shows the result of all speakers, Fig. 7 shows the result of all female speakers and Fig. 8 shows the result of all male speakers. Tables 8, 9 and 10 show the numerical value corresponding to the experimental results in Figs. 6, 7 and 8. As the results of all speakers, the reverberation time of TS was felt shorter slightly than that of ORG.

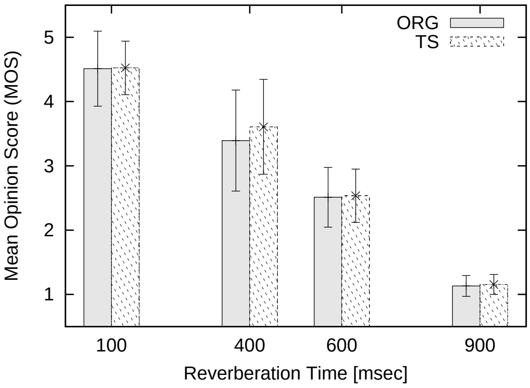

Figure 6. The results of reverberation (all speakers)

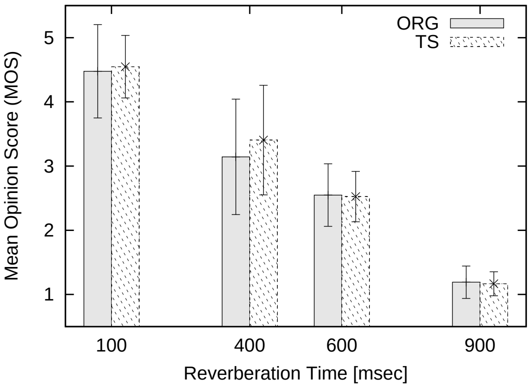

Figure 7. The results of reverberation (female speakers)

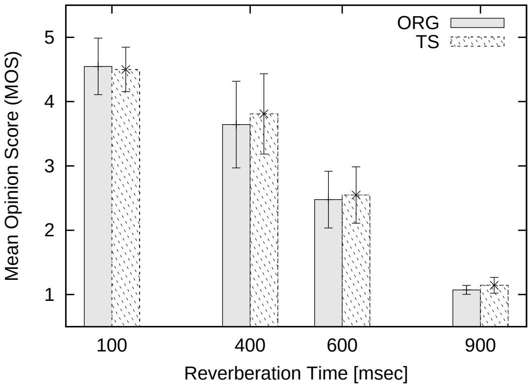

Figure 8. The results of reverberation (male speakers)

Table 8. Average and variance of reverberation (all speakers)

	Average		Variance	
T_{60}[msec]	ORG	TS	ORG	TS
100	4.51	4.52	0.58	0.42
400	3.39	3.61	0.79	0.74
600	2.51	2.54	0.46	0.42
900	1.13	1.15	0.16	0.15

Table 9. Average and variance of reverberation (female speakers)

	Average		Variance	
T_{60}[msec]	ORG	TS	ORG	TS
100	4.48	4.55	0.73	0.49
400	3.14	3.40	0.90	0.85
600	2.55	2.52	0.49	0.39
900	1.19	1.17	0.25	0.19

Table 10. Average and variance of reverberation (male speakers)

	Average		Variance	
T_{60}[msec]	ORG	TS	ORG	TS
100	4.55	4.50	0.44	0.35
400	3.64	3.81	0.67	0.62
600	2.48	2.55	0.44	0.44
900	1.07	1.14	0.07	0.12

Discussion

First, the results of quality were discussed. The all results of quality represent MOS of TS were lower than MOS of ORG. The differences of MOS between ORG and TS were nearly equal in any reverberation times for the results of all speakers. As the results of female speakers, the qualities of ORG and TS come to feel it comparable, as the reverberation time becomes longer. But in the case of male speakers, as the reverberation time was longer, the differences between ORG and TS were larger.

In the next, we discuss the results of reverberation. As the results of all speakers, the reverberation time of TS was felt shorter than the reverberation time of ORG. But we confirmed that TANDEM-STRAIGHT can synthesize almost equally reverberation of original speech signals, because the differences between ORG and TS are slightly. Variances of RT900 were lower than variances of other reverberation time for all speakers.

CONCLUSIONS

We examined the relationship between reverberation time and quality of speech signal by subjective evaluation. Two kinds of experiments were conducted. One was the subjective evaluation about quality of speech. The other was the subjective evaluation about perception of reverberation. The experimental results of quality indicate that the quality of the synthesized speech will be lower than the original regardless of the reverberation time. As the result of reverberation, TANDEM-STRAIGHT can synthesize speech without losing a reverberation characteristic of source speech.

As experiments and researches are better to use shorter reverberant speech that recorded in soundproof room and anechoic room the same as before. However, in the case of using demonstrations and entertainment, the quality of synthesized speech does not affect though using a speech recorded in typical living room. In future work we will attempt to examine the relationship between speech signal quality and background noise as a cause of reduction of the analysis performance.

ACKNOWLEDGEMENT

This work was partly supported by Grants-in-Aid for Scientific Research funded by The Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES

1 Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alainde Cheveigne. “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction” *Speech Communication, Vol.27*, No.3-4, pp.187-207 (1999).
2 H. Kawahara, M. Morise, T. Takahasni, R. Nisimura, T. Irino, and H. Banno. “TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation” In Proc. *ICASSP2008*, pp.3933-3936 (2008).
3 ITU-T Recommendation, *Methods for subjective determination of transmission quality*. p.800 (1996).
4 ITU-T Recommendation, *Mean Opinion Score (MOS) terminology*. p.800.1 (1996).
5 H. Dudley, “Remaking speech.” *J. Acoust. Soc. Am., Vol. 11, No. 2*, pp.169-177 (1939).
6 M. R. Schroeder,”New Method of Measuring Reverberation Time” *JASA, Vol.37*, pp.409-412 (1965).
7 Tatsuya Kitamura, Takeshi Saitou, “Effects of acoustic modification on perception of speaker characteristics for sustained vowels” *Acoustical Science and Technology, Vol.28, No.6*, pp.434-437 (2007).