Speech intelligibility derived from time-frequency and source smearing

Toshio Irino, Satoru Satou, Shunsuke Nomura, Hideki Banno, Hideki Kawahara

Faculty of Systems Engineering, Wakayama University, Japan.
{irino, kawahara}@sys.wakayama-u.ac.jp

ABSTRACT

We investigated speech intelligibility of four-mora word sounds degraded with a system based on a high quality vocoder, STRAIGHT, and warped-DCT. This system enables us to independently manipulate essential speech parameters for vocal tract filtering and glottal excitation. We report perceptual effects of: 1) 'temporal smearing' or reduced temporal modulation; 2) 'time-frequency smearing' or reduced resolution in both temporal modulation and spectral peak; and 3) 'source smearing' or reduced resolution of glottal pulses. By analyzing intelligibility scores from the various experiments, we quantitatively confirmed that there are linguistic dependencies of phonemes and morae within words.

1. INTRODUCTION

The intelligibility of degraded speech sounds has been reported in many studies, such as the effects of reduced contrast in spectral envelopes [1-3], reduced fluctuation in temporal envelopes [4,5], and spectral slicing with temporal misalignment [6]. Previously [7], we proposed the use of STRAIGHT [8] for smearing spectral and source excitation information since it is possible to independently manipulate the parameters corresponding to these two characteristics. We investigated the effect of frequency smearing on word perception and the relationship between phonetic and word intelligibilities. For this purpose, we used a word list controlled for familiarity to restrict the use of lexical knowledge to provide additional information about speech processing, since conventional studies used either sentences or syllables for the target speech. We confirmed a common belief that phonemes are not independent within a word even when it is difficult to use lexical information. But it was still difficult to answer whether the source of the dependency is due to physical factors (e.g., coarticulation and prosody) or linguistic factors of Japanese words (e.g. mora structure).

In this paper, we survey whether the dependency is observed even when smearing is performed in other dimensions of spectral and source representations. We report the intelligibility of degraded speech when the STRAIGHT spectra are smeared in temporal envelopes, in both temporal and frequency envelopes, and when glottal excitation is gradually degraded to noise.

2. METHOD

2.1. STAIGHT and warped-DCT

We used STRAIGHT [8] and warped-DCT for generating degraded speech as in the previous study [7]. STRAIGHT is a high quality vocoder that decomposes speech sounds into smoothed time-frequency representation about the vocal tract filter and source information about voicing and glottal vibration (or F0). It is confirmed that there are almost no artifacts in the analysis/synthesis processing, and we can produce the desired signal in accordance with the manipulation. The STRAIGHT spectrum is derived for a 1-ms frame rate and represented in linear magnitude.

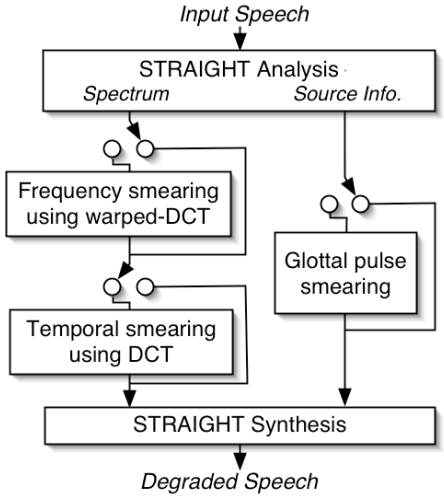

Figure 1. Speech degradation by STRAIGHT.

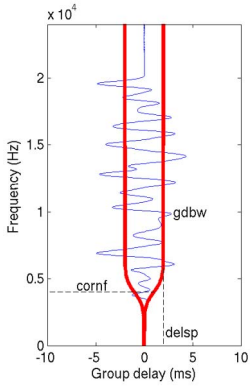

Figure 2. Glottal pulse for STRAIGHT synthesis.

Warped-DCT [7] is an orthogonal transform that accommodates both Discrete Cosine Transform (DCT) and frequency warping from linear to mel-like scales. There is a parameter, α, that controls the degree of warping; when $\alpha = 0.68$ and the sampling frequency is 48 kHz, the warped frequency is close to the ERB scale; it becomes a simple DCT when $\alpha = 0$. The original signal is recovered from the full order of the coefficients simply by applying inverse warped-DCT. The degree of smearing is determined by the upper limit of the warped-DCT order.

2.2. Signal degradation

Figure 1 shows a block diagram for producing degraded speech sounds using STRAIGHT. For frequency smearing, warped-DCT is applied vertically to each frame of the STRAIGHT spectrum. The spectrum becomes smoother as the upper limit of the warped-DCT order becomes smaller. For temporal smearing, DCT ($\alpha = 0$) is applied horizontally to each frequency bin of the STRAIGHT spectrum. We made

the upper limit of the DCT order directly correspond to the upper limit of temporal modulation in Hz. The effects of smearing in the time-frequency representation were investigated in experiments 1 and 2.

In STRAIGHT, the glottal pulse is excited at the time derived from the F0 information during the voiced parts; noise is generated during the rest. The glottal pulse is represented with an all pass signal, as shown in Fig. 2. There are three parameters: 1) 'cornf' specifies a corner frequency from impulse to noise, 2) 'delsp' specifies deviation of the group delay for the noise, and 3) 'gdbw' specifies frequency resolution of the group delay. It is possible to generate a signal from a pure impulse to widely spread noise by mainly manipulating 'cornf' and 'delsp'. The default values of cornf, delsp, and gdbw are 4000 Hz, 0.5 ms, and 70 for high quality speech synthesis. When cornf = 0 and delsp = 20, the synthesized speech sounds like whispered speech. The effects of smearing glottal pulse were investigated in experiment 3.

3. EXPERIMENTS

3.1. Conditions

3.1.1. Stimulus sounds

In this experiment, we used the speech sounds of four-mora Japanese words selected from a database controlled with respect to both word familiarity and phonetic balance [9]. The word lists were categorized in four different familiarities from low to high; each category has 20 word lists, and each word list consists of 50 words. We used the word list of the lowest familiarity in which the words are rarely used in everyday life, i.e., almost nonsense words, even though they are listed in a dictionary. This is because we intended to restrict the use of a mental lexicon for guessing what they had heard. Moreover, we minimized replication of the same word to prevent short time learning effects within the experiment.

We synthesized sets of speech sounds in accordance with the degradation conditions. For each condition, we prepared 100 degraded sounds produced from two different word lists and recorded by one male and one female speaker. In addition, we also produced a set for simple analysis/synthesis sounds without applying any degradation processing (designated as 'Inf') and also used a set for the original sounds (designated as 'Orig') to confirm the sound quality of STRAIGHT.

The signal level was digitally equalized in terms of the equal loudness level, or L_{Aeq}, which is the RMS level of the entire speech sound after A-weight filtering. The sounds were played back using M-audio FireWire 410 with 48 kHz and 24 bit and presented to the subjects in a soundproof room through headphones (Sennheiser HD-580) at a level of 70 dB A. The sounds of words were randomly presented and followed by 3 sec silent intervals for writing answers.

3.1.2. Subjects and task

Twelve Japanese subjects around the age of 22 with normal hearing thresholds between 250 and 8000 Hz participated in the experiments. These subjects were divided into two groups; six participated in experiments 1 and 2; the other six participated in experiment 3.

The subjects were instructed to write down the perceived words in Japanese 'kana' characters, which uniquely correspond to 'morae' consisting of either vowels or CV syllables. Answers were converted into phonetic sequences using a unique rule. There are 28 consonants: 'm,' 'p,' 'b,' 't,' 'd,' 's,' 'ts,' 'dz,' 'r,' 'n,' '$\int$,' '$t\int$,' 'd_3,' 'k,' 'g,' 'h,' 'hj,' 'Φ,' 'mj,' 'pj,' 'bj,' 'kj,' 'gj,' 'rj,' 'nj,' 'w,' 'j,' and 'N.' Furthermore, there is no consonant condition (represented by '.'). The vowels are: 'a,' 'i,' 'u,' 'e,' and 'o,' and there is a 'no vowel' condition (in the case of stop nasal 'N').

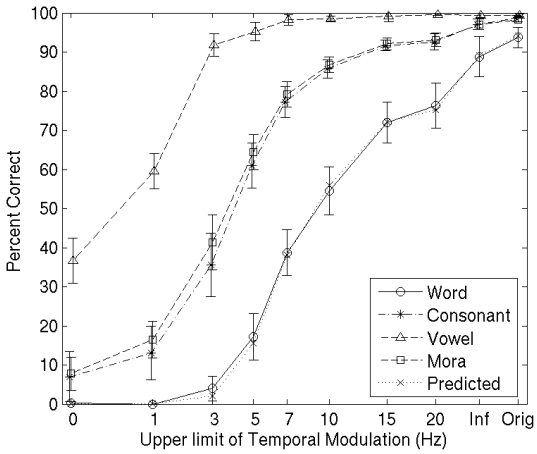

Figure 3 Intelligibility of word, consonant, vowel, and mora in temporal smearing (Exp. 1). Dotted line with cross shows the prediction result (M=3.6, N=1.7) described in section 3.5.

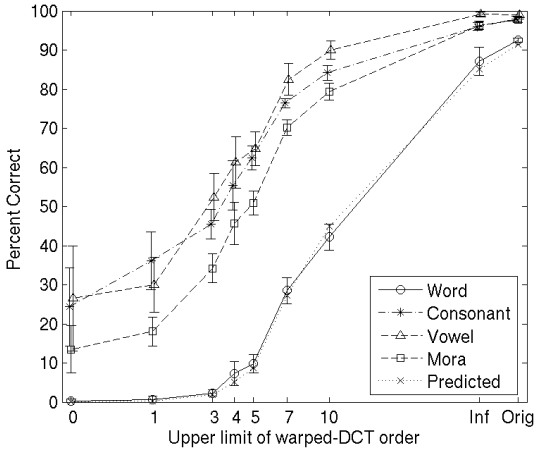

Figure 4. Results from Jiang et al. 2003 [9]. Intelligibility of word, consonant, vowel, and mora in frequency smearing.

3.2. Experiment 1: Temporal smearing

In this experiment, the process for temporal smearing using DCT (Fig. 1) was applied to the STRAIGHT spectrum. We synthesized eight sets of degraded speech sounds with respective upper limits of temporal modulation of 0, 1, 3, 5, 7, 10, 15, and 20 Hz.

3.2.1. Intelligibility

Figure 3 shows the average score and standard deviation of subjects as a function of the upper limit of temporal modulation. The word identification score for analysis/synthesis sounds ('Inf') is about 89%, which is 5% less than the score for the original sounds ('Orig'). It shows that the quality of STRAIGHT sounds is sufficiently high while some of the phonemes are degraded. The speech reception threshold (SRT) for words (the condition for a score of 50%) is about 10 Hz. The vowel identification score is more than 90 % when the upper limit is more than 3 Hz and

drops suddenly when it is less than 3 Hz. The SRT for vowels is less than 1 Hz. By contrast, the consonant identification score gradually changes across the upper limit, and the SRT is about 4 Hz. The difference indicates that consonants are more transient than vowels. The results are consistent with previous studies [4, 5].

The identification scores for morae are slightly greater than those for consonants. This is because one mora basically corresponds to either a single vowel or a CV syllable, and the high intelligibility of vowels raises the score.

3.2.2. Contrast with frequency smearing

We compared the current results with a previous study on frequency smearing by Jiang et al. [7]. Figure 4 shows the average score and standard deviation of six subjects as a function of the upper limit of warped-DCT order. It is clear that the identification scores of vowels and consonants are always about the same. Frequency smearing affects the intelligibility of vowels and consonants in a similar way.

The SRT for words is a little more than 10. This value is coincidentally similar to the SRT in temporal smearing, although the dimensions of frequency and temporal smearing are completely different and basically independent.

3.3. Experiment 2: Time-frequency smearing

In experiment 1, Jiang 2003, and many previous studies [1-5], smearing was performed in one dimension along either the time or the frequency axis. We surveyed the speech intelligibility when spectra are smeared simultaneously in both time and frequency dimensions.

As shown in Fig. 1, temporal smearing using DCT followed frequency smearing using warped-DCT. The sequence does not affect the sounds because it is linear processing. We used default source parameters and did not apply glottal pulse smearing. The upper limits of warped-DCT order were set to 0, 7, and 15; the upper limits of temporal modulation were set to 0, 7, and 15 Hz. There were eleven conditions in the experiment: nine degraded, analysis/synthesis ('Inf'), and original ('Orig').

Figures 5(a), (b), and (c) respectively show the identification scores of words, consonants, and vowels as a function of the upper limit of temporal modulation. Simple straight lines are used to connect the scores in the same warped-DCT conditions. Dashed lines with squares are the scores in experiment 1 (Fig. 3), i.e., the upper limit of temporal modulation of infinity or no smearing in frequency dimension. The scores of consonants are about 10% or a little more when the upper limit of warped-DCT order is 0; they are more than 50% when the upper limits of warped-DCT and temporal modulation are either 7 or 15. The scores of vowels are slightly above the chance level (20%) even when the upper limit of warped-DCT order is 0; they are more than 80% when upper limits are high.

Figures 5(d), (e), and (f) show the same results viewed as a function of the upper limit of warped-DCT order. Dashed lines with squares are the scores from the experiments by Jiang et al. 2003 [7], which corresponds to the upper limit of the warped-DCT order of infinity or no smearing in the temporal dimension, although the subjects were different and the sound pressure level was 78 dB A. The scores of vowels are higher than 80% and almost unchanged when the upper limits of warped-DCT and temporal modulation are either 7 or 15.

3.4. Experiment 3: Source smearing

Vocal source signals may also affect speech intelligibility. Previous studies only used either original glottal information [1, 2, 4, 5, 6] or simple white noise [3]. By using the STRAIGHT system, it is possible to access the effect of intermediate sources between them.

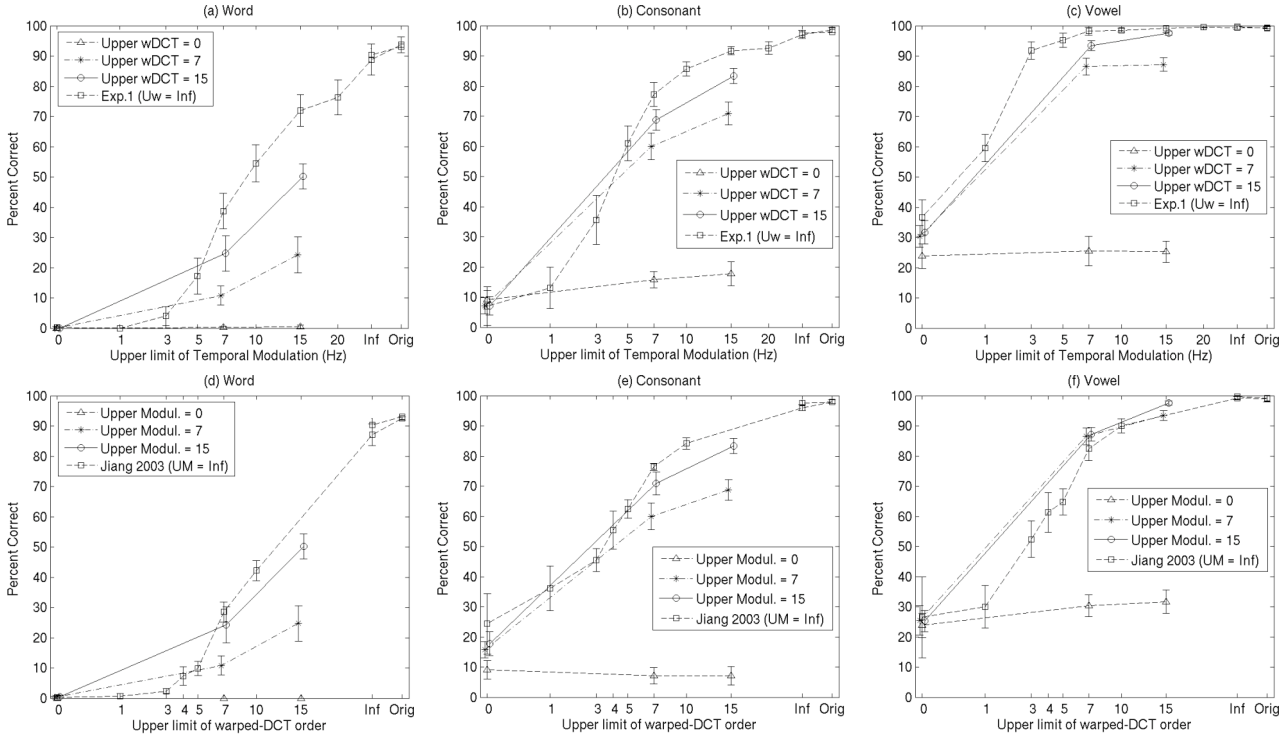

Figure 5. Intelligibility of words, consonants, and vowels by time-frequency smearing in experiment 2 with results in experiment 1 (a),(b),and (c) and in Jiang et al. 2003 (d),(e), and (f).

We produced degraded sounds in twelve conditions: three for source smearing ([delsp, cornf] = [20, 0; 20, 1500; 0.5, 4000]) and four for frequency smearing (Upper limits of the warped-DCT order: [0, 10, 15, Inf]). Figure 6 shows the identification scores of words. The scores for the intermediate condition ([20,1500]) are about the same as the clean source condition ([0.5, 4000]) and the results of Jiang 2003 [9]. The scores for the whisper condition ([20,0]) are about 15% less than these conditions. Consequently, source smearing does not have a great effect on word intelligibility when sounds were presented without background noise as in this series of experiments.

3.5. Relationship between phonemes, morae, and words

We used a database of four-mora words with controlled familiarity to investigate the relationship between the identification scores of phonemes, morae, and words.

When P_c, P_v, P_m, and P_w represent the average identification scores for consonants, vowels, morae, and *words*, we assumed that P_w is predicted with simple equations as

$$P_w = P_c^M P_v^N, \quad (1)$$

$$P_w = P_m^K, \quad (2)$$

where M, N, and K are constant power factors.

If the phonemes are completely independent in Eq. 1, then the expected values of M and N are roughly the same as the frequency of appearances in words, i.e., about 3.5 and 3.6 in this database. If the morae are completely independent in Eq. 2, then the expected value of K is 4 because of four-mora words.

We estimated these values from the identification scores of phonemes, morae, and words by using a least mean square (LMS) method. The results are shown in Table I when using the data of the individual experiments and when using all of the data together.

Note that there is very little variability across the experiments and that the estimation errors are about 2%, which is less than the variability of scores between subjects. M is about the same as the frequency of appearance (3.5) whereas N is about half of 3.6. So the phonemes are clearly not independent. Word identification is almost dominated by the identification of consonants and is hardly affected by the identification of vowels. The value of K is slightly less than 4, except in experiment 1 when the identification scores of vowels were very high and the scores of morae were better than the scores of consonants. The morae are not independent, either, since K is less than 4. We also found that the scores are significantly higher for 2nd and 4th morae than for 1st and 3rd morae in all of the experiments.

The source of dependency is due to neither the access to word lexicons nor physical attributes (spectral and source information) of degraded sounds since word familiarity is very low and the ways of degradation are completely different in accordance with the experiments. One possibility is the use of implicit knowledge for mora and 'bimora' structures that correlate to the likeliness of Japanese words [10].

4. CONCLUSION

We reported speech intelligibility of degraded four-mora words produced with a system based on STRAIGHT and warped-DCT. The results are the following: 1) 'Temporal smearing' or reduced temporal modulation mainly affects the identification scores of consonants due to their transient nature. 2) 'Time-frequency smearing' or reduced resolution in both temporal modulation and spectral peak affects the identification scores intermediately between simple frequency and temporal smearing conditions. 3) 'Source smearing' or reduced resolution of glottal pulses barely affects the identification scores. 4) We quantitatively confirmed that there are linguistic dependencies of phonemes and morae within words.

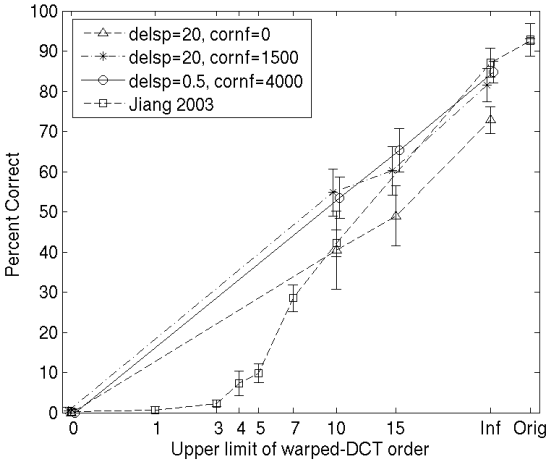

Figure 6. Intelligibility of words by source and frequency smearing.

Table I. Estimated power factors and rms errors (%)

	M	N	Error	K	Error
Jiang 2003	3.5	1.7	2.5	3.6	1.6
Exp. 1	3.5	2.5	1.7	4.1	1.5
Exp. 2	3.6	1.6	2.0	3.8	1.9
Exp. 3	3.7	1.6	2.1	3.8	1.9
All	3.6	1.7	2.1	3.8	2.0

Acknowledgments: This work was partially supported by a grant from the Faculty of Systems Engineering at Wakayama University.

REFERENCES

[1] ter Keurs, M. Festen, J. M., and Plomp, R. "Effect of spectral envelope smearing on speech reception. I," J. Acoust. Soc. Am., 93(5), pp.2872-2880, 1992. (see also references in [7])

[2] Moore, B. C. J., Glasberg, B. R., and Simpson, A. "Evaluation of a method of simulating reduced frequency selectivity" J. Acoust. Soc. Am., 91(6), 3402-3423, 1992.

[3] Shannon, R.V, Zeng, F-G, Kamath, V. Wygonski, J., and Ekelid, M. "Speech recognition with primarily temporal cues," Science, vol.270, pp. 303-340, 1995.

[4] Rosen, S.," Temporal information in speech: acoustics, auditory and linguistic aspects," Phil. Trans. R. Soc. Lond. B, 336, 357-373, 1992.

[5] Drullman, R., Festen, J. M., and Plomp, R. "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am., 95,2,1053-1064, 1994.

[6] Greenberg S., Arai, T., and Siliop, R. "Speech intelligibility derived from exceedingly sparse spectral information," ICLSP, 1998.

[7] Jiang, J., Banno, H., Kawahara, H., and Irino, T. "Intelligibility of degraded speech from smeared STRAIGHT spectrum," ICSLP 2004 (INTERSPEECH 2004) , vol. I, pp. 473-476, 2004.

[8] Kawahara, H., Masuda-Katsuse, I., and de Cheveign'e, A. "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, 27 (3-4), pp.187-207, 1999.

[9] Sakamoto, S, Suzuki, Y., Amano, S., Ozawa, K., Kondo, K., and Sone, T., "New lists for word intelligibility test based on word familiarity and phonetic balance," J. Acoust. Soc. Jpn., 54(12), 842-849, 1998 (in Japanese).

[10] Kondo, T. and Amano, S., "Correlation of bimora frequencies between spoken familiarity ranks," Proc. autumn meeting of Acoust. Soc. Jpn., vol. 1, pp.419-420, 2001 (in Japanese).