China-Japan Joint Conf. on Acoustics, 199-202, 2002

# SUPPRESSING STEADY-STATE PORTIONS OF SPEECH FOR IMPROVING INTELLIGIBILITY IN VARIOUS REVERBERANT ENVIRO MENTS

Nao Hodoshima, Tsuyoshi Inoue, Takayuki Arai, Akiko Kusumoto and Keisuke Kinoshita
Dept. of Electrical and Electronics Engineering, Sophia University, Tokyo, Japan

## ABSTRACT

In previous studies (Arai et al., 2001; Arai et al., 2002), we hypothesized that segments of an acoustic signal are masked by reverberation components of previous segments, degrading speech intelligibility. To reduce masking influences, we suppressed steady-state portions having more energy, but which are less crucial for speech perception. We have presently conducted a perceptual test with a set of artificial reverberations to explore the relationship between steady-state suppression and several reverberation conditions. The results indicated clear improvements for some reverberation conditions. We certified that Arai's technique was an effective pre-processing method for improving speech intelligibility under reverberant conditions.

## 1.INTRODUCTION

In a large auditorium, perceiving speech is often difficult. This is due to reverberation that is caused by a superposition of reflected sounds with various delays and amplitudes. Because reverberation tails affect subsequent segments, an acoustic signal of one segment is masked by the reverberation components of the previous portion, and this masking effect degrades speech intelligibility [1][2].

Syllable identification tests show that the spectral transition is crucial for syllable perception [3]. This is likely due to the fact that the information in steady-state portions of the speech signal is relatively redundant with that in transient segments [4]. Both the "delta" processing of cepstral features [3] and the RelAtive SpecTrAl (RASTA) processing [4] enhance transitions of speech and contribute to increase recognition rate in automatic speech recognition.

There are two general approaches for improving speech intelligibility in reverberant environments: pre-processing and post-processing. In the pre-processing approach, a speech signal is processed between a microphone and loudspeaker (e.g., [1][2][5][6][8]). Post-processing is a dereverberation technique applied to a signal having already been released into a room and affected by reverberation (e.g., [6][7]). As a pre-processing approach, Arai et al. suppressed the steady-state portions of speech which have more energy but which are less crucial for speech perception in order to reduce the masking influence caused by reverberation [1][2]. Another approach is modulation filtering, which alters the modulation spectrum of a signal ([5]-[7]). In light of the discovery that the important modulation frequency of a signal for speech perception is around 4Hz, Kusumoto et al. enhanced this particular frequency region in their application of the modulation filter in their pre-processing approach [5]. Both [1] [2] and [5] showed promising results for improving speech intelligibility.

As pre- and post- processing, Langhans et al. proposed the theoretical inverse modulation transfer

function (IMTF) filter, which artificially increased the modulation depth of a reverberated signal in order to account for the decrease in the modulation index of the signal from reverberation [6]. Avendano et al. also artificially increased the modulation depth as a post-processing method; they employed an IMTF filter derived from their own training data [7].

Our ultimate goal is to provide a filter for pre-processing which is suitable for an individual auditorium having a distinct reverberation time. In order to achieve this, we need to better understand the relationship between the effect of processing filter and reverberation time. To explore this relationship, a perceptual experiment has been carried out with the modulation filter and various reverberation times [8]. The results showed that the effect of the modulation filtering depended on reverberation time, and the proposed method prevented the degradation of speech intelligibility under specific conditions. While the study in [8] examined the effect of the modulation filtering, in this paper we aimed to explore the relationship between the steady-state suppression and several different reverberation conditions. By using the same steady-state suppression technique as in [1][2], we conduct a perceptual test with a set of artificial reverberations.

## 2.PERCEPTUAL EXPERIMENT

The artificial impulse responses $h_n$ were created as Eq. (1) to obtain the desired reverberation conditions [8] :

$$h_n(t) = e^{-t/\tau} h_o(t) \qquad (1)$$

where $\tau$ is a time constant. The original impulse response $h_o$ used for this study was measured in the Hamming Hall in Higashi Yamato City, Tokyo. (A reflection board was not used.) Thus, we can obtain the desired reverberation time as a function of $\tau$. Table 1 shows the set of reverberation conditions used in our experiment ($h_o$ is identical to h3 in Table 1). We used the reverberation time $T_{60}$, defined as the time the decay curve of the impulse response decreased 60 dB from steady state.

We applied the same method as in [1] [2] to suppress the steady-state portions of speech. First, an original signal was split into 1/3-octave bands. In each band the envelope was extracted. After down-sampling, the regression coefficients were calculated from the five adjacent values of the time trajectory of the logarithmic envelope of a subband. Then the mean square of the regression coefficients, *D*, was calculated. We used the *D* parameter by Furui to measure the spectral transition [3]. After up-sampling, we defined a speech portion as steady-state when *D* was less than a certain threshold. Once a portion was considered steady-state, the amplitude of the portion was multiplied by a factor 0.4.

The stimuli consisted of nonsense Consonant-Vowel (CV) syllables embedded in the Japanese carrier phrase, 'Daimoku to shite wa __ to iimasu' (It is said __ as a title). Twenty-four CVs used

Table 1. Reverberation conditions used in the experiment

| Impulse response | h1 | h2 | h3 | h4 | h5 |
|---|---|---|---|---|---|
| Rev. time (s) | 0.9 | 1.0 | 1.1 | 1.2 | 1.3 |

Table 2. CVs used in the experiment

| | Stop+V | Fricative+V | Affricate+V | Nasal+V |
|---|---|---|---|---|
| Voiceless Consonant+V | /pa/ /ta/ /ka/ /pi/ /ki/ | /sa/ /ʃa/ /ha/ /ʃi/ /hi/ | /ʧa/ /ʧi/ | |
| Voiced Consonant+V | /ba/ /da/ /ga/ /bi/ /gi/ | | /dza/ /ʤa/ /ʤi/ | /ma/ /na/ /mi/ /ni/ |

in the experiment are shown in Table 2. The original speech samples were obtained from the ATR Speech Database of Japanese. (The speaker was MAU, a 40-year-old male). The CV syllables were chosen from the monosyllable data set. The carrier was selected from two sentences in the sentence data set and we used both parts in each sentence. The beginning position of the target vowel was adjusted to 150ms from the end of the pre-target carrier phrase. The stimuli consisted of two conditions: the original signals with reverberation (Org_rev) and the processed signals with reverberation (Proc_rev).

Twenty-four normal hearing subjects (14 males and 10 females, ages 18 to 26) participated in the experiment. All were native speakers of Japanese.

The experiment, controlled by a computer, was conducted in a soundproof room. The stimuli were presented with headphones (STAX SR-303), and the sound level was adjusted to each subject's comfort level. In the experiment, a stimulus was presented at each trial. Then 24 CVs in Kana orthography were shown on the screen. Subjects were forced to choose one of 24 CVs by clicking a button on the screen with a mouse. For each subject, 240 stimuli were presented randomly (5 reverberation conditions x 24 CVs x 2 processing conditions).

## 3.EXPERINMENTAL RESULTS

The mean percent correct for each reverberation and processing condition is shown in Table 3. We analyzed the results for 22 subjects (we excluded two outliers). A 2x5 ANOVA for repeated measures was performed, confirming significant main effects of processing types ($p<.001$) and impulse response types ($p<.001$). For the comparison of means between processing types, a t-test was performed for each impulse response type. A significant difference was obtained for the h1-h4 conditions (h1: $p=.049$; h2: $p=.026$; h3: $p=.003$; and h4: $p<.001$).

Table 3. Mean percent correct in each condition

| | h1 | h2 | h3 | h4 | h5 |
|---|---|---|---|---|---|
| Org_rev (%) | 66.5 | 63.5 | 61.4 | 55.1 | 58.1 |
| Proc_rev (%) | 73.1 | 68.3 | 67.4 | 64.2 | 58.5 |

## 4.DISCUSSION

We confirmed that the rates for correct responses declined as reverberation time increased, regardless of processing type. It was found that Proc_rev performed better than Org_rev under all reverberant conditions, and a t-test showed that the differences between the correct responses are significant for conditions h1-h4 (Rev. time: 0.9-1.2 s). Because the most clear improvement was

obtained for the h4 condition (Rev. time: 1.2 s), we thought that the amount of masking of the target caused by the reverberation components of the previous portion was the smallest under the h4 condition. Our results confirmed that the steady-state suppression was useful for improving speech intelligibility as a pre-processing measure and that the effect of the steady-state suppression differed with respect to reverberation time.

## CONCLUSION

In this paper we suppressed the steady-state portions of speech based on Arai's technique [1][2] for improving speech intelligibility in reverberant environments. To explore the relationship between the steady-state suppression and several reverberation conditions, we conducted a perceptual test with a set of artificial reverberations. The results showed that the effect of the steady-state suppression depended on reverberation time and clear improvements were obtained with reverberation times of 0.9-1.2 s, maximally at 1.2 s. Thus, we certified that Arai's technique [1][2] was an effective pre-processing method for improving speech intelligibility under reverberant conditions.

## ACKNOWLEDGEMENT

We appreciate Hideki Tachibana, Kanako Ueno and Sakae Yokoyama for offering to use the impulse response data. Also, we thank the subjects who participated in our experiment.

**REFERENCES**

[1] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto and T. Kitamura, "Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments," Proc. Autumn Meet. Acoust. Soc. Jpn., 449-450 (2001)

[2] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto and T. Kitamura, "Effects on suppressing steady-state portions of speech on intelligibility in reverberant environments," Acoustical Science and Technology, **23**, (2002)

[3] S. Furui, "On the role of spectral transition for speech perception," J. Acoust. Soc. Am., **80 (4)**, 1016-1025 (1986)

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE Trans. Speech Audio Process., **2**, 578-589 (1999)

[5] A. Kusumoto, T. Arai, M. Takahashi and Y. Murahara, "Modulation enhancement of speech as a preprocessing for reverberant chambers with the hearing-impaired," Proc. IEEE ICASSP, 933-936 (2000)

[6] T. Langhans and H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering," Proc. IEEE ICASSP, 156-159 (1982)

[7] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," Proc. ICSLP, 889-892 (1996)

[8] N. Hodoshima, T. Arai and A. Kusumoto, "Enhancing temporal dynamics of speech to improve intelligibility in reverberant environments," Proc. Forum Acusticum, Sevilla (2002)