ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Speech Communication 45 (2005) 101–113

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments

Akiko Kusumoto [a,b,*], Takayuki Arai [b], Keisuke Kinoshita [b], Nao Hodoshima [b], Nancy Vaughan [a]

[a] *VA RR&D National Center For Rehabilitative Auditory Research, Portland VA Medical Center, NCRAR, 3710 SW US Veterans Hospital Road, Portland, OR 97239, USA*

[b] *Department of Electrical and Electronics Engineering, Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo 102-8554, Japan*

Received 16 May 2003; received in revised form 11 February 2004; accepted 28 June 2004

Abstract

Most listeners have difficulty understanding speech in reverberant conditions. The purpose of this study is to investigate whether it is possible to reduce the degree of degradation of speech intelligibility in reverberation through the development of an algorithm. The modulation spectrum is the spectral representation of the temporal envelope of the speech signal. That of clean speech is dominated by components between 1 and 16 Hz centered at 4 Hz which is the most important range for human perception of speech. In reverberant conditions, the modulation spectrum of speech is shifted toward the lower end of the modulation frequency range. In this study, we proposed to enhance the important modulation spectral components prior to distortion of speech by reverberation. Word intelligibility in a carrier sentence was tested with the newly developed algorithm including two different filter designs in three reverberant conditions. The reverberant speech was simulated by convoluting clean speech with impulse responses measured in the actual halls. The experimental results show that modulation filtering incorporated into a pre-processing algorithm improves intelligibility for normal hearing listeners when (1) the modulation filters are optimal for a specific reverberant condition (i.e., $T_{60} = 1.1$ s), and (2) consonants are preceded by highly powered segments. Under shorter (0.7 s) and longer (1.6 s) reverberation times, the modulation filtering in the current experiments, an Empirically-Designed (E-D) filter and a Data-Derived (D-D) filter, caused a slight performance decrement respectively. The results of this study suggest that further gains in intelligibility may be accomplished by re-design of the modulation filters suitable for other reverberant conditions.

[*] Corresponding author. Tel.: +1 503 220 8262x55949; fax: +1 503 273 5021.
E-mail addresses: akiko.kusumoto@med.va.gov (A. Kusumoto), arai@sophia.ac.jp (T. Arai), kinoshita@cslab.kecl.ntt.co.jp (K. Kinoshita), n-hodosh@sophia.ac.jp (N. Hodoshima), vaughann@ohsu.edu (N. Vaughan).

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.06.003

Keywords: Speech enhancement; Modulation spectrum; Modulation transfer function; Reverberation; Word intelligibility

1. Introduction

When people listen to a lecture in a large auditorium, for most listeners, it is difficult to comprehend speech. This is because reverberant sound, which is reflected from the walls or ceiling, masks the direct sound. Since reverberation gives us rich sound, longer reverberation time is preferred for some sounds such as classical and church music. On the other hand, less reverberation is needed in the same auditorium to keep speech intelligibility high during lectures or conversations.

Houtgast and Steeneken (1973, 1980) proposed a method based on the modulation transfer function (MTF) for objectively evaluating the intelligibility of speech in an acoustic path between the sound source and the receiver. They defined the MTF of a sound transmission system as the reduction in the modulation index (m) of the envelope intensity at the output relative to that at the input as a function of modulation frequency. The speech transmission index (STI) is a predictor of speech intelligibility derived from the MTF and it is based on the articulation index (AI) concept introduced by Fletcher and Galt (1950). Steeneken and Houtgast (1980) has shown that the STI is an objective measure to predict speech intelligibility.

The modulation spectrum is defined as spectral representation (e.g., the Fourier transform) of the temporal envelope of the speech signal. In the clean condition, the dominant component of the modulation spectrum of continuous speech lies between 1 and 16 Hz (especially 2–8 Hz) with its peak around 4 Hz in modulation frequency. This reflects the syllabic temporal structure of speech (Greenberg, 1997). It has been reported that the average modulation spectrum is similar irrespective of the language (Arai and Greenberg, 1997). The peak of the modulation spectrum shifts from around 4 Hz in clean speech to around 2 Hz in reverberant speech, and the modulation index decreases. Reverberation acts as a low-pass filter on the modulation spectrum reducing the modulation index (Houtgast and Steeneken, 1985).

It has been confirmed by perceptual experiments that modulation frequencies around 4 Hz (ranging from 1 to 16 Hz) are within the most important range for speech perception for the human auditory system. Drullman et al. (1994a,b) reported that speech intelligibility is not influenced by low-pass (16 Hz cut-off) and high-pass filtering (4 Hz cut-off) of the temporal envelopes. Arai et al. (1996, 1999) expanded Drullman's experiments utilizing band-pass filtering in addition to low-pass filter (LPF) and high-pass filter (HPF). Both experiments concluded that no significant information for perceiving speech is contained below 1 Hz and above 16 Hz on the temporal envelopes. These studies verified the importance of modulation spectral components in speech perception.

The approaches for improving the intelligibility of speech in reverberation are mainly divided into two broad types: One is a post- and the other is a pre-processing technique. Post-processing technique means that signal processing is applied after reverberation to suppress or remove the reverberation effect on speech. This is also known as a dereverberation technique. Some dereverberation techniques using multi-microphone system have been reported, such as the delay-and-sum beamformer (Flanagan et al., 1985), subband envelope estimation (Wang and Itakura, 1991), and a technique that decompose the received microphone signals into minimum-phase and all-pass components (Gonzalez-Rodrigues et al., 2000). Such techniques have shown only modest improvement in terms of reverberation reduction. Unlike those techniques, Langhans and Strube (1982) challengingly proposed a single-microphone dereverberation technique based on the concept of the MTF that utilizes acoustic information. They applied post-processing filters by inverting MTFs, which were derived from the temporal envelopes of the speech. They acquired MTFs utilizing sinusoidally modulated noise rather than speech. Avendano and Hermansky (1996) extended the work of Langhans and Strube and proposed a Data-Derived filter bank technique. They drew inverse modulation transfer functions (IMTF), which were derived from experimental data. In both techniques the envelope of the speech was filtered after applying critical band filters.

Pre-processing technique is the other approach for improving speech intelligibility in reverberant conditions. The speech signal from the microphone is filtered between the microphone and the loudspeakers. In this situation, the intelligibility of the pre-processed reverberant signal would be expected to be more robust than that of the unprocessed-reverberant signal against reverberation. The only previous use of a pre-processing technique was investigated by Langhans and Strube (1982). They used a post-processing algorithm, but adapted it to a pre-processing strategy. They saw hardly any enhancement of the modulation depth in the reverberant signal. Sentence recognition tests did not show significant improvements over the original speech in three types of artificial reverberation with five subjects.

In this paper we report a modulation filtering technique similar to RelAtive SpecTrAl processing (RASTA) which acts as a filter that passes the modulation spectrum range between 1 and 12 Hz (Hermansky and Morgan, 1994). Our filtering as a pre-processing technique is designed to enhance the important modulation spectrum region, which is deteriorated by reverberation effects, in order to reduce the degree of degradation of speech intelligibility in a reverberant environment. The modulation filters were designed based on the work of Langhans and Strube (1982) and of Avendano and Hermansky (1996). In addition, the perceptual experiments were conducted in various reverberant conditions in the current study. The purpose of the present study is to investigate whether it is possible to improve speech intelligibility by modulation filtering in a pre-processing algorithm under reverberant conditions.

2. Methods

In this section, we describe a pre-processing technique for modulation filtering including the procedure for two separate filter designs.

2.1. Signal processing for modulation filtering

While there is a variety of ways to design a modulation filter, the proposed technique in this study is similar to RASTA processing (Hermansky and Morgan, 1994) used in automatic speech recognition (ASR). The block diagram for modulation filtering is illustrated in Fig. 1. The algorithm enhances the modulation index prior to the reverberation effect in order to increase modulation depth that typically decreases due to reverberation. We focused on modifying the power spectral envelope, because it contains cues for human speech perception (Shannon et al., 1995). Although the human auditory system is relatively insensitive to phase of the original waveform (Fant, 1960), phase of the temporal envelope of speech can play a significant role in intelligibility (Greenberg and Arai, 2001). Therefore, we applied a linear-phase finite-impulse-response (FIR) filter for modulation filtering to preserve the phase of the temporal envelope. The input speech signal, sampled at 16 kHz, is divided into separate frequency channels utilizing constant-Q band-pass filters (BPF 1–16 in Fig. 1). In each channel an amplitude envelope is extracted by the Hilbert transform. The remaining carrier term to be multiplied with the resultant envelope is unchanged. Each envelope is smoothed by a lowpass FIR filter. Further, the sampling rate is downsampled by $1/M$ (the factor of $M = 160$, sampling rate 100 Hz). Then, the modulation filters 1–16 in Fig. 1 are applied as a 65-tap FIR filter. After modulation filtering, the sampling rate is upsampled to get back to the original sampling rate by the same factor of M. The negative value is removed by applying half-wave rectification. The resultant signal is multiplied by the carrier term that is stored from the envelope extraction. Then, to remove frequency components outside of the range of the channel, the same BPF that is used to divide the signal into frequency channels is applied. Finally the output signal is re-synthesized by summing up all the processed signals from each channel. In this present study, two filter designs were investigated and are described in the following section.

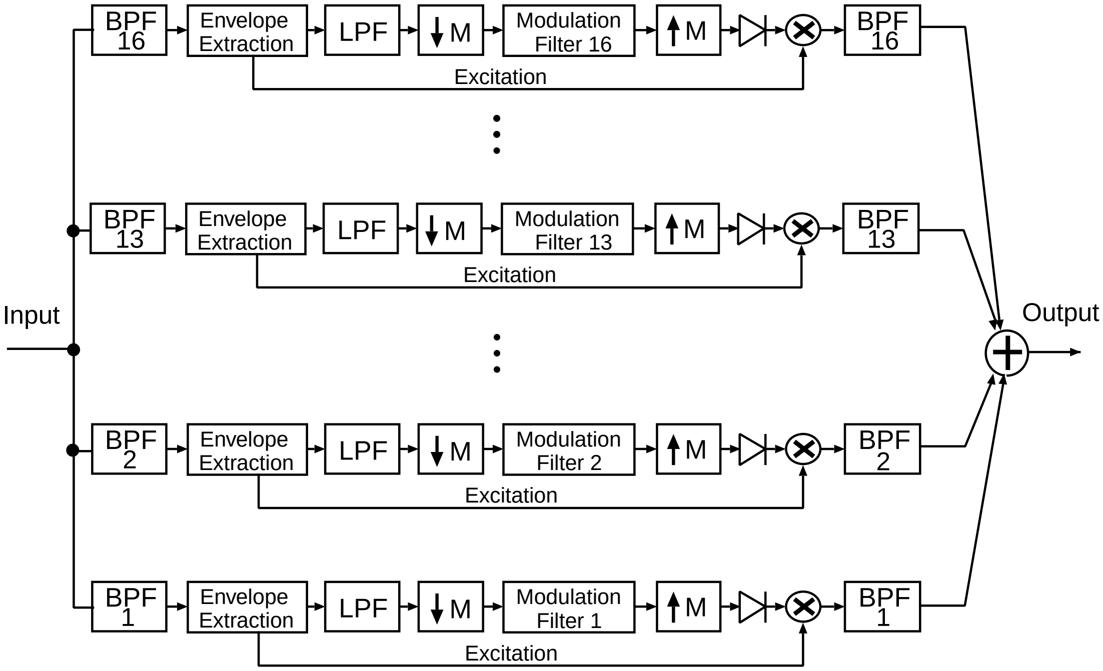

Fig. 1. A block diagram of the sequence of processing stages for modulation filtering. The spectrum of the input signal is divided into 16, 1/3-octave channels by band-pass filters (BPF). In each channel the envelope is extracted by the Hilbert transform, then lowpass filtered, down-sampled by a factor of 160, and the modulation filter is applied. The modulation filters 1–16 corresponding to the channel are applied based on the two different filter designs whose frequency responses are shown in Figs. 2 and 3. After upsampling, and half-wave rectification, the output of each channel is multiplied with the unmodified remaining carrier terms from the envelope extractions. Then, remaining frequency components outside of the channel (BPF) are removed and the final output signal is re-synthesized by adding all the processed signals from each channel.

2.2. Filter designs

We designed the modulation filters applied in Fig. 1 to enhance the important spectral components for human perception prior to the reverberation effect. It is our assumption that the modulation spectrum of the pre-processed speech, when combined with reverberation, approaches the modulation spectrum of clean speech and helps to improve speech intelligibility. Two different filter designs are presented: the Empirically-Designed (E-D) and the Data-Derived (D-D) filter described in Sections 2.2.1 and 2.2.2.

2.2.1. Empirically-Designed filter (E-D Filter)

The E-D filter enhances components around 0–8 Hz with a peak at 4 Hz to get close to the shape of the original modulation spectrum in reverberant conditions. Fig. 2 shows the symmetrical frequency response of the E-D filter on a linear scale to be applied in modulation filters 1–16 in Fig. 1. It is designed as a FIR filter based on the important modulation frequencies of the speech envelope using a Hamming window. The filter has the same frequency response for all channels 1–16.

2.2.2. Data-Derived filter (D-D Filter)

In contrast to the E-D filter, the Data-Derived filter (D-D filter) recovers the original modulations based on the MTF data derived from clean and reverberant speech. Avendano and Hermansky (1996) explored dereverberation techniques using post-processing algorithms. They applied filters to temporal envelope for each frequency channel to recover the modulations presented in the original speech. These filters were designed based on inverse MTFs (IMTF) between the filtered short time power trajectories (STPT's) of the reverberant

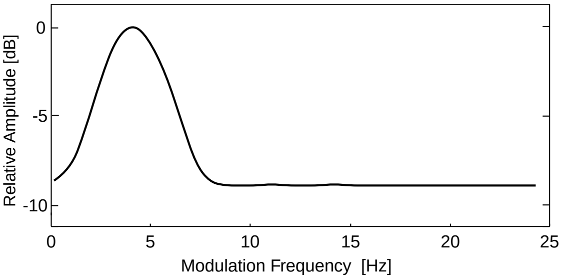

Fig. 2. Frequency response of the Empirically-Designed (E-D) filter as a function of modulation frequency. The E-D filter was applied to all modulation filters 1–16 in Fig. 1 used in the experiment.

speech and the corresponding desired trajectories of clean speech. Similar to Avendano's approach, in this study we designed the modulation filter in Fig. 1 computing IMTF from the modulation spectrum of reverberant speech and that of the corresponding clean speech.

We designed the D-D filter in the following way: First we calculated the MTF for each frequency band (band 1: 0–800 Hz (channels 1–6), band 2: 800–1600 Hz (channels 7–9), band 3: 1600–3200 Hz (channels 10–12), band 4: 3200–8000 Hz (channels 13–16)) following Arai and Greenberg (1998) in clean and reverberant conditions. Next, the MTF for each band was averaged over the 300 words with the same carrier phrase. Then, the frequency responses of the modulation filters were obtained by inverting the MTF and taking a moving average of the derived MTF. After that, a Hamming window was used to create the final FIR filter. Fig. 3 shows the frequency responses of the Data-Derived filters. The D-D filter is dependent on the reverberation condition and is therefore variable. In other words, these filters need to be designed for each reverberant condition. Whereas the same modulation filters are applied for each frequency channel using the E-D filter described in Section 2.2.1, in the D-D filter, on the other hand, a different frequency response is applied to each of four corresponding bands.

2.3. Effects of modulation filtering on speech

Fig. 4 represents the average modulation spectra of 300 words and carrier sentences before (Clean) and after (Org) reverberation effects (T_{60} = 1.1 s). They were analyzed in 4 bands (band 1: 1–800 Hz, band 2: 800–1600 Hz, band 3: 1600–3200 Hz, band 4: 3200–8000 Hz). It can easily be seen that after reverberation the modulation indices at the peaks are severely reduced and modulation frequencies of the peaks are less than 2 Hz in all bands as Houtgast et al. stated in (Houtgast and Steeneken, 1985).

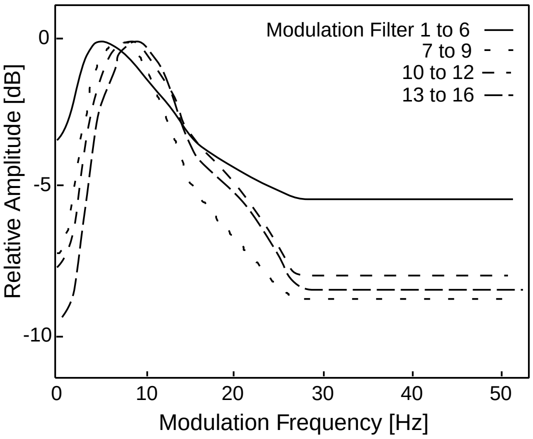

Fig. 3. Frequency responses of the Data-Derived (D-D) filters as a function of modulation frequency. The D-D filters were applied to modulation filters 1–16 in Fig. 1. In this case, four frequency bands were derived for modulation filters 1–6, 7–9, 10–12, and 13–16. The frequency responses were variable depending on the halls.

In Fig. 5 the changes in the modulation spectra of clean speech by the E-D filter before reverberation effects are shown in the same format as Fig. 4 except the solid line represents the processed speech (E-D Filter) and dashdot line represents Clean. Taking a closer look at these peaks in the figure, it can be seen that the indices of the modulation spectra using the E-D filter are intensified as contrasted with those of Clean. This is because we designed the modulation filters to compensate for the decrease in modulation indices due to reverberation effect.

Fig. 6 represents the modulation spectra of speech with- and without the E-D filter after reverberation effects (Org) rather than before (T_{60} = 1.1 s). Comparing the results in Figs. 5 and 6 demonstrates that the addition of reverberation in Fig. 6 reduces the over all effect of the E-D filter. How-

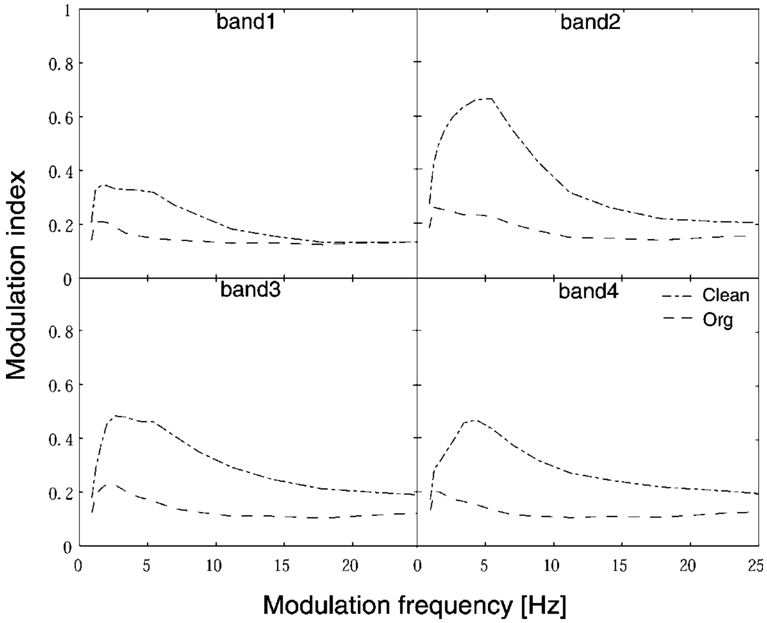

Fig. 4. Average modulation spectra of 300 rhyming-words including the carrier sentence for each frequency band before (dashdot lines) and after (dashed lines) reverberation effects at T_{60} = 1.1 s (Rev B). Frequency regions are constituted with band 1 (0–800 Hz), band 2 (800–1600 Hz), band 3 (1600–3200 Hz), and band 4 (3200–8000 Hz). Speech before reverberation is referred to as Clean, and after reverberation is referred to as Org.

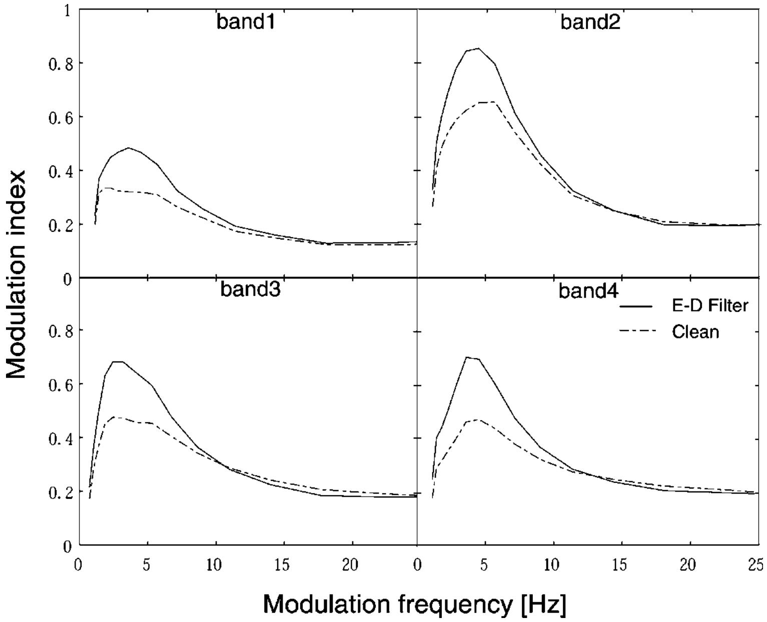

Fig. 5. Average modulation spectra of 300 rhyming-words including the carrier sentence for each frequency band before reverberation effects (Rev B). The E-D filter (solid lines) and Clean (dash dot lines) are compared before reverberation in each frequency. Frequency regions are constituted with band 1 (0–800 Hz), band 2 (800–1600 Hz), band 3 (1600–3200 Hz), and band 4 (3200–8000 Hz).

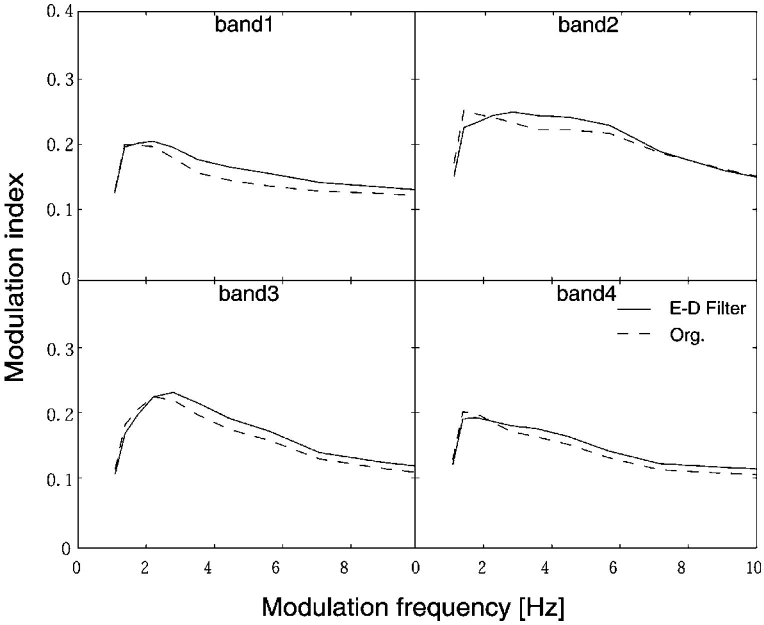

Fig. 6. Average modulation spectra of 300 rhyming-words including the carrier sentence for each frequency band after reverberation effect. The E-D filter (—) and Org (– – –) are compared after reverberation in each frequency band. Frequency regions are constituted with band 1 (0–800 Hz), band 2 (800–1600 Hz), band 3 (1600–3200 Hz), and band 4 (3200–8000 Hz).

ever, in Fig. 6 the E-D filter still provides some enhancement around the 4 Hz regions, particularly in bands 1 and 2 although the effect is not as pronounced as in Fig. 5. Recall that this is the important range for human speech perception.

3. Perceptual experiment

Perceptual experiments were conducted to examine whether it is possible to reduce the degree of degradation of intelligibility with pre-processed speech. We hypothesized that, under a specific reverberant condition, the word intelligibility of modulation-filtered speech would be better than that of the unprocessed speech. It was the purpose of the experiments to clarify the relationship between modulation filtering and reverberation time by seeking the optimal modulation filter suitable for a specific reverberant condition. For that purpose, we used both the E-D and D-D filter designs in several reverberant conditions to see which combinations of filter designs and reverberation yield greater improvement.

3.1. Subjects

Thirty-two people who were all native speakers of American English participated in this experiment. None of them reported hearing loss. The age limit was set between 20 and 40 years in order to avoid hearing loss due to aging. All participants signed the consent form explaining the aim and the required time and monetary reward for this experiment.

3.2. Speech material

The Modified Rhyme Test (MRT), originally developed by House et al. (1965), is a closed set test consisting of six different lists (List A–F), each composed of 50 American English monosyllabic words, which were either consonant–vowel–consonant (C_1VC_2), consonant–vowel (C_1V), or vowel–consonant (VC_2). The speech material was recorded in our lab by a male who is a native speaker of American English. The speech signal was sampled at 44.1 kHz and quantized with 16 bit resolution. The stimuli were downsampled to

16 kHz. We used the version of the MRT in which several words were replaced by Kruel et al. (1968) compared with the original test. "*Number*—You will mark the *WORD*, please." was the carrier phrase. The MRT was chosen because the target words and the foils were the same structure. For example, when the word-initial consonant is the discrimination task, the VC_2 within the target word stays the same in all six answer choices. Similarly, for word-final consonant discrimination tasks, the C_1V stays the same. It is also important to have one carrier phrase so that the segment preceding the target does not vary. This is because the target word is always overlaid by the reverberant components of the previous segment (overlap masking).

3.3. Processed speech stimuli

Reverberation times (Table 1) were measured in three auditoria in Japan: Hamming Hall in Higashi-yamato City (Rev A), Hamming Hall in Higashi-yamato City without reflection boards (Rev B), and Strings booth in the NHK Hall (Rev C). The early decay time (EDT), which was used to extrapolate reverberation time by measuring the required time for the first 10 dB down of the decay curve, was employed for measuring reverberation time (Jordan, 1980). The speech signals (Clean) and also processed by the Empirically-Designed filter, and the Data-Derived filter were convoluted with the impulse response data from each of the three auditoria. After convolution the three processing types were designated Org, E-D Proc and D-D Proc, respectively. This resulted in nine sets of stimuli (three processing types × three Rev) utilized in this experiment.

Table 1
Reverberation times of three conditions used in the experiments

Reverberation	T_{60} (s)
Rev A	1.6
Rev B	1.1
Rev C	0.7

Reverberant signals were simulated by convoluting stimuli with impulse response data, which was obtained from actual halls.

3.4. Procedures and apparatus

Each subject was tested with three lists of the MRT consisting of 50 items each. Each list was assigned to one of the reverberant conditions such as Rev A, B and C. The initial or final consonant discrimination tasks were randomized in order to reduce a potential learning effect. Within each list, all three processing types (Org, E-D Proc and D-D Proc) were counterbalanced across subjects, so that by the end of the experiment, each processing type was heard an equal number of times in each reverberant condition. Reverberation times were always presented from shortest to longest in order to familiarize subjects gradually with increasing reverberation time.

The six-rhyming-word groups were provided on the answer sheet as choices. Subjects were instructed to choose the most likely stimulus out of six choices after listening to each presentation. Subjects operated the experimental interface on the laptop PC that was set up by the Matlab graphical user interface (GUI) and a mouse input device. Each subject listened to the stimuli at his/her most comfortable level through digital stereo headphones. Before the tests, practice sessions were held to adjust the volume of stimuli and to familiarize subjects with the operation of the GUI. Subjects advanced through the experiment at their own pace by clicking on the "Play again" button or the "Next" button with the mouse. They were asked in advance to pay attention to the target "*WORD*". They were instructed to choose their answer based on the first impression, although they were able to repeat each sentence as many times as they wanted. After each list was finished, subjects were asked to take a 5-min break.

4. Experimental results

Fig. 7 shows the subjects' performance for word initial consonant tasks in three simulated reverberant conditions (labeled Rev A, Rev B, and Rev C on the abscissa). The parameter in this figure is the filter types applied to the speech stimuli. Filled symbols represent the unprocessed reverberated

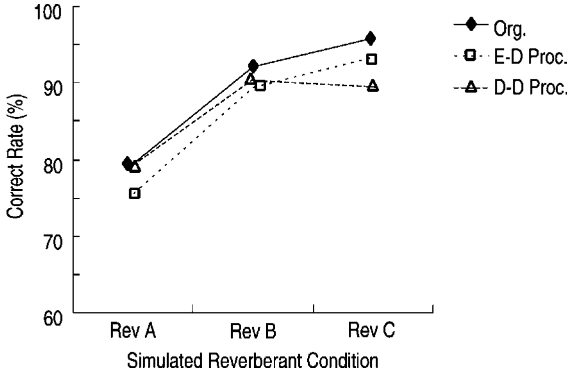

Fig. 7. Mean percent correct recognition scores for consonants in word-initial position from 32 subjects with three types of processing (filled diamond: unprocessed signal (Org), open square: processed signal (E-D Proc), open triangle: processed signal (D-D Proc)) in three simulated reverberant condition labeled Rev A ($T_{60} = 1.6$ s), Rev B ($T_{60} = 1.1$ s) and Rev C ($T_{60} = 0.7$ s).

utterance (Org). Open symbols represent the two filter designs used to process the speech stimuli, E-D Proc and D-D Proc. Results show that for consonants in word-initial position, processing with either E-D Proc and D-D Proc did not improve performance. There was a slight, but significant decrease in performance for D-D Proc compared to Org for Rev C, in which reverberation time was shortest ($t(23) = 2.065$, SE = 0.0299, $p = 0.05$).

Fig. 8 follows the same format as Fig. 7, except performance is shown for consonants in word-final position. For final consonants, there was approximately a 6% improvement in word discrimination for both E-D Proc and D-D Proc compared to unprocessed speech for condition Rev B. These 6% improvements, however, were not statistically significant ($p = 0.298$ (E-D), 0.270 (D-D)). For Rev A, in which reverberation time was longest, E-D Proc reduced performance significantly ($t(23) = 2.078$, SE = 0.0521, $p = 0.049$).

Comparing results in Figs. 7 and 8 suggests that the effect of the processing on initial and final consonant discrimination depends on the particular reverberant condition used. The E-D Proc caused a slight performance decrement in the longest reverberation time tested (Rev A), whereas D-D Proc reduced performance in the shortest reverberation time (Rev C). There is a tendency for improvement in word intelligibility in the middle range of reverberation time (Rev B). Results did not strongly support the hypotheses that modulation filtering improves consonant discrimination at the word level for reverberant speech. In our initial analysis, confusion matrices did not reveal any consistent confusion between Org and processed results. However, the results suggest that modulation filtering maybe useful for a limited range of reverberation times; i.e., between 0.7 and 1.6 s. In this experiment the MRT rhyming-words lists were presented in order from short to long reverberation times. The task with the long reverberation time was a challenging task, therefore the subjects were presented the short reverberation time first to familiarize them with reverberant speech. Otherwise the results with the longer reverberation time (Rev A) would have been even worse than the results in current study.

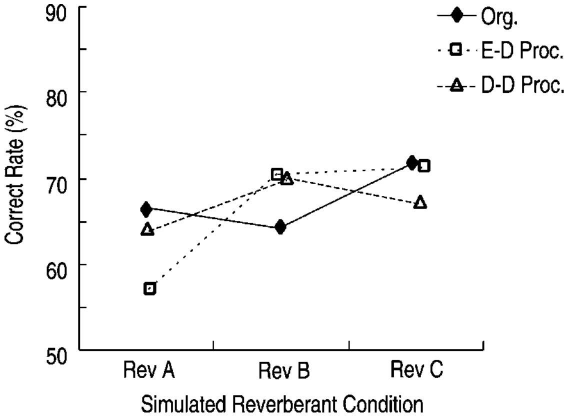

Fig. 8. Mean percent correct recognition scores for consonants in word final position from 32 subjects with three types of processing (filled diamond: unprocessed signal (Org), open square: processed signal (E-D Proc), open triangle: processed signal (D-D Proc)) in three simulated reverberant condition labeled Rev A ($T_{60} = 1.6$ s), Rev B ($T_{60} = 1.1$ s) and Rev C ($T_{60} = 0.7$ s).

4.1. Discussion

In the perceptual experiment with 32 listeners in three reverberant conditions, modulation filtering did not produce a significant improvement in word intelligibility. However, for consonant-final-position discrimination tasks, both E-D Proc and D-D Proc resulted in improved mean performance levels for the condition in which reverberation time was 1.1 s (Rev B). Perhaps the most useful findings of this study are that for both word-initial and word-final consonant discrimination tasks, E-D Proc did not adversely affect word intelligibility for short ($T_{60} = 0.7$ s) reverberation times and D-D Proc did not adversely affect word intelligibility for long ($T_{60} = 1.6$ s) reverberation times. This suggests that future efforts to modify each filter based on each reverberation time could further improve their effectiveness.

The effectiveness of modulation filtering varied depending on whether the consonant task was word-initial or word-final. Recall that the both E-D and D-D pre-processed speech in Rev B provided about 6% improvement when the consonants differed as to final position, but not initial position. The overlap-masking of reverberation could explain why the processing was effective in the word-final position. Overlap-masking occurs when an initial segment of a word is masked by a preceding segment in a reverberant environment (Bolt and MacDonald, 1949; Náběelek and Robinette, 1978; Náběelek et al., 1989). Knudsen (1929) conducted recognition tests with Consonant–Vowel–Consonant (CVC) syllables in a reverberant environment and reported a greater degradation of speech intelligibility in final consonants than in initial consonants. The study showed that the pre-ponderance of errors among final consonants was the result of overlap-masking produced by the reverberation of the preceding vowel. In our experiment, within C_1VC_2, reverberating C_1V masks the final consonant C_2. The previous segment V before the final consonant C_2 is a vowel containing high power. The modulation filtering in this study may suppress the power of the steady state of C_1VC_2 segment in order to reduce reverberating C_1V components. Reducing masker components, which mask the final consonant C_2, can improve intelligibility.

In contrast, when the initial consonant is the test item, the effectiveness of the modulation filtering is reduced because there is little power in the previous segment. That is because the /schwa/, a highly reduced vowel, occurs immediately before the initial consonant in the word, "the" ("you will mark the *WORD*, please"). The apparent difference in the masking effect of speech segments prior to the initial and final consonants may account for the difference in the effectiveness of the modulation filtering in these two discrimination tests.

Arai et al. (2002) reported a technique for reducing the masking effect similar to the effect of modulation filtering. In this technique the parameter D, the mean square of the regression coefficients for each time trajectory of the logarithmic envelope in each band (1/3-octave band), is defined to represent the degree of steadiness of speech following Furui (1986). The steady-state portions of speech are detected when the parameter D is under a pre-determined threshold. Then, the power of those segments is suppressed to prevent masking of the preceding segments. Hodoshima et al. (2002, 2003) conducted the perceptual experiments to confirm the effectiveness of Arai's technique using Japanese consonant-vowel (CV) syllables in a carrier phrase under the following reverberant condition: 0.8, 0.9, 1.0, 1.1 and 1.2 s. The results showed that the steady-state suppressed speech was significantly intelligible as opposed to the unprocessed speech.

Modulation filtering enhances the modulation component 0–8 Hz in each frequency channel using a linear process on the envelope. In contrast, steady-state suppression attenuates the temporal envelope by non-linear processing. However, modulation filtering for enhancement around the 0–8Hz regions is similar to the steady-state suppression in that it effectively suppresses the dc components of the temporal envelope. It can be said, then, that modulation filtering does suppress steady-state components substantially. Hence, effectiveness of speech processed by modulation filtering and by steady-state suppression should be compared with the same subjects in future investigations.

In the future, we plan to investigate the relationship between the effects of overlap-masking and modulation filtering. While we tested word intelligibility with one carrier phrase, we will try the same approach with a variety of carrier phrases to test the effects of the phoneme that occurs immediately before the target word. In addition, the experimental results indicate that modulation-filtering is effective for a limited range of reverberation times between 0.7 and 1.6 s. As stated above, future efforts will be directed toward redesigning each filter to modify the E-D filter for shorter reverberation and the D-D filter for longer reverberation.

5. Implementation in hearing-impaired listeners

The difficulties of understanding reverberant speech are much more serious for people who are hard of hearing because of aging, and for people who wear hearing aids to make use of their residual hearing. Disturbances such as reverberation or noise deteriorate their hearing ability, even though they can hear well at a short distance. In a large auditorium the amplification system is often a built-in FM or infra-red transmitter that works with or without hearing aids. For hearing aids the system operates on direct audio input which eliminates the effects of room acoustics on speech intelligibility.

5.1. Pilot study with hearing-impaired listeners

We examined in a small sample of four hearing-impaired subjects the effectiveness of modulation filtering as a pre-processing method as a small pilot test. Participants who were long-term hearing-aid users were four Japanese females with severe sensorineural hearing loss. Table 2 shows their self-reported hearing levels for each hearing-impaired subjects (unaided). The Japanese sentence, “Terebi gemu ya pasokon de gemu o shite asobu (playing with video games and PCs)” was employed as the speech material. It was processed with the Empirically-Designed (E-D) filter described in Section 2.2.1. The experiment was conducted at the St. Ignatius church in Tokyo. In the Cathedral the measured reverberation time was approximately 3.1–3.2 s across all frequency regions. Participants sat in the cathedral and listened to stimuli from loudspeakers located on the ceiling. They wore their hearing aids. During the test, participants were forced to choose one of two stimuli on the basis of which one was easier to hear. As a result of this preference test, all four hearing impaired listeners preferred processed speech.

Table 2
Self-reported hearing level of hearing-impaired listeners who participated in a small pilot study (unaided)

Subject	Left (dBHL)	Right (dBHL)
A	75	75
B	83	82
C	105	95
D	105	95

5.2. Discussion

The results of the preference test with hearing-impaired listeners indicated that the pre-processed speech could facilitate comprehension of reverberant speech. Notice that they preferred the pre-processing stimuli with much longer reverberation times than were used with normal hearing listeners. This suggests that testing with much longer reverberation times than 1.6 s might be useful. It is difficult to ascertain whether, in fact, hearing-loss or longer reverberation times or some combination of both variables was responsible for the perceived improvement in intelligibility. The results were not quantitative, and statistical analyses were not conducted because of the small sample population. This finding leads to further investigation of the effectiveness of current algorithms with other test procedures for hearing-impaired and elderly listeners. Future investigations of the effectiveness of the current algorithms should focus on other speech test procedures and comparisons of hearing-impaired and normal-hearing, young and old listeners.

Hearing-impaired listeners are affected more by reverberation than normal hearing people, so the effectiveness of processing may be observed more readily in that population. The advantage of the pre-processing strategy is that speech intelligibility can be improved without the use of assistive listening devices besides hearing aids for events such as lectures in large auditoria. The methods for improving speech intelligibility in this study are notable for application to hearing aids users and elderly listeners without hearing aids, in particular.

A number of studies have examined the effects of reverberation on children and elderly listeners (e.g., Nábělek and Robinson, 1982), non-native listeners (e.g., Nábělek and Donahue, 1984), and hearing-impaired listeners (e.g., Nábělek and Pickett, 1974). Recognition accuracy decreases in all cases compared with listeners with normal hearing.

6. Conclusions

When speech is heard in a large hall with reverberant effects, the audience has difficulty understanding the contents due to speech intelligibility degradation. In this study an algorithm for signal processing was proposed to reduce the degree of degradation of speech intelligibility prior to reverberation effects. The goal of the algorithm was to enhance the important components of the modulation spectrum for human perception. Two different filter designs were incorporated in the algorithm. The Empirically-Designed (E-D) filter was designed to enhance the 0–8 Hz regions with a 4 Hz peak in the modulation spectrum domain based on degradation of the modulation spectrum by reverberation previously described by Houtgast and Steeneken (1985). The second method is the Data-Derived (D-D) filter developed to recover the original modulations employing inverse modulation transfer function (IMTF) on the basis of derived data. We conducted perceptual experiments to examine whether word intelligibility was improved by modulation filtering designed in this current study. Word intelligibility in a carrier sentence was improved approximately 6% for normal hearing listeners using both E-D Proc and D-D Proc under 1.1 s reverberation time when the word-final consonant discrimination was a task. We concluded that modulation filtering by pre-processing is effective under the following conditions:

(1) When the modulation filters are optimal for specific reverberation times.
(2) When consonants are preceded by highly powered segments.

In other reverberation conditions, the effect of modulation filtering was minimal with both E-D and D-D filtering. The E-D filtering caused a slight performance decrement in long reverberation time (1.6 s), whereas D-D filtering reduced performance in short reverberation time 0.7 s. The word intelligibility results of the study suggest the need for further investigations to re-design the modulation filters suitable for other specific reverberant condition. These filters could be compared with other technique such as steady-state suppression.

The informal pilot study with hearing-impaired listeners showed preferences for the pre-processed speech sentences. This preference suggests that processed speech may provide more benefit for hearing-impaired than for normal hearing listeners. Further experiments will be conducted with subjects with hearing disabilities. Improvement of speech intelligibility in reverberant condition for each group of people will be a significant challenge in the future.

Acknowledgements

We would like to thank the members of the Anthropic Signal Processing Group (ASP) and Center for Spoken Language Understanding (CSLU) at the Oregon Graduate Institute (OGI, currently Oregon Health and Science University, OGI school of Science and Engineering) for their assistance in data collection for this project. Our gratitude is further extended to Hynek Hermansky, Professor at OGI, for his useful comments. We would also like to thank Hideki Tachibana, Kanako Ueno and Sakae Yokoyama at the University of Tokyo for providing the impulse response data. We are grateful to Chikashi Michimata of the Department of Psychology at Sophia University for his guidance on statistical analyses. A part of this research was supported by Grants-in-Aid for Scientific Research (A-2,16203041) from the Japan Society for the Promotion of Science.

References

Arai, T., Greenberg, S., 1997. The temporal properties of spoken Japanese are similar to those of English. In: Proc. of Eurospeech, pp. 1011–1014.

Arai, T., Greenberg, S., 1998. Speech intelligibility in the presence of cross-channel spectral asynchrony. In: Proc. IEEE Internat. Conf. on Acoust., Speech and Signal Process. (ICASSP), pp. 933–936.

Arai, T., Kinoshita, K., Hodoshima, N., Kusumoto, A., Kitamura, T., 2002. Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments. Acoust. Sci. Technol. 23 (4), 229–232.

Arai, T., Pavel, M., Hermansky, H., Avendano, C., 1996. Intelligibility of speech with filtered time trajectories of spectral envelopes. In: Proc. Internat. Conf. on Spoken Lang. Process. (ICSLP), pp. 2490–2493.

Arai, T., Pavel, M., Hermansky, H., Avendano, C., 1999. Syllable intelligibility for temporally filtered LPC cepstral trajectories. J. Acoust. Soc. Amer. 105 (5), 2783–2791.

Avendano, C., Hermansky, H., 1996. Study on the dereverberation of speech based on temporal envelope filtering. In: Proc. Internat. Conf. on Spoken Lang. Process. (ICSLP), pp. 889–892.

Bolt, R.H., MacDonald, A.D., 1949. Theory of speech masking by reverberation. J. Acoust. Soc. Amer. 21, 577–580.

Drullman, R., Festen, J.M., Plomp, R., 1994a. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Amer. 95, 2670–2680.

Drullman, R., Festen, J.M., Plomp, R., 1994b. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Amer. 95, 1053–1064.

Fant, G., 1960. Acoustic Theory of Speech Production. Mouton.

Flanagan, J.L., Johnson, J.D., Zahn, R., Elko, G.W., 1985. Computer-steered microphone arrays for sound transduction in large rooms. J. Acoust. Soc. Amer. 78 (5), 1508–1518.

Fletcher, H., Galt, R.H., 1950. The perception of speech and its relation to telephony. J. Acoust. Soc. Amer. 22, 89–151.

Furui, S., 1986. On the rule of spectral transition for speech perception. J. Acoust. Soc. Amer. 80, 1016–1025.

Gonzalez-Rodrigues, J., Sanchez-Bote, J.L., Ortega-Garcia, J., 2000. Speech dereverberation and noise reduction with a combined microphone array approach. In: Proc. IEEE Internat. Conf. on Acoust., Speech and Signal Process. (ICASSP), pp. 953–956.

Greenberg, S., 1997. On the origins of speech intelligibility in the real world. In: Proc. of ESCA Workshop on Robust speech recognition for unknown communication channels, France, pp. 23–32.

Greenberg, S., Arai, T., 2001. The relation between speech intelligibility and the complex modulation spectrum. Proc. of the European Conf. on Speech Communication and Technology (Eurospeech), 1, pp. 473–476.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. on Speech and Audio Process. 2 (4), 578–589.

Hodoshima, N., Inoue, T., Arai, T., Kusumoto, A., 2002. Suppressing steady-state portions of speech for improving intelligibility in various reverberant environments. Proc. of China-Japan Joint Conf. on Acoust. pp. 199–202.

Hodoshima, N., Arai, T., Inoue, T., Kinoshita, K., Kusumoto, A., 2003. Improving speech intelligibility by steady-state suppression as pre-processing in small to medium sized halls. Proc. of the European Conf. on Speech Communication and Technology (Eurospeech), pp. 1365–1368.

House, A.S., Williams, C.E., Hecker, M.H.L., Kryter, K.D., 1965. Articulation testing methods: consonantal differentiation with a closed-response set. J. Acoust. Soc. Amer. 37, 158–166.

Houtgast, T., Steeneken, H.J.M., 1973. The modulation transfer function in room acoustics as a predictor of speech intelligibility. Acoustica 28, 66–73.

Houtgast, T., Steeneken, H.J.M., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J. Acoust. Soc. Amer. 77 (3), 1069–1077.

Houtgast, T., Steeneken, H.J.M., Plomp, R., 1980. Predicting speech intelligibility in rooms from the modulation transfer function I. General room acoustics. Acoustica 46, 60–71.

Jordan, V.L., 1980. Acoustical Design of Concert Halls and Theatres. Applied Science Publishers.

Knudsen, V.O., 1929. The hearing of speech in auditoriums. J. Acoust. Soc. Amer. 1, 56–82.

Kruel, E.J., Nixon, J.C., Kryter, K.D., Bell, D.W., Lang, J.S., 1968. A proposed clinical test of speech discrimination. J. Speech Hear. Res. 11, 536–552.

Langhans, T., Strube, H.W., 1982. Speech enhancement by nonlinear multiband envelope filtering. In: Proc. IEEE Internat. Conf. on Acoust., Speech and Signal Process. (ICASSP), pp. 156–159.

Nábělek, A.K., Donahue, A.M., 1984. Perception of consonants in reverberation by native and non-native listeners. J. Acoust. Soc. Amer. 75 (2), 632–634.

Nábělek, A.K., Letowski, T.R., Tucker, F.M., 1989. Reverberant overlap- and self-masking in consonant identification. J. Acoust. Soc. Amer. 86, 1259–1265.

Nábělek, A.K., Pickett, J.M., 1974. Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing impaired listeners. J. Speech Hear. Res. 17 (4), 724–739.

Nábělek, A.K., Robinette, L., 1978. Influence of precedence effect on word identification by normally hearing and hearing-impaired subjects. J. Acoust. Soc. Amer. 63, 187–194.

Nábělek, A.K., Robinson, P.K., 1982. Monaural and binaural speech perception in reverberation for listeners of various ages. J. Acoust. Soc. Amer. 71 (5), 1242–1248.

Shannon, R.V., Zeng, F.G., Wygonski, J., Kamath, V., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304.

Steeneken, H.J.M., Houtgast, T., 1980. A physical method for measuring speech-transmission quality. J. Acoust. Soc. Amer. 67, 318–326.

Wang, H., Itakura, F., 1991. An approach of dereverberation using multimicrophone sub-band envelope estimation. In: Proc. IEEE Internat. Conf. on Acoust., Speech and Signal Process. (ICASSP), pp. 953–956.