Enhanced Speech Yielding Higher Intelligibility for All Listeners and Environments

Takayuki Arai[1] *and Nao Hodoshima*[2]

[1]Department of Information and Communication Sciences
Sophia University, Tokyo, Japan
arai@sophia.ac.jp
[2]Department of Information Media Technology
Tokai University, Tokyo, Japan
hodoshima@tokai-u.jp

Abstract

The current paper discusses two approaches to enhanced speech in reverberation/noise: machine signal processing and human speech production. We reviewed the speech enhancement techniques, including steady-state suppression and compared the modulation spectra of speech signals before and after processing. We also introduced the Lombard-like effect of speech in reverberation, and compared the characteristics of speech signals, including the modulation spectra between speech signals uttered in quiet and reverberation. We found that the enhanced speech signals have distinct characteristics that yield higher speech intelligibility.

Index Terms: speech enhancement, steady-state suppression, modulation spectrum, intelligibility of speech, reverberation, Lombard effect

1. Introduction

In the past, researchers have explored speech enhancement techniques to improve the intelligibility of speech in different listening conditions. Many studies have focused on signal processing after speech signals were degraded by environmental noise and/or reverberation. In such studies, the main goal of the signal processing is to recreate a signal that is as close as possible to the original signal, that is, "speech restoration". Other techniques have focused on signal processing to "enhance" speech signals, so that the processed signal is more intelligible than the original one, that is, "speech feature exaggeration."

There are several types of speech feature exaggeration in both the spectral and temporal domains. One technique in the spectral domain is formant enhancement (e.g., [1]). This technique emphasizes spectral peaks, such as formants, that are important for speech perception, and it is reported that formant enhancement yields better intelligibility especially for elderly and/or hearing-impaired listeners. "Frequency compression" is another spectral domain enhancement technique (e.g., [2]). Frequency compression can be useful for listeners who do not have flat audiograms. Another frequency compression technique is called "critical-band compression" [3]. Critical-band compression improves the intelligibility of speech for hearing-impaired listeners by compressing frequency components within each critical band towards the center frequency [3, 4].

There are also several speech enhancement techniques that operate in the temporal domain. One temporal domain technique is "time-scale modification" (TSM; e.g., [4]), where an input speech signal is elongated without lowering the fundamental frequency. This technique changes the length of the utterance, so special treatment is needed for real-time processing. Another temporal domain technique is "modulation filtering" (e.g., [5]). In our previous studies [6], we conducted perceptual experiments where the intelligibility of processed speech by modulation filtering was tested in simulated reverberant environments. In these studies, the speech signal processed with modulation filtering was more intelligible for normal and hearing-impaired listeners.

Arai *et al.* [7, 8] further proposed steady-state suppression (SSS) to improve speech intelligibility. This technique deemphasizes steady-state portions of speech that are less important for speech perception. It is believed that SSS reduces temporal masking, and it has been reported that speech processed with SSS has higher intelligibility than the original signal for elderly listeners [9, 10]. It is also pointed out that SSS reduces so-called "overlap-masking" in reverberation, which is known to be one of the main factors degrading the intelligibility of speech [11]. Hodoshima *et al.* [12] showed that SSS significantly improved the intelligibility of speech in reverberant environments.

Combining TSM and SSS turns out to be synergistic. Arai *et al.* [13] pointed out that a simple time-scale elongation is not the best way to reduce overlap-masking, because vowel portions are usually elongated in slow speech but have more speech energy. The combination of TSM and SSS yielded better performance in reverberant speech.

We have found it useful to take into account the nature of speech production when researching how to improve speech intelligibility in severe environments. This is because when people speak, they adapt their speech to compensate for noise and reverberation. Many studies have looked at noisy speech and shown that the acoustic characteristics in the temporal and spectral domains (e.g., intensity, duration, F0, F1 and F2) are different than those of speech spoken in quiet [14, 15]. This is known as the Lombard effect (e.g., [16]). It has also been discussed that speech uttered in noise is more often intelligible than that uttered in quiet. However, the same scrutiny has not been applied to reverberant speech. Precisely how people modify the acoustic characteristics of their speech in reverberant environments has not been clarified. [14-16].

One of our goals is to provide intelligible speech announcements in noisy and/or reverberant public spaces, such as train stations. In this paper, we discuss "enhanced speech" with two approaches (machine signal processing and human speech production) that focus on yielding higher speech intelligibility. In the current paper, we review the SSS technique, comparing the modulation spectra of the original

Copyright © 2010 ISCA

and processed speech signals. Then, we compare the enhanced speech signals in noise/reverberation with ones in quiet. Finally, we use listening tests for young and elderly participants to show that the enhanced speech sounds are intelligible in reverberation as well as in noise. We show that steady state suppression yields intelligible speech for all listeners (from young through elderly) and all environments (in noise and reverberation).

2. Steady-state suppression

2.1. Background

Speech sounds reflect the static as well as dynamic aspects of speech production. Speech may be seen as the sequence of syllables where the onset and coda of each syllable are dynamically changing in time, because of co-articulation, while the nucleus, although it does contain some information, is more steady and has more energy. It has been shown that the information in steady-state portions of a speech signal, such as syllable nuclei, is relatively insignificant compared with the information in transient portions, like the syllable onset and coda [17]. RelAtive SpecTrAl (RASTA) processing, which relies on this observation, enhances transitions of speech to improve performance of automatic speech recognition (ASR) [18]. Like RASTA, modulation filtering also enhances the modulation frequency components between 1 and 16 Hz, that are important for both human speech perception [19] and ASR [20].

As in RASTA processing, SSS enhances the transitions, while the steady-state parts are suppressed or attenuated. While RASTA and modulation filtering apply a linear filter on the temporal envelope domain, SSS detects steady-state portions by using Furui's measure [21] and suppresses them by directly manipulating the temporal envelopes.

2.2. Comparison of the modulation spectra

Figures 1 and 2 show the modulation spectra of the original speech and speech processed by SSS, respectively. The modulation spectrum is used because it is reported that this measure reflects speech dynamics and is well correlated with the intelligibility of speech [22]. In this figure, 32 sentences were used to compute the modulation spectra. For each sentence, band pass filters were used to divide a speech signal into four frequency bands (Band 1: 0-800 Hz; Band 2: 800-1600 Hz; Band 3: 1600-3200 Hz; and Band 4: 3200-8000 Hz). For each band-passed signal, the temporal envelope was extracted and the modulation indices were computed based on Houtgast and Steeneken's definition [23]. Finally, all the modulation spectra among 32 sentences were averaged within each frequency band.

As shown in Fig. 1, the main peak in the modulation frequency is located around 4 Hz, which reflects the syllabic rate of speech. For the processed speech, the strong increase in the modulation index around 10 Hz, which reflects phonemic rate of speech, is observed, as discussed in [12].

3. Lombard-like effect in reverberation

3.1. Acoustic analysis of speech produced in noise/reverberation

In this section, we show how a Lombard-like effect was observed in speech spoken in reverberant environments as reported in noise [14-16]. This section references Hodoshima *et al.* [24] for its acoustic analysis of speech. We recorded ten words (4 morae each) in a carrier sentence spoken by four young native speakers of Japanese in a sound-proof room. The recording conditions were quiet (Q), with white noise (N) and reverberation (R1 and R2) provided to the speakers through their headphones. Reverberation time (T60) of the impulse responses were 3.6 s (R1) and 12.3 s (R2). The impulse responses were recorded at a church and a tunnel, respectively.

Figure 3 shows the mean acoustic characteristics of speech produced in Q, N, R1 and R2. The results showed that (1) modification in speech production was observed in reverberation (i.e., an increase in word duration, overall intensity, F0, F1, and a decrease in consonant-vowel intensity ratio), (2) a similar modification was observed both in reverberant and noisy environments, although the changes are not exactly the same, and (3) reverberation time slightly affected the degree of the modification. The results suggest that speakers modify their speech production in specific ways depending on the acoustic environment (i.e., noise and

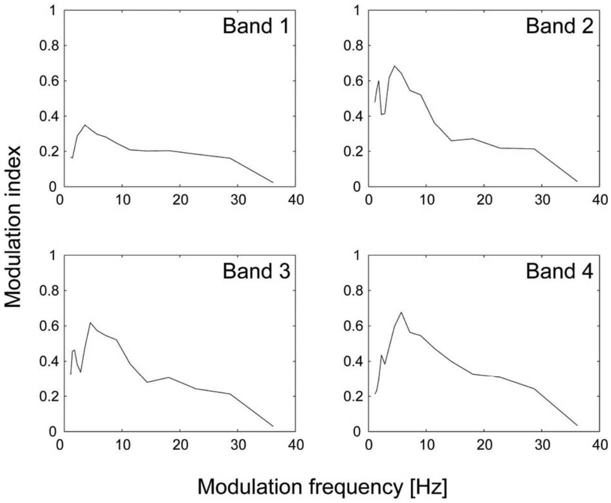

Figure 1: *Modulation spectra of the original speech signals.*

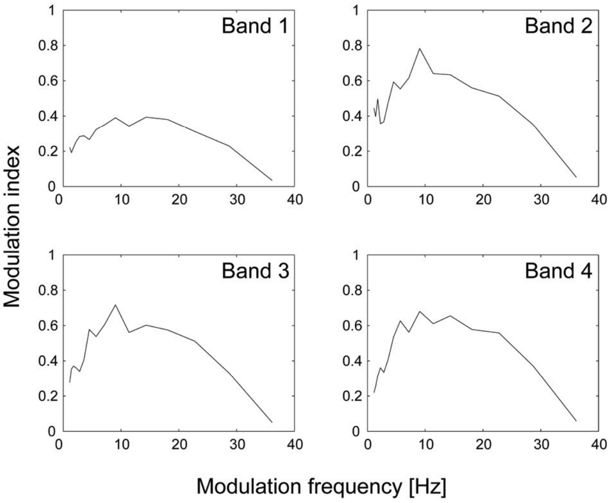

Figure 2: *Modulation spectra of the speech signals after steady-state suppression.*

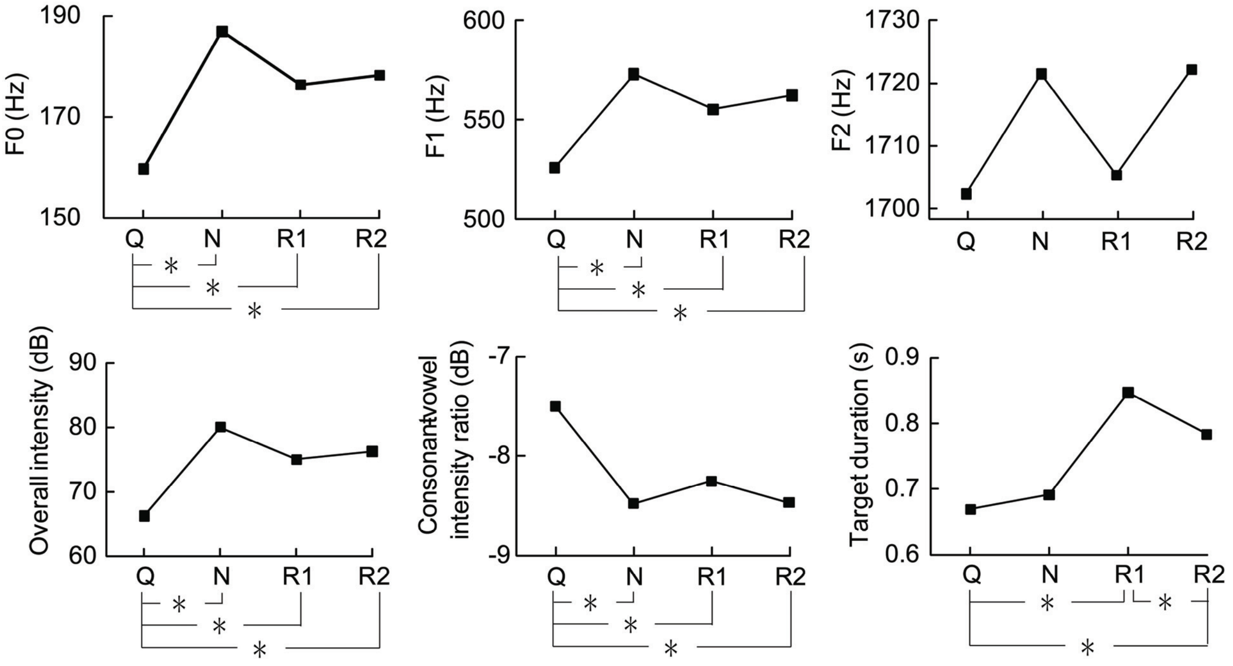

Figure 3: *Mean acoustic characteristics of speech spoken in quiet (Q), noise (N) and reverberation (R1 and R2). An asterisk shows a significant difference at p<0.01. (Data are replotted from [24].)*

reverberation). A larger inter-speaker difference in reverberant environment (not shown in the figure) indicates that different speakers use different strategies of speech enhancement to cope with reverberant sounds than they use to cope with relatively stationary white noise.

3.2. Intelligibility of speech spoken in noise/reverberation

The intelligibility of the speech spoken in reverberation/noise was compared with quiet speech by means of listening tests carried out in reverberant/noisy listening environments. The recording conditions were essentially the same as in Section 3.1 except that 32 new words in a different carrier sentence, 2 new speakers, and 1 reverberation condition (R1) were used. Participants were 32 young (4 males and 28 females, 23 years old on average) and 32 elderly (11 males and 21 females, 72 years old on average). All were native speakers of Japanese. For the listening condition, either white noise was added or an impulse response (Ra, Rb) was convolved with the speech spoken in each of Q, N or R. This yields 6 listening conditions: Q_N, N_N, Q_Ra, Q_Rb, R1_Ra, and R1_Rb. For example, Q_N refers to when the recording was in quiet (Q) and the listening test was conducted in noise (N); the other abbreviations are interpreted in the same way. Signal-to-noise ratio was -2 dB for the young participants and 2 dB for the elderly participants. Reverberation times were 2.6 s (Ra) and 3.6 s (Rb) for the young participants and 1.5 s (Ra) and 2.5 s (Rb) for the elderly participants.

Figure 4 shows the mean percent correct of mora of the young participants. The results showed that N_N, R1_Ra and R1_Rb had significantly higher mora identification scores than Q_N, Q_Ra and Q_Rb, respectively. This indicates that speech spoken not only in noise but in reverberation was more intelligible than speech spoken in a quiet environment.

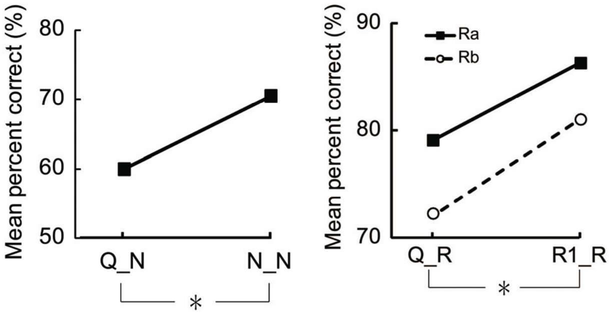

Figure 4: *Mean percent correct of mora of young participants (left: listening test in noise, right: listening test in reverberation). An asterisk shows a significant difference at p<0.01. (Data are replotted from the part of [25].)*

The results of the elderly participants showed similar tendencies. Since elderly people generally have more difficulty understanding speech with noise/reverberation as compared to quiet, the results imply that speech enhanced in this way will help the elderly understand speech in noisy/reverberant public spaces.

3.3. Comparison of the modulation spectra

Figure 5 shows the modulation spectra of the speech signals in quiet and in reverberation. In this figure, again, the same 32 sentences were used to compute the modulation spectra as in Section 2.2. As shown in this figure, the main peak in the modulation frequency is located around 4 Hz as observed in Fig. 1. The slight increase in the modulation index around 4 Hz of the speech in reverberation (especially in the 4th band) is also observed.

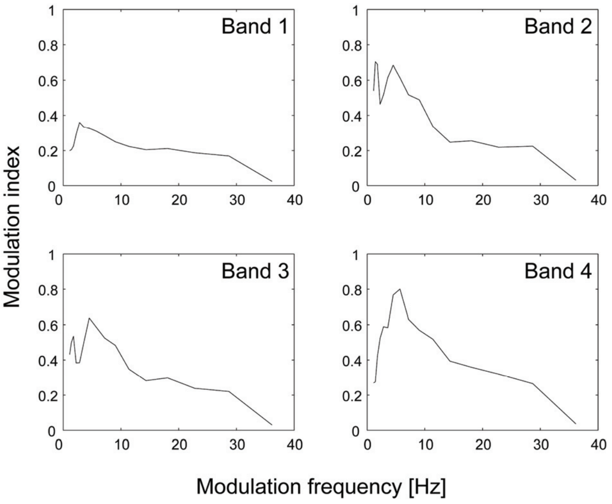

Figure 5: *Modulation spectra of speech signals uttered in reverberation.*

4. Conclusions

In this paper, we reviewed steady-state suppression and compared the modulation spectra of speech signals before and after processing. We also discussed the Lombard-like effect of speech in reverberation from both the speech production and speech perception perspective. We showed that speakers talking in reverberant environments changed the acoustic characteristics of their speech and were more intelligible as compared to speech produced in quiet. In the future, we would like to discuss distinct characteristics of enhanced speech signals and compare them with the results of perceptual experiments. A possible application of these studies would be to develop signal processing of announcements or instructions in all environments, and specifically in public spaces, where relatively higher speech intelligibility is required.

5. Acknowledgments

This work was partially supported by Sophia University Open Research Center from MEXT, Grant of Mitsui Sumitomo Insurance Welfare Foundation (Welfare of the Elderly Section), and Grant-in-Aid for Young Scientists (B) from MEXT (21700203).

6. References

[1] Simpson, A. M., Moore, B. C. and Glasberg, B. R., “Spectral enhancement to improve the intelligibility of speech in noise for hearing-impaired listeners,” *Acta Otolaryngol*, Suppl. 469: 101-107, 1990.

[2] Reed, C. M., Hicks, B. L., Braida, L. D. and Durlach, N. I., “Discrimination of speech processed by low-pass filtering and pitch-invariant frequency lowering,” *J. Acoust. Soc. Am.*, 74: 409-419, 1983.

[3] Yasu, K., Hishitani, M., Arai, T. and Murahara, Y., “Critical-band based frequency compression for digital hearing aids,” *Acoust. Sci. Tech.*, 25(1): 61-63, 2004.

[4] Kulkarni, P. N., Pandey, P. C. and Jangamashetti, D. S., “Multi-band frequency compression to reduce the effects of spectral masking,” *Int. J. Speech Tech.*, 10: 219-227, 2007.

[5] Roucos, S. and Wilgus, A., “High quality time-scale modification of speech,” *Proc. ICASSP*, 10: 493-496, 1985.

[6] Kusumoto, A., Arai, T., Kinoshita, K., Hodoshima, N. and Vaughan, N., “Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environment,” *Speech Communication*, 45(2): 101-113, (2005).

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

[8] Arai, T., Kinoshita, K., Hodoshima, N., Kusumoto, A. and Kitamura, T., “Effects on suppressing steady-state portions of speech on intelligibility in reverberant environments,” *Acoust. Sci. Tech.*, 23(4): 229-232, 2002.

[9] Kobayashi, K., Yasu, K., Hodoshima, N., Arai, T. and Shindo, M., “A study of syllable enhancement for elderly listeners by suppression energy of steady-state portions of vowels,” *J. Acoust. Soc. Jpn.*, 64(5): 278-289, 2008 (in Japanese).

[10] Hodoshima, N., Miyauchi, Y., Yasu, K. and Arai, T., “Steady-state suppression for improving syllable identification in reverberant environments: A case study in an elderly person,” *Acoust. Sci. Tech.*, 28(1), 53-55, 2007.

[11] Nábělek, A. K., Letowski, T. R. and Tucker, F. M., “Reverberant overlap- and self-masking in consonant identification,” *J. Acoust. Soc. Am.*, 86(4): 1259-1265, 1989.

[12] Hodoshima, N., Arai, T., Kusumoto, A. and Kinoshita, K., “Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments,” *J. Acoust. Soc. Am.*, 119(6): 4055-4064, 2006.

[13] Arai, T., Nakata, Y., Hodoshima, N. and Kurisu, K., “Decreasing speaking-rate with steady-state suppression to improve speech intelligibility in reverberant environments,” *Acoust. Sci. Tech.*, 28(4): 282-285, 2007.

[14] Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow R. I. and Stokes, M. A., “Effects of noise on speech production: Acoustics and perceptual analysis,” *J. Acoust. Soc. Am.*, 84, 917-928, 1988.

[15] Junqua, J.-C., “The Lombard reflex and its role on human listeners and automatic speech recognizers,” *J. Acoust. Soc. Am.*, 93, 510-524, 1993.

[16] Lane H. and Tranel B., “The Lombard sign and the role of hearing in speech,” *J. Speech and Hear. Res.*, 14, 677-709, 1971.

[17] Strange, W., Jenkins, J. J. and Johnson, T. L., “Dynamic specification of coarticulated vowels,” *J. Acoust. Soc. Am.*, 74(3): 695-705, 1983.

[18] Hermansky, H. and Morgan, N., “RASTA processing of speech,” *IEEE Trans. Speech Audio Process.*, 2(4): 578-589, 1994.

[19] Arai, T., Pavel, M., Hermansky, H. and Avendano, C., “Syllable intelligibility for temporally filtered LPC cepstral trajectories,” *J. Acoust. Soc. Am.*, 105(5): 2783-2791, 1999.

[20] Kanedera, N., Arai, T., Hermansky, H. and Pavel, M., “On the relative importance of various components of the modulation spectrum for automatic speech recognition,” *Speech Communication*, 28: 43-55, 1999.

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

[22] Greenberg, S. and Arai, T., “What are the essential cues for understanding spoken language?,” *IEICE Trans. on Information and Systems*, E87-D(5): 1059-1070, 2004.

[23] Houtgast, T. and Steeneken, H. J., “A review of the MTF concept in room acoustics and its use of estimating speech intelligibility in auditoria,” *J. Acoust. Soc. Am.*, 77(3): 1069-1077, 1985.

[24] Hodoshima, N., Arai, T. and Kurisu, K., “Speaker variabilities of speech in noise and reverberation,” IEICE Technical Report, SP2009-69: 43-48, 2009 (in Japanese).

[25] Hodoshima, N., Arai, T. and Kurisu, K. “Intelligibility of speech spoken in noise and reverberation.” *Proc. International Congress on Acoustics*, 2010 (to be published).