DIFFERENCES IN THE EFFECT OF TIME-EXPANDED AND TIME-CONTRACTED SPEECH ON INTELLIGIBILITY BY PHONETIC FEATURE

Toru Shibuya [1), 2)] *, Yosuke Kobayashi* [1)]*, Hitomi Watanabe* [1)]*, and Kazuhiro Kondo* [1)]

[1)] Graduate School of Science and Engineering, Yamagata University
[2)] NEC Communication Systems, Ltd.

ABSTRACT

We have been investigating the effect of speech rate alteration on Japanese speech intelligibility, especially the differences by phonetic feature. We evaluated this difference in intelligibility using the Japanese Diagnostic Rhyme Test (DRT) on artificially speed-altered speech, such as time-expanded speech (1.6 times the original length, or 60% time expansion) and time-contracted speech (0.6 times the original length). Artificial speaking rate alteration was shown to have some effect or degradation on the Japanese speech intelligibility depending on the phonetic feature, initial consonant feature, and succeeding vowel context. Nasal and unvoiced plosives were relatively not affected by speed alteration. Syllables with vowel context /i/ showed 10% higher intelligibility when time-expanded. These results suggest guidelines for feature-dependent intelligible speed alteration methods.

Index Terms— *Speech rate alteration, Japanese speech intelligibility, DRT (Diagnostic Rhyme Test), Phonetic feature, Speech recognition*

1. INTRODUCTION

Broadcast speech is often hard to understand when the speaking rate is high. Speech rate control is often used to improve intelligibility. Such speech rate conversion technologies have been studied for elderly people, and have been shown to be effective [1]. However, in the real-time processing of broadcast signals, there are difficulties in the time synchronization between video and speaking rate-reduced speech. Solutions for this problem were studied, such as inter-word pause control [2], and adaptive speech rate control method that enables rate reduction in the first portions and gradual speed increase in the latter portions [3]. These methods enabled the reduction of speech delivery speed while maintaining the voice pitch and quality without lengthening the entire program duration. Other attempts at intelligibility improvement without speech rate alteration are the steady-state suppression method [4], which improved the intelligibility for elderly people in reverberant noise.

Nejime et al. developed a real-time portable device that slows the speaking rate without changing the pitch, and applied this to hearing aids [2,5]. They also showed that speech-rate slowing would not improve the intelligibility of speech in noise for hearing-impaired people who have cochlear damage [6]. Other evaluation results of the time-expanded speech of consonant in monosyllables were studied, and some intelligibility improvement in nasal, and semivowel were shown [7].

On the other hand, time-contracted speech, *i.e.*, quick speech was studied for fast and intelligible text-to-speech application for visually impaired person (*e.g.* screen readers) [8]. These studies aim for ultra-high speech speed rate, such as 20 mora/sec.

In our discussion, we focus our attention on the different effects of speech rate on Japanese speech intelligibility by various phonetic features for users with normal hearing. In this paper, we first explain the experimental conditions for the DRT intelligibility tests using three artificially speed-altered speech. Time-expanded speech (60% time expansion, "1.6x-slow"), time-contracted speech (time contraction to 60% of the original length, "0.6x-quick"), and the original speed were tested. Six phonetic features of the initial consonant, as well as the succeeding vowel context were compared. Detailed discussions about the effect of speech speed control by phonetic feature and vowel context is given. Possible method for adaptive speech speed alteration on slow and quick browsing is given.

2. EXPERIMENTAL CONDITIONS FOR THE SUBJECTIVE SPEECH INTELLIGIBILITY TEST

2.1. DRT (Diagnostic Rhyme Test) and DRT word-pairs

The DRT is one of the well-known speech intelligibility tests for subjective quality measurement. DRT uses word-pairs that are different only by one initial consonant as the evaluation stimuli. DRT is an intelligibility test method with two-to-one forced selection in which the subject hears one word, and is presented a choice of two rhyming words from which the subject must choose one. The initial consonants in the DRT word-pairs are arranged to evaluate the intelligibility of one of the six phonetic features; voicing, nasality, sibilation, sustention, graveness, and compactness [9]. In this study, we chose the DRT to measure the subjective word speech intelligibility of the speed-altered speech. We also plan to measure the sentence intelligibility using other testing methods.

2.2. Artificial speed alteration

Artificially speed-altered speech signals were created using Praat [10] which is a free software for audio manipulation. Praat uses the classic Pitch Synchronous Overlap Add (PSOLA) method to alter the length of speech samples without changing the pitch. All 120 words in the DRT set were processed.

In this evaluation, four different speeds were used. In the artificially time-contracted speech, denoted as "0.6x-quick," original speech is sped up to 0.6 times the length of the original speech. The time-expanded speech, denoted as "1.6x-slow," is

slowed to 1.6 times the original speech. The "1.0x-altered" sample is the same length as the original but resynthesized using PSOLA. Finally, we also included the unaltered (without PSOLA) speech for reference, denoted "original".

2.3. DRT conditions

To evaluate the word intelligibility under the noise environment, babble noise (multi-talker noise) [11] was mixed with speech. The competing noise levels are set to 0, 4, 8, 12 and 16 dB relative to the signal level, and also no-noise.

The DRT words were presented to the subjects using headphones. Each subject adjusted the level to their comfortable listening levels. The subjects were six males, all in their early twenties with normal hearing. The speaker of the test speech was female, and average speaking rate of the DRT words (2 mora for each word) was 6.1 mora/sec.

3. EVALUATION RESULTS

3.1. Intelligibility by phonetic features

Figure 1 shows the noise level (0 to 16dB relative to the signal level, and no-noise) vs. speech intelligibility as the average of CACRR (chance-adjusted correct response rate) over all phonetic features, with the four types of speed–altered speech, *i.e.*, "0.6x-quick", "1.0x-altered", "1.6x-slow" and "original". As shown, with the "0.6x-quick," some intelligibility degradation compared to other speeds at almost all noise levels tested were seen.

Table 1 shows the summary of the intelligibility difference between speed-altered speech and the "1.0x-altered" speech by the six phonetic feature, where statistical significance of 1% was confirmed with the analysis of variance (ANOVA). Additionally, the difference between "1.0x-altered" and "original" was not significant in all analyses mentioned below.

According to Table 1, in the "1.6x-slow" speed, some intelligibility improvements were seen with "Graveness". In the "0.6x-quick" speed, significant intelligibility degradations were seen with some phonetic features, while some phonetic features were not affected by speed alteration.

These observations imply the possibility of a speech rate increase without degradation in intelligibility, such as for quick browsing, and also possibility of speech rate decrease for improved intelligibility, such as for hearing improvement.

3.2. Intelligibility by initial consonant phonetic features

Figure 2 shows the noise level vs. speech intelligibility among speed-altered speech, as the average CACRR difference between "1.0x-altered" and "1.6x-slow" speech, and the difference between "1.0x-altered" and "0.6x-quick" speech.

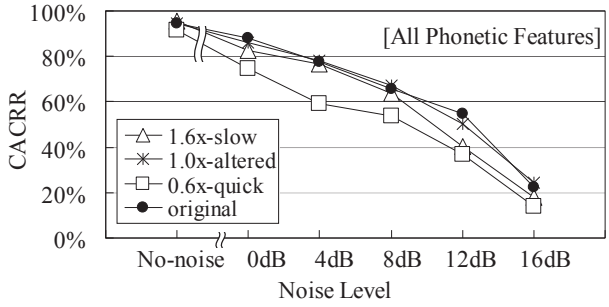


Figure 1. DRT result of over all phonetic features

Table 1. Summary of the intelligibility difference

Phonetic features	Word example	1.6x-slow	0.6x-quick
Voicing	zai-sai	0	- -
Nasality	man-ban	0	0
Sustention	hashi-kashi	- -	- -
Sibilation	jyamu-gamu	0	-
Graveness	waku-raku	+	0
Compactness	piza-kiza	0	- -

+ +: intelligibility improvement p <1%, +: p <5%

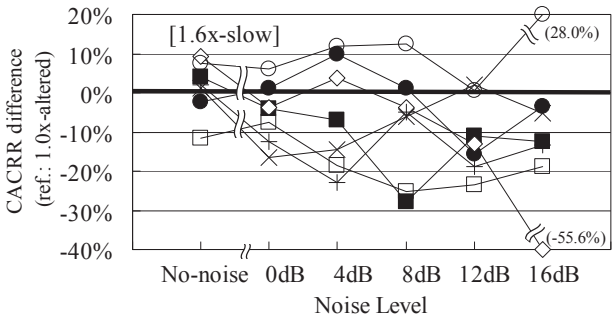
0 : no difference

- -: intelligibirity deterioration p <1%, -: p <5%

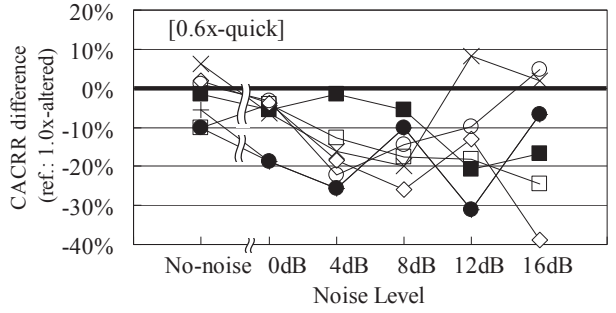
The seven consonant features shown here were reclassified from the six phonetic features defined in the DRT.

Figure 3 also shows the difference of the noise-averaged CACRR by the reclassified seven consonant features. The statistical significance level of 1% was confirmed for the marked conditions using ANOVA. The difference between "1.0x-altered" and "original" was again not significant.

As can be read in these figures, in the "1.6x-slow" speed, unvoiced plosives show intelligibility improvement, but the others did not. In the "0.6x-quick" speed, significant intelligibility degradations were seen in some consonants (*e.g.* voiced plosives, semi-vowels, and liquids), while some consonants were relatively not affected by speed alteration (*e.g.* nasals, voiced affricates). Also, significant degradation in both time-expanded speech and



(a) 1.6x slow speech



(b) 0.6x quick speech

-×- Nasal, -■- Voiced affricate, -●- Voiced plosives,
-□- Semivowel and liquid, -□- Unvoiced affricate,
-○- Unvoiced plosives, -+-Glide

Figure 2. Average CACRR difference for 1.6x-slow speech and 0.6x-quick speech

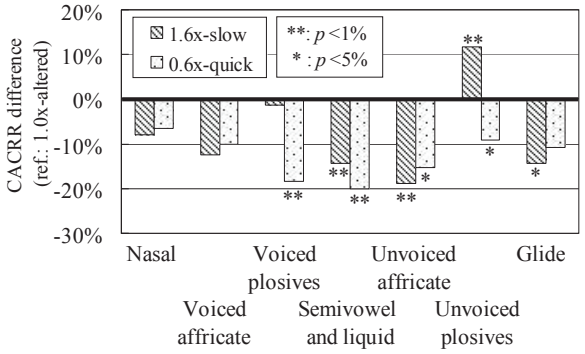


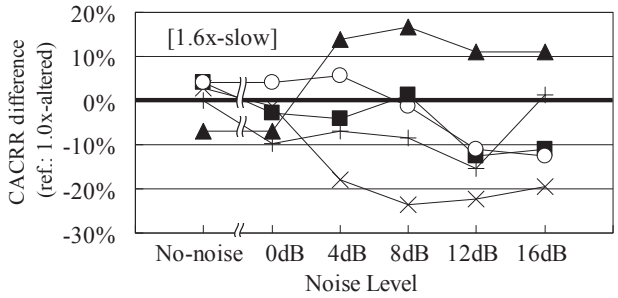
Figure 3. Noise-averaged CACRR by seven consonant features

time-contracted speech were seen with some consonants (*e.g.* semi-vowels, liquids, unvoiced affricates). These observations imply the possibility of a speech rate control using carefully selected consonant feature to alter speed without impacting the intelligibility.

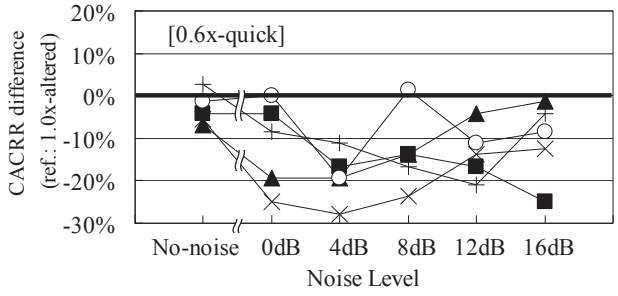
3.3. Intelligibility by succeeding vowel context

DRT word contains all (five) Japanese vowel context equally, *i.e.* 24 words for each vowel context. Figure 4 shows the noise level vs. speech intelligibility as the average CACRR difference between "1.0x-altered" and the speed-altered speech by the five vowel contexts.

Figure 5 also shows the difference of the noise-averaged CACRR by vowel context, with significance level of 1% by ANOVA. According to these figures, these observations were made:



(a) 1.6x slow speech



(b) 0.6x quick speech

-■-: /a/, -▲-: /i/, -×-: /u/, -+-: /e/, -○-: /o/

Figure 4. Average CACRR difference in 1.6x-slow speech and 0.6x-quick speech

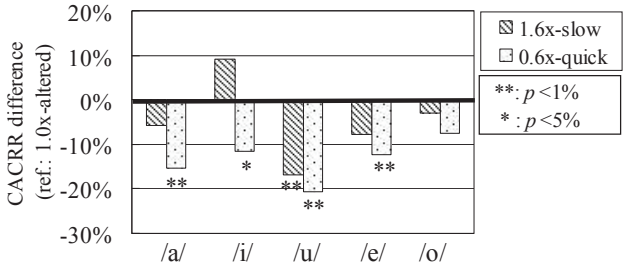


Figure 5. Noise-averaged CACRR by five vowel features

- The vowel context /u/ shows significant degradation in both speeds.
- In the "0.6x-quick" speech, context /o/ was not significantly affected while the other contexts were.
- In the "1.6x-slow", context /i/ show significant improvement.
- These observations again imply the possibility of a speech rate control without degradation by selecting the vowel context.

4. DISCUSSION

4.1 Comparison to previous studies

According to Table 1, the "1.6x-slow" speech shows no significant difference compared to "1.0x-artered" or "original" speech. This is the same as was shown in the prior research [6]. However, if we closely examine the results by phonetic features, there are some improvements in unvoiced plosives and vowel context /i/ as shown in Figure 3 and 5.

In [7], with the consonant time-expansion, there were intelligibility improvements in nasals, and semivowels. We also observed this, as well as improvements in unvoiced plosives in our evaluation.

4.2. The effect of speech alteration on unvoiced plosives

The unvoiced plosives were shown to have higher intelligibility by time expansion. Figure 6 shows the detailed result of unvoiced plosives divided into each consonant, /p/, /t/ and /k/. The /p/ and /t/ have significant improvement in "1.6x-slow" while /k/ did not. These results were confirmed with ANOVA. Furthermore, in /p/, the CACRR difference between "1.6x-slow" and "0.6x-quick" was significant, with p=0.0003<1%.

The DRT word "piza" is a combination of /p/ and vowel context /i/, which both were shown to improve intelligibility by time-expansion, and is a word included in the DRT. Figure 7 shows the waveform and formant of "piza". The /p/ is a consonant that has relatively rich low frequency components, and the /i/ is a palatal sound that has low F1 and high F2 compared to other vowels. Thus, the time length of a formant transition in /p/ to /i/ is stable, making time expansion enhance this transition to help its identification. This may be one of the scenarios in which time-extended speech may have effect on the intelligibility of palatal phonemes.

4.3. Suggested guidelines for an adaptive speech rate control method

Our focus is to come up with some implications for adaptive speech alteration method which preserves or enhances speech intelligibility. Table 2 lists some suggestions for adaptive speech

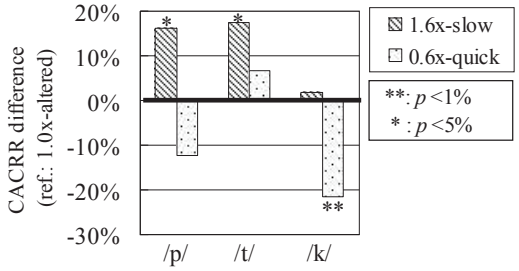


Figure 6. Noise-averaged CACRR for unvoiced plosives

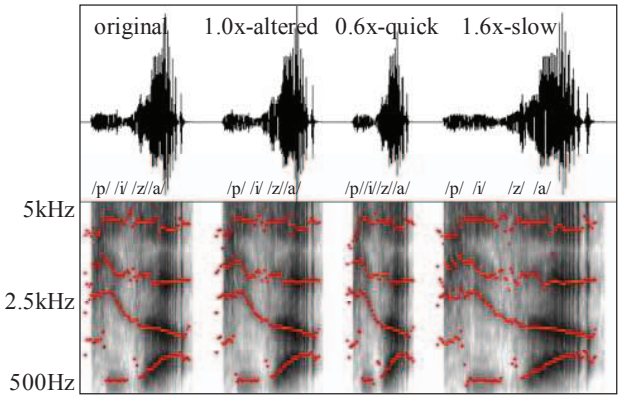


Figure 7. Waveforms and formants of "piza" at various speeds

Table 2. Guidelines for adaptive speech rate control

Initial Consonant	Succeeding vowel									
	Time expansion					Time contraction				
	/a/	/i/	/u/	/e/	/o/	/a/	/i/	/u/	/e/	/o/
/m/, /n/		+				-	-	-	-	-
/z/, /dz/		+				-	-	-	-	-
/b/, /d/, /g/		+				0		0	0	
/y/, /w/, /r/	0	+	0	0	0	0		0	0	
/s/, /ts/	0	+	0	0	0	0		0	0	
/p/, /t/, /k/	+	+	+	+	+	-	-	-	-	-
/h/	0	+	0	0	0	-	-	-	-	-

+: expand, 0: no-expansion / contraction, -: contract
non-marked: irrelevant

rate control using the phonetic features of consonant and vowel context.

By following these guidelines, improvement in the intelligibility by time-expansion for hearing aids should be possible. Likewise, time-contraction for quick browsing without severely degrading the intelligibility should also be possible.

5. CONCLUSION

We have been investigating the effect of speech rate on Japanese speech intelligibility by phonetic feature. We evaluated this difference in intelligibility using the Japanese DRT on artificially speed-altered speech.

According to the results, some intelligibility improvements or degradation were seen by artificial speech rate alteration. The speech intelligibility improves with time-expanded speech in unvoiced plosives and vowel context /i/.

In the time-contracted speech, significant intelligibility degradations in most consonant features and vowel features were seen. However, there were no significant degradation in nasals, voiced affricates and syllables with vowel context /o/.

These results suggest guidelines for feature-dependent intelligible speed alteration methods. They also imply the possibility of a speech rate increase without degradation in intelligibility, such as for quick browsing. There is also possibility of speech rate decrease for improved intelligibility such as for hearing improvement for elderly people and/or hearing aids. However, this needs testing using hearing-impaired patients.

These implications are for word intelligibility only. We obviously need confirmation with sentence intelligibility, and further testing is necessary with read sentence speech as test material. We plan these in the near future.

Finally, we would like to implement the suggested phonetic feature-adaptive speed alteration, and measure its effect on intelligibility.

6. REFERENCES

[1] A Imai, "Speech rate conversion technology and its applications," *IEICE Tech. Rep.,* CQ2007-29, July 2007 (in Japanese).

[2] Y. Nejime, T. Aritsuka, T. Imamura, T. Ifukube and J. Matsushima, "A portable digital speech-rate converter for hearing impairment," *IEEE Trans. Rehabilitation Engineering*, vol.4, No.2, June 1996.

[3] A Imai, N. Seiyama, T. Takagi, and E. Miyasaka, "Evaluation of speech rate conversion for elderly people," *Proc. Int. Workshop on Gerontechnology*, 1-03, pp.31-32, March 2001.

[4] K. Yasu, Y. Miyauchi, N. Hodoshima, N. Hayashi, T. Inoue, T. Arai and M. Shindo, "Evaluation of steady-state suppression of speech for elderly people in reverberant environments," *IEICE Tech. Rep.,* SP2004-154, Mar.2005 (in Japanese).

[5] Y. Nejime, H. Ikeda, T. Imamura, T. Izumi, T. Ifukube and J. Matsushima, "Evaluation of speech-rate conversion method by hearing-impaired listeners," *IEICE Tech. Rep.*, SP92-150, Mar.1993 (in Japanese).

[6] Y. Nejime and B.C.J.Moore, "Evaluation on the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss," *J. Acoust.. Soc. Amer.*, 103(1), Jan. 1998.

[7] T. Adachi, K. Kedera, K. Terashima, N. Maekawa and M. Tateto, "Effect of consonant extension in digital speech processing," *Audiology Japan*, 47, pp181-191, 2004 (in Japanese).

[8] K. Ohshima, T. Nishimoto, and T. Watanabe, "Relationship between oral comprehension and hearing experience of fast TTS for the visually disabled," *IEICE Tech. Rep.*, SP2005-81, WIT2005-43, Oct. 2005 (in Japanese).

[9] K. Kondo, R. Izumi, M. Fujimori, R. Kaga, and K. Nakagawa, "On a two-to-one selection based Japanese speech intelligibility test," *J. Acoust. Soc. Jap*., 63, 196-204, Apr. 2007 (in Japanese).

[10] P. Boersma and D. Weenink,"Praat: Doing phonetics by computer, " http://www.praat.org/

[11] Rice University: Signal Processing Information Base (SPIB), http://spib.rice.edu/spib/select_noise.html