J. Acoust. Soc. Jpn. (*E*) **21**, 5 (2000)

PAPER

The auditory-oriented spectral distortion for evaluating speech signals distorted by additive noises

Mitsunori Mizumachi* and Masato Akagi**

* *ATR Spoken Language Translation Research Laboratories,*
2-2-2, Hikaridai, Seika-chou, Souraku-gun, Kyoto, 619-0288 Japan
** *School of Information Science, Japan Advanced Institute of Science and Technology (JAIST),*
1-1, Asahidai, Tatsunokuchi, Nomi-gun, Ishikawa, 923-1292 Japan

(*Received 11 November 1999*)

This paper proposes an objective speech distortion measure as a substitute for human auditory systems. Simultaneous and temporal masking effects are introduced into this measure called auditory-oriented Spectral Distortion (ASD). We calculate the ASD using spectral components over masked thresholds in the same way as the Spectral Distortion (SD). We confirmed that the ASD is more compatible to subjective mean opinion score that represents distortions on auditory impression than the SD. We applied the ASD to optimize a noise reduction algorithm proposed by the authors, and confirmed that this optimized algorithm reduces noises appearing to the ear. ASD is sure to be an available guide to design noise reduction algorithms.

Keywords: Objective distortion measure, Additive noises, Masking effects, Noise reduction, Spectral subtraction

PACS number: 43. 50. Hg, 43. 66. Dc, 43. 71. Gv

1. INTRODUCTION

The demand of noise reduction technique is increasing as speech processing techniques become increasingly practical, especially with regard to noise reduction for automatic speech recognizers (ASRs). Although almost all recent ASRs work perfectly in ideal environments, their performances tend to decline drastically in noisy environments.[1] There is also a demand to decrease distortions on auditory impressions for hearing aids, cellular phones and so on. It is very difficult for many noise reduction methods to reduce non-stationary noises in daily life. A guide that estimates the quantity of distortion in daily life is necessary to design noise reduction algorithms. According to Kahrs and Brandenburg[2] there are some objective distortion measures introduced auditory masking phenomena. They are almost divided into telephone-band coded speech measures (300–3,400 Hz) or wideband high quality audio measures (20–20,000 Hz). Since we need an objective distortion measure for hearing aids whose available frequency ranges are generally from 100 Hz to 6,000 Hz, the telephone-band coded speech measure is not sufficient and the high quality audio measure is too strict. Furthermore, we should evaluate a great variety of distortions caused by various noises unlike evaluation of codec speech signal. Aim of this distortion measure is not to predict the subjective evaluation itself as a ratio scale, but to evaluate the abilities of noise reduction algorithms. It is enough for this measure to be an interval scale.

This paper proposes an objective distortion measure, called the auditory-oriented Spectral Distortion (ASD),[3] that takes account of both simultaneous and temporal masking effects in the same way as the above measures. ASD thinks over general ideas in our auditory systems such as auditory filters more exactly than other distortion measures. We conduct some listening tests to optimize the parameter of the ASD and to examine effectiveness of the ASD. Experimental results show that the optimized ASD is more compatible to subjective evaluation than the Spectral Distortion (SD).

The authors have proposed a noise reduction method, and confirmed its effectiveness as a front-end of ASRs.[4,5] In this paper, we try to apply this method to various acoustic equipment that requires reducing noises appearing to the ear such as hearing aids. We also conduct some experiments to optimize and evaluate this noise reduction algorithm using two objective distortion measures, the LPC-SED[4] for ASRs and the ASD for hearing aids.

In Section 2, we describe the implementation of an objective distortion measure (that is the ASD) based on the human auditory characteristics. In Section 3, we describe the verification of the ASD. Then in Section 4, we apply the ASD to optimize noise reduction algorithm. Finally, conclusions are drawn in section 5.

2. OBJECTIVE DISTORTION MEASURE

2.1 System Overview

The auditory-oriented Spectral Distortion (ASD) is constructed by introducing auditory masking effects into the Spectral Distortion (SD). Characteristics of both simultaneous and temporal masking phenomena are included in the ASD. Although they actually depend on sound pressure levels, fixed characteristics are introduced into the ASD to simplify its calculation. Figure 1 illustrates a process of calculating a masking threshold, as an A-weighted logarithm amplitude spectrum by a thin solid line, detected masker components by circles, masking patterns for each masker by dotted lines, a masking threshold for temporal masking by a broken line, and an integrated masking threshold by a thick solid line. We describe how to calculate them below.

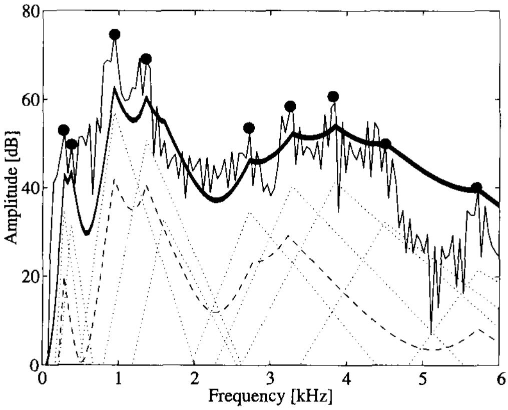

Fig. 1 An illustration of calculating a masking threshold. An A-weighted amplitude spectrum is shown by a thin solid line, detected masker components by circles, masking patterns for each masker by dotted lines, a masking threshold for temporal masking by a broken line, and an integrated masking threshold by a thick solid line.

The ASD is calculated using spectral components over the masked thresholds in each short term frame whose length is 21.3 ms and period is 5.3 ms. In this paper, every signal is sampled by 48 kHz with 16 bit accuracy.

2.2 Implementation of Simultaneous Masking Effect

Masker candidates are detected from a target signal, so that the masking pattern for each masker can be calculated. A target signal $x(t)$ passes through the A-characteristic filter adopted in sound level meters. An A-weighted logarithm amplitude spectrum $x(\omega)$ must be like the human perception in comparison with an amplitude spectrum itself. When a spectral component $X(k)$ satisfies all the conditions in Eq. (1), it is detected as one of masker candidates.

$$\begin{cases} X(k)>X(k-1), \\ X(k)\geq X(k+1), \\ X(k)-X(k+j)>3\ [\text{dB}],\ j=2,3,\ \dots,\ J, \end{cases} \qquad (1)$$

where k and j are discrete frequencies. The search range J is an equivalent rectangular bandwidth (ERB) which equals to the critical band of an auditory filter. The ERB is defined as follows:

$$\mathrm{ERB}(f)=24.7(4.37\cdot f/1{,}000+1)\quad [\text{Hz}], \qquad (2)$$

that is the bandwidth of an auditory filter whose center frequency is f Hz.[6] We reduce the number of masker candidates detected in Eq. (1) on the assumption that there is a masker at the most in each auditory filter. Only the largest masker in the logarithm amplitude scale (circles in Fig. 1) is survived in each auditory filter.

A masking pattern for each masker candidate is calculated as an approximation of the masking pattern for narrow-band noises measured by Egan and Hake.[7] For practical purposes, the masking pattern is approximated by the three following points:

$$\begin{cases} \mathrm{A}: (k, X(k)-18), \\ \mathrm{B}: (k-2\cdot\mathrm{ERB}(k), X(k)-48), \\ \mathrm{C}: (k+4.5\cdot\mathrm{ERB}(k), X(k)-48), \end{cases} \qquad (3)$$

where $X(k)$ means the sound pressure level in dB of a masker candidate in k Hz, and $\mathrm{ERB}(k)$ is calculated by Eq. (2). Each masking pattern (each triangle by dotted lines in Fig. 1) is calculated as the blacken region in Fig. 2.

2.3 Implementation of Temporal Masking Effect

The temporal masking effects are implemented by attenuating masking thresholds calculated in the past frames. For example, a broken line in Fig. 1 shows an attenuated masking threshold of the past frame. We adopt a post-masking curve measured by Zwicker.[8] Figure 3 shows a post-masking curve that determines the elapsed attenuation level in dB for each masking threshold, for example a thick solid line in Fig. 1. The effect of backward masking is disregarded since it is not as influential as the effect of forward masking.[9]

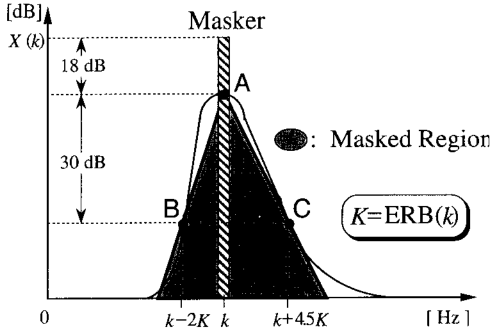

Fig. 2 A masking pattern as an approximation of the masking pattern for narrow-band noises measured by Egan and Hake.

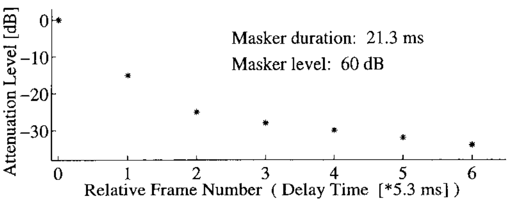

Fig. 3 A post-masking curve for burst tones measured by Zwicker.

2.4 Calculation of the ASD

We can obtain the current masking threshold (a thick solid line in Fig. 1) by integrating all masking patterns in both simultaneous masking (dotted lines in Fig. 1) and temporal masking (a broken line in Fig. 1) using the Power-law model presented by Lutfi.[10] Assuming that $Y_{\omega_i}(\omega)$ is a masking pattern for a single masker in simultaneous masking whose frequency is ω_i, and $Y_t(\omega)$ is an attenuated masking threshold in temporal masking where t is a frame number related to delay time, the total masking threshold $Y_{\mathrm{total}}(\omega)$ is calculated as:

$$Y_{\mathrm{total}}(\omega)=F^{-1}\left[\sum_{\omega i}F[Y_{\omega_i}(\omega)]+\sum_{t=1}^{6}F[Y_t(\omega)]\right],$$
$$F(z)=z^p, \qquad (4)$$

where p is a constant, and $Y_{\omega_i}(\omega)$, $Y_t(\omega)$, and $Y_{\mathrm{total}}(\omega)$ are not logarithmic.

Lutfi has originally confirmed that $p=0.33$ is the most desirable as compared with results of perceptual experiments when he used only a few narrow-band noises as masker signals.[10] We should reconsider the most suitable value of parameter p, since we use more complicated speech signals as masker signals. We discuss this problem in Section 3.

Finally, assuming that we cannot perceive distortions if their spectral levels are under the masking thresholds $Y_{\mathrm{total}}(\omega)$, the ASD is calculated as follows.

$$\mathrm{ASD}=\sqrt{\underset{i}{\mathrm{MEAN}}\{S_{\mathrm{target}}(i)-S_{\mathrm{clean}}(i)\}^2}\ [\mathrm{dB}], \qquad (5)$$

where MEAN calculates the arithmetical mean, $S_{\mathrm{target}}(i)$ and $S_{\mathrm{clean}}(i)$ are logarithm amplitude spectra of a target speech signal and a clean speech signal, and i indicates the discrete frequency that satisfies

$$S_{\mathrm{target}}(i)>20\log_{10}Y_{\mathrm{total}}(i),$$
$$100\,[\mathrm{Hz}]\leq i\leq 6{,}000\,[\mathrm{Hz}]. \qquad (6)$$

ASD equals to the SD when the discrete frequency i in Eq. (5) comes under every discrete frequency in the range that is from 100 Hz to 6,000 Hz.

3. VERIFICATION OF THE ASD

3.1 Aim

We examine the verification of the ASD. A criterion of verification is the linearity between subjective evaluation and objective evaluation. We can easily estimate how much a speech signal is distorted when the linearity is well preserved. The relationship between the mean opinion score (MOS) as subjective evaluation and the ASD or the SD as objective evaluation is examined.

3.2 Procedure

A continuous Japanese vowel /ao/ in the ATR speech database[11] is prepared and distorted by adding a random noise (from 2,000 Hz to 3,000 Hz). A clean speech signal and a noisy speech signal are drawn in Fig. 4. Figure 4 also shows objective distortions of the noisy speech signal in SD or ASD (p=0.6). The best suited value of parameter p in the ASD is 0.6, and it is explained in Section 3.3. Mean distortion was 16.7 dB in SD, 21.5 dB in ASD (p=0.6), and −2.3 dB in SNR. It was obviously evaluated as very noisy (MOS was distinctly 0 on the five point scale: 4[clean]-0[noisy]) in pre-listening tests, so we determined that this signal was a noise maximum speech signal. Test signals are made by adjusting at even intervals on each objective distortion measure. For all practical purposes, six noisy speech signals whose noise levels are 4.2 dB, 8.4 dB, 12.6 dB in SD, and 5.4 dB, 10.8 dB, 16.1 dB in ASD (p=0.6), are prepared as shown in Fig. 5.

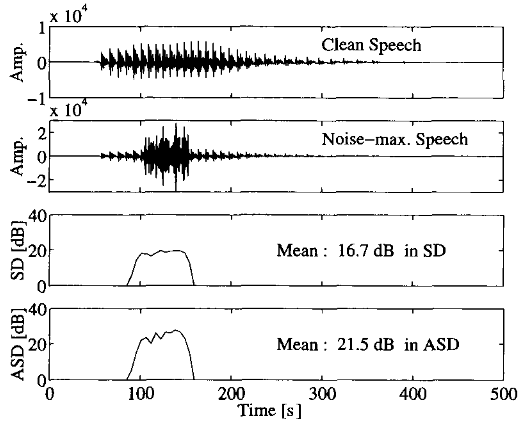

Fig. 4 A clean speech signal, the most noisy speech signal, and distortions in SD and ASD (p=0.6).

Each signal is randomly presented four times to each subject through the headphone (STAX Lambda Nova Signature). Subjects are eight postgraduate students who have no hearing impairments. The sound pressure levels of a clean speech signal and a noise maximum speech signal are 66 dB (A) and 75 dB (A). Subjects listen to a clean speech, a noise maximum speech, and a speech to evaluate in this order, or a noise maximum speech, a clean speech, and a speech to evaluate in this order, and give the MOS on the five point scale toward the third speech signal.

3.3 Experimental Results

Figure 6 shows the MOSs for all stimuli on the

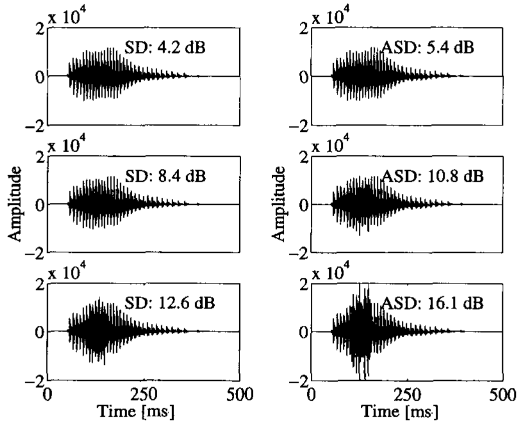

Fig. 5 Noisy speech signals prepared for the verification of objective distortion measures.

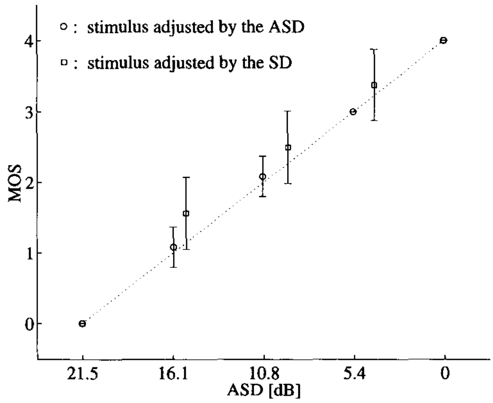

Fig. 6 Relationship between the ASD and the MOS for all stimuli.

ASD ($p=0.6$) scale, and Fig. 7 shows them on the SD scale. In both Fig. 6 and Fig. 7, circles mean stimuli adjusted at even intervals on the ASD scale and squares mean them on the SD scale, and error bars mean standard deviations of the MOSs that eight subjects gave.

The standard deviations of the MOSs are small for stimuli (plotted by circles) adjusted at even intervals in the ASD scale. On the other hand, they are comparatively large for stimuli (plotted by squares) adjusted at even intervals in the SD scale. Then, stimuli are sparsely arranged on the ASD scale, but they are closely arranged on the SD scale. In Fig. 7, two stimuli evaluated differently in subjective listening tests are almost the same on the SD scale, for example two stimuli when the SD is 12.6 dB are given distinct different MOSs.

We should also examine the best suited value of parameter p in the ASD. The relationship between the ASD and the MOS appears to be an interrelated curve. If the ASD under a certain condition is well suited to subjective MOS, the interrelated curve becomes a straight line, which is called the best desirable line (BDL). An index to evaluate the distortion measure is the correlation coefficient between an interrelated curve and the BDL. Here, the parameter p in the ASD varies from 0.3 to 0.9 in 0.1 steps. Figure 8 shows the correlation coefficients between the interrelated curve for the ASD in each p and the BDL. We found that the best suited value of parameter p was 0.6 (correlation coefficient was 0.992), whereas the correlation coefficient for the SD was 0.951.

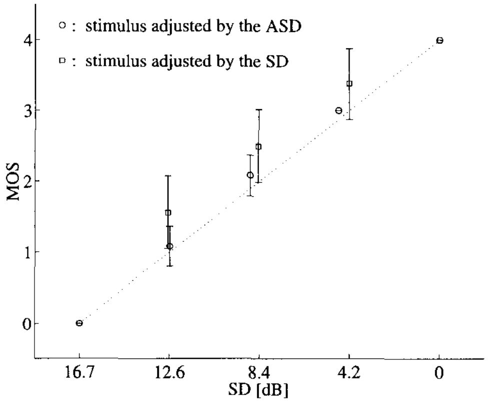

Fig. 7 Relationship between the SD and the MOS for all stimuli.

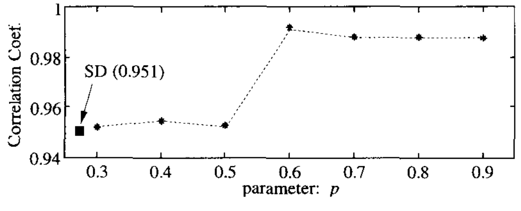

Fig. 8 Correlation coefficients between the interrelated curve for the ASD on each p and the BDL.

3.4 Discussion

In Fig. 6, ASD is appropriate as an interval scale because the MOS increases in proportion as the ASD decreases as shown. In Fig. 7, although two stimuli are evaluated equally on the SD, their MOSs as subjective evaluations are different. This suggests that SD does not reflect the subjective evaluation.

It is desirable that the standard deviation of the MOS is small. If a distortion measure is compatible to subjective evaluation, MOS shuold be an integer in this experimental condition. A large standard deviation means that subjects take great pain to evaluate a stimulus on the MOS that must be an integer. Therefore, ASD is superior to the SD as regards the correspondence to subjective evaluation.

On optimizing the parameter p, this result supports the assertion by Lutfi *et al.* that p should be larger than 0.33 if the number of maskers increases.[12,13]

4. APPLICATION TO OPTIMIZE NOISE REDUCTION ALGORITHM

4.1 Outline of the Noise Reduction Algorithm

The authors have proposed a noise reduction method,[4,5] consisting of three modules: estimation of signal directions, estimation of the noise spectrum, and subtraction of noise spectrum using the Spectral Subtraction.[14] In the third module, we can obtain an amplitude spectrum $\hat{S}_{\mathrm{speech}}(\omega)$ of the noise reduced speech signal using $\hat{S}_{\mathrm{noise}}(\omega)$ that is an amplitude spectrum of the estimated noise signal and $\hat{S}_{\mathrm{received}}(\omega)$ that is an amplitude spectrum of the received noisy signal as follows.

$$\hat{S}_{\mathrm{speech}}(\omega)=\begin{cases} S_{\mathrm{received}}(\omega)-\alpha\cdot\hat{S}_{\mathrm{noise}}(\omega), \\ \qquad\qquad S_{\mathrm{received}}(\omega)\geq\alpha\cdot\hat{S}_{\mathrm{noise}}(\omega), \\ \beta\cdot S_{\mathrm{received}}(\omega), \quad \text{otherwise,} \end{cases} \qquad (7)$$

where α is the subtraction coefficient, and β is the flooring coefficient that is fixed as 0.001 in this paper.

4.2 How to Optimize the Noise Reduction Algorithm

The parameter of the noise reduction algorithm is optimized to bring out the greatest ability in each purpose. One of the most dominant parameter of this method is the subtraction coefficient α in Eq. (7). We adopt two objective distortion measures as criteria to optimize it. One is the ASD supposing hearing aids, and the other is the LPC log Spectrum Envelope Distortion (LPC-SED), that is a distortion measure for a front-end of ASRs[4] and calculates the distortion in the range that is from 100 Hz to 6,000 Hz. We already confirmed that the LPC-SED was compatible for speech recognition rates when $\alpha=1.0$.[5] However, we did not confirm whether $\alpha=1.0$ is valid.

The test signals are the same as those signals found in Section 3.2, and coordinated to set the SNR as 10 dB. Noises in these signals are reduced by proposed noise reduction method when the subtraction coefficient α varies from 0.0 to 2.0 in 0.1 steps. Note that we do not execute the process of noise reduction when $\alpha=0$.

4.3 Experimental Results

In Fig. 9, we can see the relationships between the parameter α and distortions after noise reduction on the LPC-SED and the ASD ($p=0.6$). It is obvious that this method can reduce distortions caused by non-stationary noises on both distortion measures. The experimental results show that the suitable value of α should be obviously different in each purpose.

It is confirmed that former experiment results[4,5] were valid, because Fig. 9 (upper panel) shows that the former parameter setting ($\alpha=1.0$) is the most suitable. In order to improve the auditory impressions, it is desirable to set α more than 1.0. To examine this tendency in detail and verify the ASD again, we conducted additional listening tests for conversational speech signals extracted from concurrent conversational signals. The conversational signals are from the ATR speech database,[11] and a target speech is "*tsuuyakudenwa kokusaikaigi jimukyoku desu*" and a disturbance speech is "*daimoku no shimekiri wo oshietekudasai.*" Eight subjects

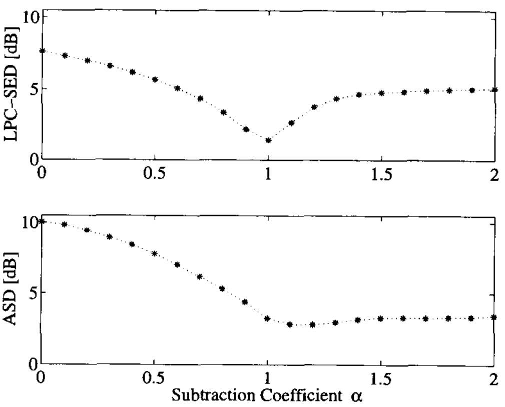

Fig. 9 Distortions on the LPC-SED (upper panel) and the ASD (lower panel) of noise-reduced speech signals for each subtraction coefficient α.

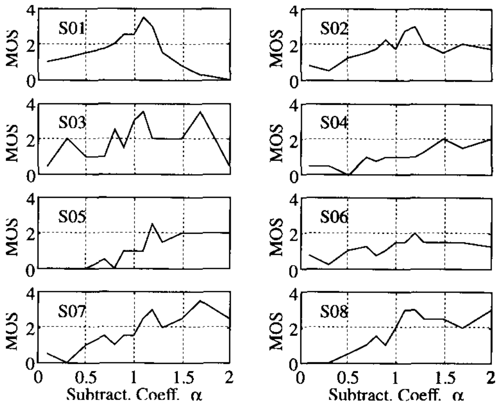

Fig. 10 The mean MOS for each subject.

randomly listened to each extracted speech signal four times, and gave the MOS on the five point scale.

The experimental result for each subject is individually shown in Fig. 10. Each panel in Fig. 10 shows the mean MOS for each extracted target speech signal, and has the maximum value or a local peak of the MOS in the range $1.1\leq\alpha\leq1.5$. A primary factor for scattering the MOS ($0.1\leq\alpha\leq0.3$) may be the separation of a disturbance speech in listening to concurrent conversational signals. Figure 11 shows mean MOSs for all subjects, and has the maximum MOS in $\alpha=1.2$. There is no significant difference between $\alpha=1.0$ and $\alpha=1.2$ in a statistical test, since the dynamic range of the MOS is different for each subject. However, almost all

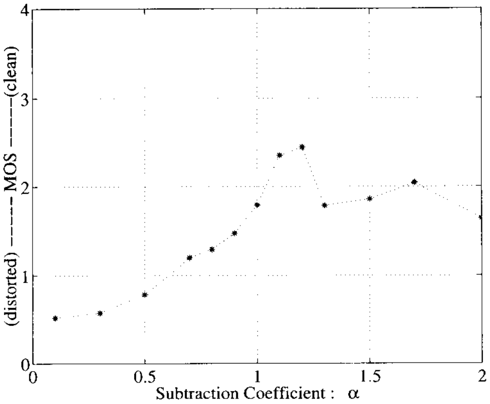

Fig. 11 The mean MOS for all subjects.

subjects suggest that it is desirable that a is larger than 1.0 as shown in Fig. 10.

4.4 Discussions

It is confirmed that the suitable value of the subtraction coefficient a is different in each application and the ASD works well.

When a parameter a is larger than 1.0, the noise reduced speech signal barely includes noise, however, it is shaved to some extent. A human auditory system may be sensitive to noises and may be insensitive to the lack of spectra in the speech signals. This over-subtraction equally shaves spectral peaks and spectral dips, but spectral dips are inaudible owing to masking effects. In other words, spectral peaks are dominant and spectral dips are not influential for speech perception. These knowledges prove that the ASD fits the human auditory perception more accurately than the SD again, since the ASD is an objective distortion measure considering that spectral peaks are important and the SD equally evaluates distortions in all frequencies.

5. CONCLUSIONS

This paper proposed an objective distortion measure, called ASD, that takes account of simultaneous and temporal masking effects. ASD is more compatible to subjective evaluation (MOS) than the SD. The experimental results suggest that the suitable parameter for the noise reduction algorithm is different in each application such as a front-end of ASRs or hearing aids. Unlike the commonplace setting such as a front-end of ASRs, over-subtraction of noise spectra should be desirable to decrease distortions on auditory impressions.

ACKNOWLEDGMENTS

This work was partly supported by the CREST project under the sponsorship of Japan Science and Technology Corporation, and by a research grant from the Ministry of Education, Science and Culture of Japan (No. 10680374).

REFERENCES

1) J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition* (Kluwer Academic Publishers, Boston, 1996).
2) M. Kahrs and K. Brandenburg, *Applications of Digital Signal Processing to Audio and Acoustics* (Kluwer Academic Publishers, Massachusetts, 1998).
3) M. Mizumachi and M. Akagi, "An objective distortion estimator for hearing aids and its application to noise reduction," Proc. EUROSPEECH '99, 2619–2622 (1999).
4) M. Mizumachi and M. Akagi, "Noise reduction by paired-microphones using spectral subtraction," Proc. ICASSP '98, 1001–1004 (1998).
5) M. Mizumachi and M. Akagi, "Noise reduction method that is equipped for a robust direction finder in adverse environments," Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, 179–182 (1999).
6) B. R. Glasberg and B. J. C. Moore, "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138 (1990).
7) J. P. Egan and H. W. Hake, "On the masking pattern of a simple auditory stimulus," J. Acoust. Soc. Am. **22**, 622–630 (1950).
8) E. Zwicker, "Dependence of post-masking on masker duration and its relation to temporal effects in loudness," J. Acoust. Soc. Am. **75**, 219–223 (1984).
9) E. Zwicker, *Psychoakustik* (Springer-Verlag, Berlin, 1982) [Japanese edition: Y. Yamada (Nishimura, Niigata, 1992)].
10) R. A. Lutfi, "Additivity of simultaneous masking," J. Acoust. Soc. Am. **73**, 262–267 (1983).
11) K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe, and H. Kuwabara, "Speech database user's manual," ATR Tech. Rep. TR-I-0028 (1988) (in Japanese).
12) R. A. Lutfi, "A power-law transformation predicting masking by sounds with complex spectra," J. Acoust. Soc. Am. **77**, 2128–2136 (1985).
13) L. E. Humes and W. Jesteadt, "Models of the additivity of masking," J. Acoust. Soc. Am. **77**, 1285–1294 (1989).
14) S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process. **ASSP-27**, 113–120 (1979).

Mitsunori Mizumachi was born in Fukuoka, Japan on September 20, 1972. He received a B.E. degree from Kyushu Institute of Design in 1995, and M. E. and Ph.D. degrees from Japan Advanced Institute of Science and technology (JAIST) in 1997 and 2000, respectively. He is currently an invited researcher of the ATR Spoken Language Translation Research Laboratories. His research interests include signal processing of speech and objective evaluation of speech quality. He is a member of the IEICE and ASJ.

Masato Akagi was born in Okayama, Japan on September 12, 1956. He received a B. E. degree from Nagoya Institute of Technology in 1979, and M. E. and D. E. degrees from Tokyo Institute of Technology in 1981 and 1984, respectively. In 1984, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT). From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been with School of Information Science, JAIST, and he is currently a professor of JAIST. His research interests include speech perception, modeling of speech perception mechanisms of humans, and signal processing of speech. During 1988, he joined the Research Laboratories of Electronics, MIT as a visting researcher, and in 1993, he studied at the Institute of Phonetic Science, Univ. of Amsterdam. He received the Sato paper Award from the ASJ in 1998. He is a member of the IEICE, ASJ, ASA and ESCA.