Applied Acoustics 70 (2009) 689–694

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Using HDC to evaluate signal similarity for information masking

Satoru Gotoh [a,*], Mitsuo Matsumoto [b], Yoshio Yamasaki [a]

[a] Graduate School of Global Information and Telecommunication Studies, Waseda University, Honjo, Saitama 367-0035, Japan
[b] 3-15-17 Nanakuma, Fukuoka-shi, Fukuoka-ken, Japan

ARTICLE INFO

Article history:
Received 23 November 2007
Received in revised form 9 June 2008
Accepted 21 August 2008
Available online 26 October 2008

Keywords:
Informational masking
Sound similarity
Speech interference

ABSTRACT

In the many studies done on informational masking, interfering speech reduces speech intelligibility. This effect is often used to secure privacy in public spaces. These applications require estimates of how much masking is required. In general, masking effects are estimated by using spectrum information as excitation patterns. However, estimates of informational masking can hardly be obtained by only using spectrum information. Therefore, we estimated the effects of informational masking using time-domain information. Then, we calculated the cepstra of the envelopes' magnitude histograms. If these cepstra are different between the target and the masker, the signals are not similar in the time-domain. Furthermore, the effect of informational masking would be low. Therefore, we considered the histograms' cepstra distances (HCD) to estimate signal similarities. The signal similarities in our first experiment were estimated using five maskers by utilizing the HCD. These maskers were random noise, music, female speech, male speech, and target speaker's speech. Male and female speech were more similar to the target speech than music and noise. Also, the same speaker's speech was the most similar in the set of maskers. A listening test was carried out in the second experiment to verify the HCD. A double masker was used in this experiment as an effective informational masker. It has similar characteristics to reversal speech. The listening test results suggest the double-masker's masking effects has the same relation with HCD. This suggests informational masking can be estimated by signal similarity using the HCD.

1. Introduction

Masking effects are used to secure privacy in public spaces [11]. Two kinds of masking effects, energetic masking and informational masking are often applied. It seems that informational masking is more effective than energetic masking for masking speech. Energetic masking is generally a quantity that relates well to the signal-to-noise ratio (SNR). Informational masking, on the other hand, degrades target detection when a listener's auditory processing fails to separate the target's audio stream from that of the interference [5,1,9].

The mechanism for speech perception in the auditory systems is still an extensive topic of research. Masking effects are also closely related to the mechanism in the auditory system. Masking effects are generally estimated by SNR in the frequency domain that is related to excitation patterns [4]. SNR is an adequate measure to estimate effect of masker from energetic point of view. However, if a masker has similar frequency characteristics to that of the target speech, for example, the reversal speech [8] which is well known as an effective informational masker, is a masker; human perception can hardly separate both speeches even at SNR of 0 dB. Therefore, the mechanism for masking effects in the auditory system is not only related to spectrum features but also to temporal changes in the time-domain waveform. Signal similarity in informational studies is considered to be closely related to informational masking. For example, Durlach et al. reported that decreasing target-masker similarity tends to reduce the masking effects [3]. Therefore, we investigated the informational masking effect from the time-domain-similarity point of view.

The envelopes of time-domain waveforms represent temporal changes in signal dynamics. In addition, narrow-band temporal envelopes are closely related to speech intelligibility [6]. Therefore, we considered signal similarity to evaluate masking effects using the energy distribution of narrow-band temporal envelopes. Furthermore, we introduced histogram's cepstrum distance (HCD), that is derived from cepstra of histograms of the envelopes. First, the cepstra of the envelopes' magnitude histograms are calculated from targets and maskers. If these cepstra are different between the target signal and masker, the signals are not considered similar in the time-domain. Temporal envelopes are well known to informational masking cue [8]. From informational masking point of view, targets and maskers would be separated easily, if the signals are not similar in time-domain. Finally, we conducted a subjective test to confirm the relation between the HCD and informational masking. This suggested the HCD is closely related to the separation of mixed signals in the auditory system.

* Corresponding author. Tel./fax: +81 495 24 5935.
E-mail address: ahjinsei@yahoo.co.jp (S. Gotoh).

0003-682X/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apacoust.2008.08.006

2. Signal-similarity estimates using temporal envelopes

Signal similarity is closely related to the sound separation of masking effects in the auditory system. If the target signal and masker are not similar, it is easy to separate and detect both signals. Therefore, we estimated signal similarity by using temporal envelopes.

2.1. Cepstrum for envelope's histogram

Masking effects are generally estimated using the frequency domain's SNR, which is related to excitation patterns [4]. In contrast, reversal speech is well known as an example of an effective informational masker [8]. If reversal speech is a masker, the target signal has exactly the same magnitude spectrum. It will be an SNR of exactly 0 dB. It is also easy to detect the target signal. However, target separation is very difficult compared with random noise, which has the same magnitude spectrum. Furthermore, target uncertainty is high. This suggests the mechanism for understanding and separating mixed signals is not only due to the frequency domain's SNR but also very much due to time-domain information. Therefore, we have focused on temporal envelopes in this article to estimate informational masking. Narrow-band temporal envelopes are closely related to speech intelligibility [6]. In addition, we considered that the histograms of envelope dynamics are closely related to the difficulty to discriminate target signal.

Nakashima et al. proposed that the probability density function (PDF) of mixed signals can be separated by cepstrum deconvolution [7]. If target signal X was mixed with uncorrelated masker Y, the power envelope histogram as a PDF of mixed signal Z is calculated by

$$p(Z) = p(X) * p(Y), \tag{1}$$

where $p(*)$ denotes each PDF. The Fourier transform of the PDF is known as a characteristic function. Consequently, the convolution of PDF can be

$$F[p(Z)] = F[p(X)]F[p(Y)] \tag{2}$$

$F[*]$: *Fourier transform*. If the cepstrum can be calculated from the characteristic function, the PDF can be separated into target and masker characteristic functions. Thus, the PDF cepstrum $C_p(Z)$ of the target and masker are given by

$$C_p(Z) = F^{-1}[\ln F[p(X)]] + F^{-1}[\ln F[p(Y)]] \tag{3}$$

$$= C_p(X) + C_p(Y). \tag{4}$$

Then, we can obtain the PDF of target and masker features separately in the cepstrum domain. The distance between these PDF cepstra is given by

$$D_c = \sqrt{\frac{1}{N}\sum (C_p(X) - C_p(Y))^2}, \tag{5}$$

where D_c is the PDF cepstrum distance. We culculate histogram's cepstrum distance (HCD) as the PDF cepstrum distance D_c using $p(X)$ as the power envelope dynamics histogram of the target and $P(Y)$ as the masker to verify signal similarities. Consequently, if the HCD increase, the signal similarity and masking effects are considered to decrease.

2.2. Signal-similarity evaluation using HCD

We surmised that the HCD might be represented as signal similarity. Therefore, we investigated signal similarity by comparing speech signals used as targets. We prepared one speech signal as the target, which was spoken by a male and about 2-s-long. In this experiment, we used five different maskers. These maskers were random noise, which had the magnitude spectra of the target signals, music, female speech, male speech, which was spoken by a non-target speaker, and utterances by the male target speaker.

Fig. 1 shows the 1/4 octave band's power envelopes for the target and maskers at a center frequency of 500 Hz. The horizontal axis denotes the time and the vertical axis denotes the dynamics.

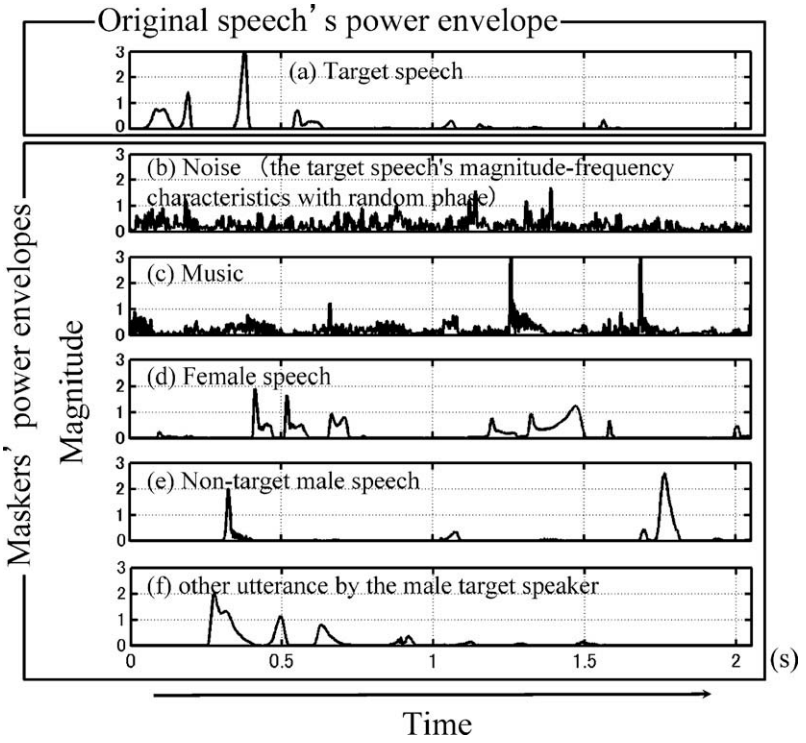

Fig. 1. Narrow-band squared envelopes of maskers (fc: 500 Hz).

Panel (a) represents the target-speech-signal's power envelope.
Panel (b) shows the power envelope of noise that has the target speech's magnitude–frequency characteristics with random phase.
Panel (c) shows the music power envelope.
Panel (d) shows the power envelope for female speech.
Panel (e) shows the power envelope for non-target male speech.
Panel (f) shows power envelope for other utterance by the male target speaker.

The power envelopes contain temporal dynamics information. This suggests it is difficult to evaluate sound similarities using unprocessed envelopes. Because shapes of stationary noise envelopes are constant. Shapes of music envelopes are different depends on tempo, rhythm and melody. And, Shapes of speech's envelopes differences are depends on contents of speech signals. For example, in Fig. 1a have different peaks compare with (f) even in speech signal which is same speaker's same sentence.

Fig. 2 shows power envelope histograms and their cepstra at a center frequency of 500 Hz.

Panel (a) represents the target-speech-signal.
Panel (b) shows the noise that has the target speech's magnitude–frequency characteristics with random phase.
Panel (c) shows the music.
Panel (d) shows the female speech.
Panel (e) shows the non-target male speech.
Panel (f) shows other utterance by the male target speaker.

The left panels are the power envelope histograms. The horizontal axis denotes the magnitude ratio that defines the power envelopes' maximum energy for 100 percent. The vertical axis denotes the envelopes' energy distributions. Although there are differences between the sound signals, these histograms still exhibit similar trends. The right panels show the cepstra of the histograms. There are differences between the sound signals, especially biggest of noise and music histogram cepstra are almost 10 times bigger than biggest of the histogram cepstra of speech signals. Furthermore, speech signals have similar trends. Therefore, we investigated HCDs for signal similarities.

Fig. 3 plots the signal similarities compared with male speech as the target signal. The horizontal axis denotes 1/4 octave band center frequencies. The vertical axis denotes HCDs. Music and noise in this figure have much greater distances than speech signals. This suggest that the HCD is reliable for estimating signal similarities.

Fig. 4 plots HCDs that were averaged over the entire 1/4 octave band distances. The results showed speechs' HCDs are around −2.4 to −2.3. On the other hand, HCD of the music and noise are over −1.4. This easily confirms signal similarities and suggests music and noise are not similar to speech target compare with speech signals. However, no speech traveled a long distance, which suggests that the various speech interfered with one another.

We discussed the HCD for estimating signal similarities. Consequently, We surmise there is the relation between signal similarities by HCD and informational masking. We next evaluated the masking effect by using the HCD.

2.3. HCD demonstration using babble speech

Babble speech is known as informational masker for speech [10] and is constructed using several speech signals. If many vocalizations are used to make babble speech, informational masking decreases[2]. This is because the masker's speech information also decreases due to multi-speech interference. Therefore, we used babble speech to investigate the HCD for signal similarities.

Fig. 5 plots the results of HCD for babble speech. The horizontal axis denotes the numbers of utterances for making babble speech. The vertical axis denotes the HCD that were averaged over the en-

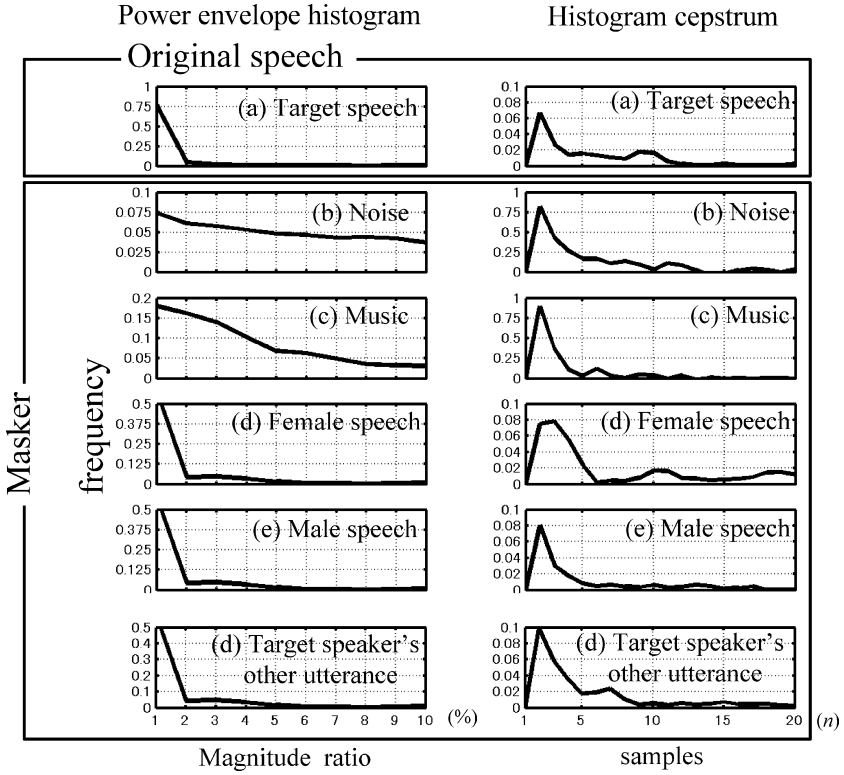

Fig. 2. Envelope power histograms and their cepstra (fc: 500 Hz).

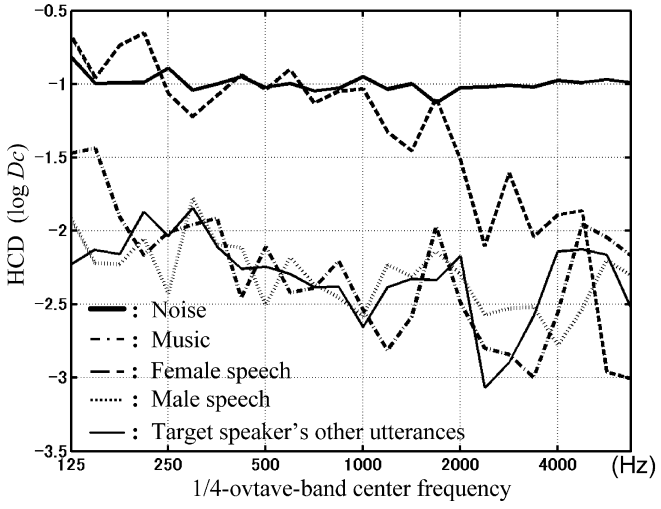

Fig. 3. HCD in 1/4-octave band.

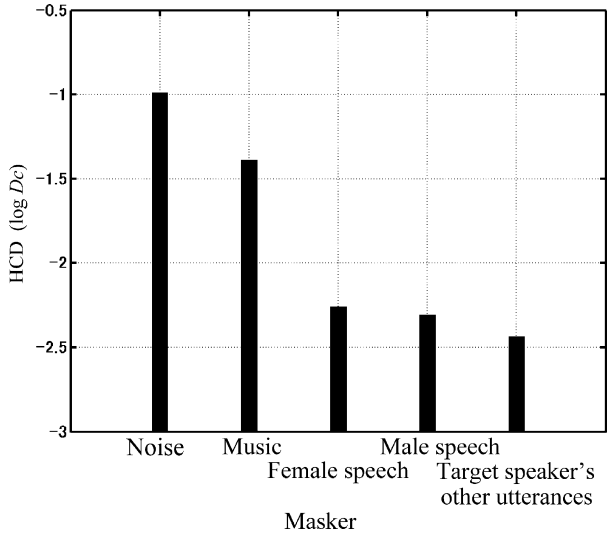

Fig. 4. HCD for unprocessed maskers.

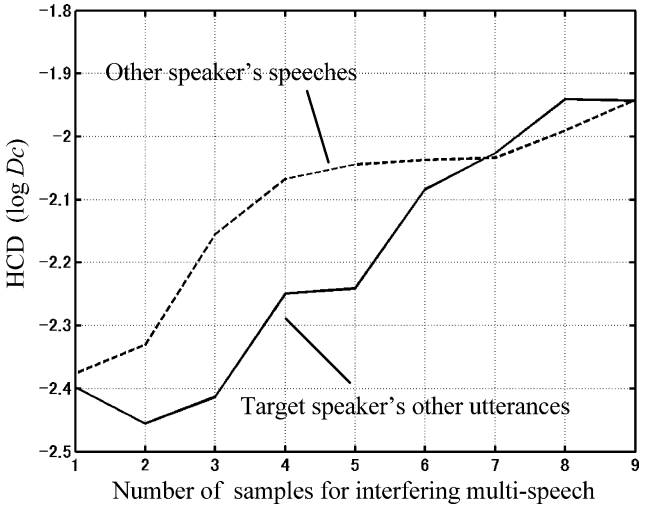

Fig. 5. HCD for competing multi-speech.

tire 1/4 octave band distances. The babble speech plotted by the bottom solid line was made by the target speaker uttering different sentences. The babble speech plotted by the top dotted line was made by the non-target speakers. We can see there is a relation between the number of utterances and increasing HCD. Also, if the sound signals are from the same speaker, the HCDs are almost all shorter than those for babble speech by different speakers. However, there is a similar trend in HCD to babble speech by different speakers when more than seven utterances are made.

We confirmed the HCD can be used to estimate signal similarity for informational masking in this section. However, the relation between HCD and informational masking in the auditory system is still unclear. We next conducted a listening test to confirm the relation between HCD and speech recognition in the auditory system.

3. Informational masking listening test using double masker

We performed an experiment to estimate informational masking to confirm the HCD by preparing a double masker that combined two different signals using the characteristics of reversal speech.

3.1. Characteristics of reversal-speech signals

Reversal speech is known to an effective informational masker [8]. If target speech is mixed with reversal speech, it would be difficult to separate and understand the target signal. We focused on these reversal speech characteristics to make an effective masker.

Where an original speech spectrum is $X(k)$, a reversal speech spectrum, $\widehat{X}(k)$, must be

$$\widehat{X}(k) = X^*(k), \tag{6}$$

where k denotes the number of frequencies. Consequently, the reversal-speech spectrum is a complex conjugate of the original-speech spectrum. Therefore, the mixed signal of original and reversal speech must be

$$X(k) + X^*(k) = 2 \times \mathrm{real}[X(k)] \tag{7}$$

The mixed signal, which only has the real part of the spectrum, is an even signal folding back at half signal length. We next prepared a double masker that had reversal-signal characteristics to confirm the relation between the HCD and speech recognition.

3.2. Preparation of double masker

We prepared a double masker to clarify the relations between HCDs and informational masking by listening tests. Fig. 6 outlines the procedure involved in making the double masker. We needed two sound sources. First, both signals A and B were divided into a spectrum of real and imaginary parts by Fourier transform. Then, the imaginary part was swapped mutually and -1 times. Therefore, the reconstructed signal (double masker) had the characteristics of both signals. We prepared four types of double maskers in these experiments. These double maskers were speech and noise, different speech pairs, music and speech, and different speech pairs including the target. The masking effect was evaluated through the listening tests. The speech samples used for stimuli used no blank linked speech or original music for the double masker.

3.3. Estimation of masking effect using adequate reply rate

Uncertainty is often used in informational masking studies to verify whether target speech has been detected and separated from mixed signals [3]. Therefore, we considered an adequate reply rate could be used to measure the uncertainty of informational masked

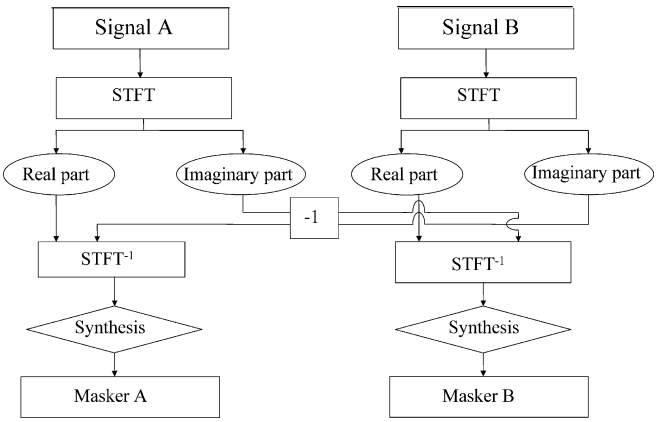

Fig. 6. Method of producing double masker.

speech. The speech intelligibility score is based on a dictation test and is generally used to estimate the masking effect. However, the conversation content in dialogue can often be understood even if speech intelligibility is low. Consequently, it is difficult to estimate informational masking effects using speech-intelligibility scores. However, an adequate reply rate can be established using a sufficient number of subject answers and comparisons do not need to be made with the task question. If target speech is uncertain due to the influence of informational maskers, subjects cannot respond adequately. Therefore, when there is a masking effect, the adequate reply rate decreases.

3.4. Results of estimating masking effect

The listening test was conducted using four types of double maskers. The double maskers components were speech with, i.e. (a) a noise masker that had a speech-magnitude spectrum, (b) a music masker, (c) a speech masker, and (d) a speech masker using the target signal's imaginary part. We recorded 120 questions as the stimuli in an anechoic room. These questions were mixed with maskers for the listening test's stimuli. All questions were easy-to-answer personal questions. Most could be answered with one word. Number of stimli is 120 samples and 10 samples for each condition. The SNRs between the target and maskers were −6, −3, and 0 dB. The subjects were seven males, aged 22–27, who were native Japanese with normal hearing. They listened to the 120 recorded samples through audio headphones (AKG-K240) choosing preferable listening levels under diotic conditions. They then wrote down answers to the stimuli except for "yes" or "no" questions.

Fig. 7 plots the listening test results. The horizontal axis denotes the double-masker components with speech. The vertical axis denotes the adequate reply rate. The solid line denotes the SNR between the target and masker of −6 dB. The dashed line denotes the SNR between the target and masker of −3 dB. The dashed–dotted line denotes the SNR between the target and masker of 0 dB. Analysis of variance (ANOVA) was performed on the results. One-way ANOVA was ran, for 4 signal types and 3 SNR conditions. Of the results, differences in "signal type" were significant ($F(3,8) = 54.31, p < 0.001$). On the other hands, differences between "SNR conditions" were not significant ($F(2,9) = 0.14, p > 0.5$). The results suggest that the signal types are more effective than SNR conditions. These results reconfirm that speech content is almost entirely uncertain, when the double masker is used on target speech. However, it is an impractical masker. This is because the target is not generally determined. Nevertheless, we can see the speech masker is more effective than the other maskers.

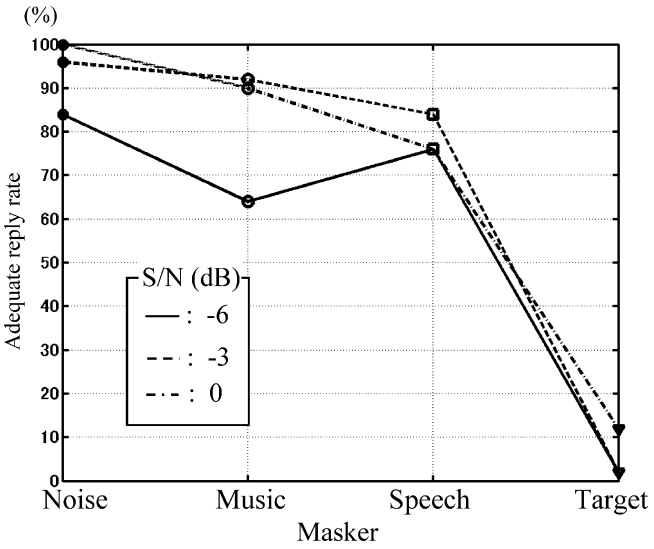

Fig. 7. Listening test results evaluated by adequate reply rate.

Fig. 8 plots signal similarity obtained by using the HCD of the double maskers. The horizontal axis denotes double-masker components with speech. The vertical axis denotes the HCD. Signal similarities of HCDs do not depend on SNR. Therefore, we averaged HCDs over entire SNR conditions. In this results, the noise masker has longest distance in four signal types. The music masker have longer distance compare with speech maskers. And, masker using target speech have smallest distance in the results. HCDs do not use SNR information to estimate signal similarity. It seems difficult to compare results with listening tests. In despite of, these HCDs distances are accord with listening test Fig. 7. Consequently, HCD has similar trend to the listening test results. These results suggest the HCD is suitable for evaluating informational masking. The results also suggest that the HCD of the center frequency of 500 Hz fits the listening test results. Noise and music compared with speech, HCD of center frequency 2 kHz is larger distance than HCD of 500 Hz. This suggests that the mean of the entire narrow-band HCD is stable. We reconfirmed the relation between informational masking and signal similarity.

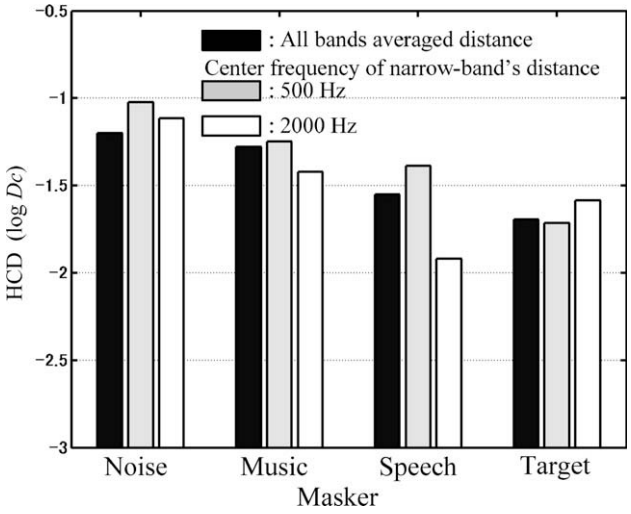

Fig. 8. HCDs of double maskers compared with target speech.

4. Conclusion

We investigated signal similarity for informational masking. If similar signals are mixed, this sound-source information is difficult to understand and separate in auditory systems. We studied this signal similarity using the cepstrum distance from a histogram of temporal envelopes. Signal similarities could be estimated using the magnitude histogram's cepstrum distance (HCD). If the HCD increase, the signal similarity and masking effects are considered to decrease. The signal similarity from the HCD was also reconfirmed using babble speech. Babbling by different speakers extended over a greater distance than babbling by the same speaker. However, when there were more than seven babbling speakers, it was the same distance. The results suggested that the HCD fit to evaluate informational masking. Because many vocalizations are used to make babble speech, informational masking decreases. We performed an experiment to estimate informational masking to confirm the HCD. We used a double masker in the listening test, which has reversal-speech characteristics. The results suggested that the double masker using target speech was very effective for informational masking. Also, the double masker created more informational masking than either noise or music. Furthermore, the HCDs also suggested that speech maskers cover less distance than either noise or music. This suggests informational masking is closely related to signal similarity.

Acknowledgement

We particularly thank Mikio Tohyama for discussions and advices concerning this work. We would like to thank Toru Taniguchifor good advices this work. This research was partially supported by Yamaha Corporation.

References

[1] Arbogast T, Mason C, Kidd Jr G. The effect of spatial separation on informational and energetic masking of speech. The Journal of the Acoustical Society of America 2002;112:2086.
[2] Drullman R, Bronkhorst A. Speech perception and talker segregation: Effects of level, pitch, and tactile support with multiple simultaneous talkers. The Journal of the Acoustical Society of America 2004;116:3090.
[3] Durlach N, Mason C, Shinn-Cunningham B, Arbogast T, Colburn H, Kidd Jr G. Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. The Journal of the Acoustical Society of America 2003;114:368.
[4] Espinoza-Varas B, Cherukuri S. 1995. Evaluating a model of auditory masking for applications in audiocoding. Applications of signal processing to audio and acoustics, IEEE ASSP Workshop, p. 195–197.
[5] Freyman R, Helfer K, McCall D, Clifton R. The role of perceived spatial separation in the unmasking of speech. The Journal of the Acoustical Society of America 1999;106:3578.
[6] Houtgast T, Steeneken H, Plomp R. Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics. Acustica 1980;46(60–72).
[7] Nakashima N, Tagaeto T, Tohyama M, Lyon R. 1996. Cepstrum deconvolution for estimating probability density functions of compound signals. In: Proceedings of Internoise, p. 2821.
[8] Rhebergen K, Versfeld N, Dreschler W. Release from informational masking by time reversal of native and non-native interfering speech. The Journal of the Acoustical Society of America 2005;118:1274.
[9] Schmitz C, Iyer N. On the reduction of masking effects while preserving competing binaural audio streams. Signals, Systems and Computers, 2003. Conference record of the thirty-seventh asilomar conference on 1. 2003.
[10] Van Engen K, Bradlow A. Sentence recognition in native-and foreign-language multi-talker background noise. The Journal of the Acoustical Society of America 2007;121:519.
[11] Wang C, Bradley J. Prediction of the speech intelligibility index behind a single screen in an open-plan office. Applied Acoustics 2002;63(8):867–83.