Article
Chinese Science Bulletin
Acoustics
July 2012 Vol.57 No.20: 2548–2553
doi: 10.1007/s11434-012-5273-1

Chinese speech identification in multi-talker babble with diotic and dichotic listening

PENG JianXin[1], ZHANG HongHu[2*] & WANG ZiYou[1]

[1] *Department of Physics, School of Sciences, South China University of Technology, Guangzhou 510640, China;*
[2] *Department of Architecture, Zhejiang University, Hangzhou 310058, China*

Received February 21, 2012; accepted May 2, 2012

To explore Chinese Mandarin speech identification in babble of spatially separated talkers, subjective speech identification tests of word and sentence were made with diotic and dichotic listening respectively. The result shows that the speech identification scores changed non-monotonically with the masker number *N* increasing from 1 to infinity, first declining gradually until reaching their minimums and then rising. Statistical difference was found between the scores of diotic and dichotic listening. For all the values of *N* checked, dichotic listening achieved higher scores than diotic listening, showing that dichotic effect has an advantage for reducing babble masking. And the scores of sentence test are significantly higher than that of word test with whether diotic or dichotic listening, indicating that the linguistic connection in sentence can help listeners get a better perception of the target speech in babble masking.

Chinese Mandarin, speech identification, multi-talker babble, diotic listening, dichotic listening

Citation: Peng J X, Zhang H H, Wang Z Y. Chinese speech identification in multi-talker babble with diotic and dichotic listening. Chin Sci Bull, 2012, 57: 2548–2553, doi: 10.1007/s11434-012-5273-1

In everyday listening conditions, our ears often receive a mixture from multiple concurrent sound sources. As an important case, listening to a speech of interest in presence of competing speech babble (the summation of the simultaneous speech signals other than the interested) has been the subject of considerable studies over years [1–5]. Agus et al. [6] recognized that the competing babble might mask the target speech significantly. The relationship between the perception of the target speech and the competing babble was found quite complex. Carhart et al. [7] reported that the masking was strongly related to the number (*N*) of simultaneous masking talkers, which first grew as *N* increasing from 1 to 3 and then decreased until it became stable as *N* exceeding 64. Brungart et al. [8] also found that *N*=2 or 3 might produce considerably more masking than *N*=1 at low signal-to-noise ratios. Yost et al. [2] observed that, for a divided attention task, *N*=3 created much more difficulty than *N*=2. Simpson and Cooke [9] measured consonant identification rate with diotic listening, using a closed set of vowel-consonant-vowel speech tokens gated by multi-talker babble. In their experiment, babble noises of $N = \{1, 2, 3, 4, 6, 8, 16, 32, 64, 128, 512, \infty\}$ were employed, where $N=\infty$ representing a speech-shaped steady noise. The identification rate was found non-monotonic, which first decreased gradually as *N* increasing from 1 and reached its minimum at *N*=8 and then kept almost stable between *N*=8 and 128 until a recovery to the level of speech shaped noise at *N*=512. Hoen et al. [10] further investigated with diotic listening the differential effects of acoustic-phonetic and lexical contents on target word identification between natural and time-reversed speech babbles of *N*=4, and the results showed that the identification rate was poorer in the natural babbles than in the reversed ones.

Moreover, the perception of the target speech would get improved in conditions that the target and the competing speeches coming from different positions than that all the speeches coming from a same position [11]. Arbogast et al. [4] found a large release from masking caused by spatially

*Corresponding author (email: zhanghh@zju.edu.cn)

separating the target talker and competing maskers, which implied that the listener was able to take the binaural advantage to achieve substantially better perception. Yost et al. [2] argued that spatial cues were particularly helpful in resolving a condition of three talkers, compared to the conditions of two talkers, and suggested that binaural processing may be more important for solving the "cocktail party" problem with over two concurrent maskers there.

In comparison to the anechoic environment, room reverberation was found by some researchers to be able to disrupt listeners' ability to distinguish spatial locations of competing voices [12,13]. The amounts of masking and spatial release may depend on the characteristics of the room and the masking both. When the masking was primarily energetic, spatial release decreased from a maximum of about 8 dB in the least reverberant rooms to about 2 dB in the most reverberant rooms. For the informational masking, a larger release of 15–17 dB was observed without being affected by reverberation [14].

By now, the researches on the speech identification in *N*-talker babble have been almost all made for western languages. In this work, Chinese Mandarin is the objective language. Subjective tests of word and sentence were carried out by use of acoustical simulation and auralization techniques with software Odeon [15,16]. Monaural and binaural room impulse responses (MRIRs and BRIRs) between the spatially separated virtual talkers and the listener in a virtual room of reverberation time about 1.0 s in middle frequencies were simulated. The acoustical stimuli were convolved with the MRIRs and BRIRs and then reproduced through headphone to generate target and competing speeches for diotic and dichotic listening in the tests respectively. Conditions of 9 masker numbers of $N = \{1, 2, 4, 6, 8, 12, 16, 32, \infty\}$ were considered.

1 Experiment

1.1 Simulation of room impulse responses

The symmetrical plane of the virtual room and 8 masking conditions with finite masker number $N = \{1, 2, 4, 6, 8, 12, 16, 32\}$, are illustrated in Figure 1. The room is 25 m long and 6 m high, with its front and back walls being 18 m and 23 m wide respectively. All the sources and the receiver are located on a square grid of size 1 m×1 m, 1.2 m high above the floor. The target source and the receiver are on the central axis of the room. All the sources are omni-directional with unit sound power.

The impulse responses from the target and each masking source to the receiver, including the MRIRs and BRIRs, were simulated respectively with the receiver facing to the target source. The BRIRs were calculated by ODEON using the default Head Related Transfer Function adopted in the software. By properly arranging sound absorption of room surfaces, the reverberation time (RT) of the virtual room was controlled to be around 1.0 s in 500–1000 Hz octave bands as listed in Table 1. The values of RT in Table 1 were calculated from the MRIR between the receiver and the target source. In every band, the difference between the value in Table 1 and that derived from the MRIR between the receiver and any another masking source was trivial.

1.2 Listening materials

The phonetically balanced word lists as specified by GB 15508–1995 [17] were used to generate target speeches in the word test. Every list consists of 25 rows. Each row as a carrying phrase contains 3 single-syllable target words, reading as "The – row is ×××". The "–" stands for row number and "×××" for the 3 target words. The 3 syllables in

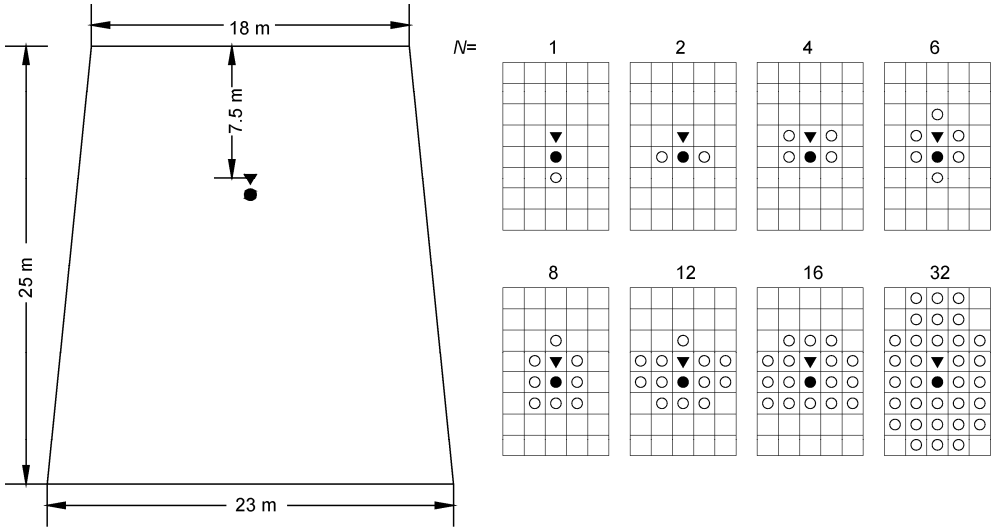

Figure 1 Plane sketch of virtual room and configurations of 8 masking conditions of finite masker number *N*. ●, target source; ▼, receiver; ○, masking sources.

Table 1 Reverberation time of virtual room in the octave bands from 125–4000 Hz

Oct-band (Hz)	RT (s)
125	1.30
250	1.21
500	1.03
1000	1.09
2000	0.98
4000	0.75

each row are randomly arranged and nonsense. And the total 75 syllables in each list keep the same balance of difficulty level and phonemic characteristic without repetition. The sentence lists as specified by the standard Chinese Mandarin Hearing in Noise Test (MHINT) [18] were used for target speeches in the sentence test. Every list contains 10 sentences and each sentence is composed of 10 target words; for example, "这个球队终于打入决赛(This team enters the finals at last)". Some Mandarin speech materials of sentences other than that for the target speeches, were used for producing the babble speeches [19]. All dry signals of the target and masking speeches were recorded in anechoic chamber with microphone of instrumental grade at sampling rate of 44.1 kHz, using native talkers of Chinese Mandarin. Two talkers, a male and a female, were used to record the target speeches respectively and some other male and female talkers were used to record the masking speeches. The speeches uttered by each masker at a different position in the virtual room were recorded by a different talker in anechoic chamber. The contents of speeches used in different masking conditions were different from one another, helping avoid influence of memory of listeners. For the masking condition of N=1, the masker corresponding to the male or the female target talker was of the same gender as the target talker. For the masking conditions of N = {2, 4, 6, 8, 12, 16, 32}, the gender ratio of maskers was fixed as 50% male to 50% female as was adopted in the research of Hoen et al. [10], and the two genders were almost symmetrically distributed around the receiver's position with the central axis of the room as the symmetric axis. For the masking condition of N=∞, a speech-shaped noise was used as the masking speech signal and the masking source was the same one as in the condition of N=1.

For every masking condition, the dry signals of target and masking speeches were convolved with the simulated MRIRS and BRIRs respectively after headphone equalization. For every different masking source, a different masking signal was used. A level adjustment was applied for every convolved signal according the sound pressure level caused by the source of the signal at the receiver. The level estimation was based on the overall A-weighted RMS value and corrected for the effect of silent periods by the application of a threshold [20,21]. First, the convolved masking signals were mixed together to generate the babble signal. Then, the babble signal was mixed with the convolved signal of the target at SNR of 0 dBA.

1.3 Listening procedure

The subjective listening tests were carried out in a quiet experimental room, where the level of background noise was lower than 35 dBA. Total 16 listeners, 10 male and 6 female, attended the tests. They were chosen from undergraduate students aged from 20 to 24, all native speakers of Chinese Mandarin, and without known hearing problems.

The 9 different masker numbers of N = {1, 2, 4, 6, 8, 12, 16, 32, ∞}, 2 listening ways (diotic and dichotic), and 2 types of listening content (word and sentence), all resulted in 9×2×2=36 test conditions. The experiments were made 9 times. Each time, 4 different test conditions were dealt with, by a group of 4 listeners selected from the total 16. The experiment of each test condition took about 10 min. After the experiments of 2 test conditions, the listeners took a break of about 20 min, and then attended the experiments of the rest 2 test conditions.

In the experiments, the mixtures of babble and target speeches were amplified by Edirol UA-25 USB Sound Card and Symetrix 304 Headphone Amplifier, and then reproduced through headphone (Sennheiser HD580) at about 70 dBA. For each test condition, target speeches of 2 different lists were listened to, one for the male target talker and the other for the female. The listeners were asked to write down the words which pronounced as the target words that they had heard in the carrying phrases of word test or in the sentences of sentence test. The speech identification score of the test condition was calculated as the average percentage of the correct identification of the group, namely, the mean value of the 4 (listeners)×2 (lists)=8 scores of the group.

Each listener participated in the experiments 2–3 times. In order to prevent the effect of memory, the tests were carefully arranged to guarantee that, there was an interval of at least a week between any 2 successive times of the experiments that any listener participated in, and every listener did not meet identical target speech lists in the experiments.

2 Results

The identification scores of the word and sentence tests, with the error bars representing the 95% standard deviations, are given in Figures 2 and 3. In both the tests, the scores of two listening ways (diotic and dichotic) changed non-monotonically with the masker number N increasing. The scores first declined gradually with N increasing from 1 to 6, and then rose up with N increasing further except for a slight drop occurring at N=32 in the sentence test with diotic listening. At every value of N, the scores of dichotic listening were higher than that of diotic listening in both the tests and the scores of diotic listening in the sentence test were

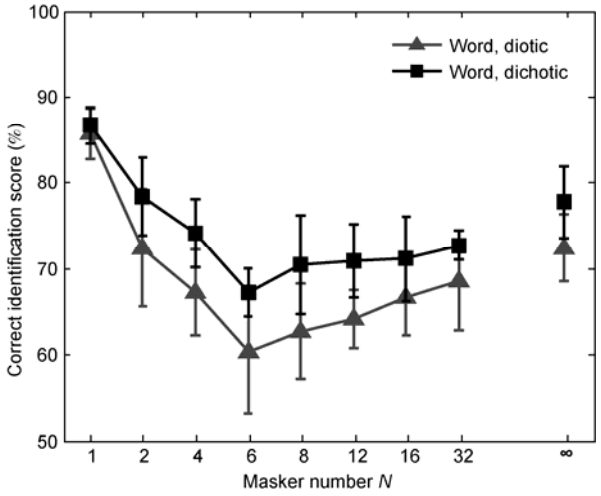

Figure 2 Chinese Mandarin speech identification scores of the word test in babble masking.

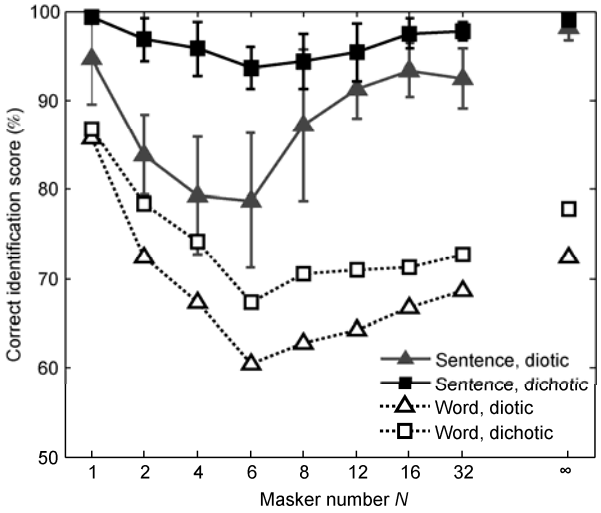

Figure 3 Chinese Mandarin speech identification scores of the sentence test in babble masking, with the results of the word test (dashed lines) as reference.

even higher than that of dichotic listening in the word test.

Analysis of variance was made for the factors N and the way of listening (diotic or dichotic). In the word test, both N, with $F_{(8,126)}$=15.6 and p<0.001, and the way of listening, with $F_{(1,126)}$=23.8 and p<0.001, had significant effect on the scores; but their interaction, with $F_{(8,126)}$=23.8 and p=0.933, did not. In the sentence test, the two factors and their interaction all had significant effect, with $F_{(8,126)}$=8.4 and p<0.001 for N, $F_{(1,126)}$=64.6 and p<0.001 for the way of listening, and $F_{(8,126)}$=3.5 and p<0.001 for the interaction of the two factors. And in both tests, the scores of diotic listening were statistically different from that of dichotic listening.

In the word test:

With diotic listening, significant differences of the scores existed between the condition of N=1 and that of all the other N values, and between the condition of N=∞ and that of N = {1, 6, 8, 12}. The scores of N = {4, 6, 8, 12, 16, 32} were not significantly different (p>0.05).

With dichotic listening, the score of N=1 differed from that of all other N values. The scores of N = {4, 6, 8, 12, 16, 32} were statistically equivalent, apart from the pairs (4, 6) and (6, 32). The score of N=∞ was not statistically different from that of the other N values except for N = {1, 6}.

In the sentence test:

With diotic listening, if a pair of N values both belonged to one of sets of N = {1, ∞}, N = {4, 6, 8}, and N = {1, 8, 12, 16, 32}, the corresponding scores were not significantly different (p>0.05); otherwise, they were.

With dichotic listening, the scores of every pair of N, except (1, 4), were not significantly different, if and only if the pair both belong to either one of the two sets of N = {1, 2, 4, 16, 32, ∞} and N = {2, 4, 6, 8, 12}.

3 Discussion

The result that the identification scores varied non-monotonically with N for both word and sentence tests of Chinese speech, agrees with the former findings for western languages by researchers such as Carhart et al. [7], Danhauer and Leppler [22], and Simpson and Cooke [9]. In fact, babble produces a combination of energetic and informational masking [23]. At low values of N, the listener appears to be able to take advantage of listening in the gaps of babble, where the level of babble is low, to overcome some linguistic confusions [8]. As N gets larger, the gaps are filled in and the spectro-temporal saturation of babble [8–10] also increases, resulting in a monotonically increasing of energetic masking. Combined with effect of informational masking, the situation becomes very complex. It appears that at some a value of N the informational masking would begin to decline with N increasing. Carhart [7] reported that the conditions of N = 4–32, and even of N=64, caused more masking than that of steady-state noise, although the informational masking was found in his study to reach its maximum at N=3 and thereafter decrease. The non-monotonic manner of identification scores changing with N increasing in this study also reflected the complex effects of the combination of energetic and informational masking. It is difficult for us to exactly determine at which value of N the informational masking in this study reached its maximum. From the experimental data obtained, we could only say that the value was within the domain of N < 8, because the increasing of the identification score of N = 8 over N = 6, which meaning the decreasing of the total masking, could be attributed to the decreased informational masking while the energetic masking generally keeping increasing monotonically.

It is worthwhile to compare more in detail the results in this paper and in [9] of word tests with diotic listening. The

comparison seems even more plausible when noticing the similarity in the experimental methods of the two studies which were based on identification of target words of nonsense syllables. Although it might be reasonable to view the essence of the scores defined in the two studies as the same, there could still exist difference in the degree of measuring speech identification between a pair of scores of a same value from the two different studies. The potential difference may limit the direct comparison between the scores of the two tests to some extent. However, if we ignore the potential difference temporarily, the observation that at every value of N the score of this paper is higher than that of [9] could be explained by the higher SNR in this study. Nevertheless, whether the difference is ignorable or not, the comparison between the trends of the scores is always reasonable. The first trend is at what N the scores reached their minimums. The value of $N = 6$ in this study is smaller than that of N=8 in [9] (although the condition of N=7 was not included in this study, N=7 is still smaller than 8, even if the minimums of score might actually have occurred there). It may be partly because of the reverberant listening environment of this study, which produced additional noticeable image sources with early and strong reflections and some other more masking with later reflections. Another thing is that in [9], the scores of a large range of N from 8 to 128 were not significantly different from each other, and different from that of N=512–∞. In this paper, the scores of $N \geqslant 8$ were not different significantly from that of N=∞. The reason might also mainly be the reverberation which smoothed the perception of the change of large N.

With both diotic and dichotic listening, the scores of sentence test were higher than that of word test at all the values of N. That means the linguistic connection of the words in a sentence can help the listener understand the sentence better. Someone at first thought might make a hypothesis that the linguistic connection (of course as informational linkage by its nature) in the target speeches would perhaps be more effective in reducing the informational masking of the babble than reducing the energetic masking. But by comparing the condition of N=6 with those of N>6, where the latters had less informational masking and more energetic masking than the former, we can see that the improvement of the latters was higher than the former, implying the linguistic connection in sentence might be even more effective in reducing energetic masking, or at least the above hypothesis seemed lack of supportive evidence. It seems also plausible to compare the conditions of N=∞ and N=1. At N=∞, where the energetic masking was predominant, the scores were about 70% in the word test and near 100% in the sentence test, the difference between the scores being significantly large. At N=1, where the energetic masking was at its minimum, the difference between the scores was much less. Partly, it could be explained by the fact that at N=1 the score was already as high as about 90% in the word test, thus leaving little room for the score to rise further. It could also be explained as the linguistic connection in sentence may be less effective in reducing informational masking.

In the word test of this study, the scores at all N values were improved by dichotic listening over diotic listening. The least improvement was at N=1. However, in the condition of N=1, the masking source was located on the same symmetrical axis of the virtual room as the target source and the receiver were. The target source (and so the masking source) gave identical signals to the two ears of the listener, which might make it difficult for the advantage of dichotic listening to take effect. For larger finite N values, the improvement of dichotic listening were obvious. And the effect of spatial separation of the talkers, which was further enhanced by the talkers sending out different signals, may be helpful for the listener to trace the target speech more easily. The conditions of N=∞ and N=1 had the same masking source position, but the former achieved a little more improvement than the later. It seems that under a stead state interference, the dichotic listening maybe more helpful for the listener to concentrate on the target source, leading to a better perception of the target speech. In the sentence test, the scores at all N values were also improved by dichotic listening over diotic listening, even at N=1. At small N values in the domain of N <8 where the N value causing maximum informational masking lied in, the improvement in the sentence test, brought by dichotic listening, was obviously larger than that in the word test, while at larger N values where the energetic masking got stronger, the improvement in the sentence test was not obviously larger than that in the word test. But still, it is difficult for one to assume that the dichotic listening might be more effective in reducing informational masking in the sentence test, because the scores at large N values were already high, leaving very small margins for further improvement, especially at N=∞ where the score with diotic listening was near to 100% already.

At last, it should be pointed out that, this paper basically intends to study the identification of target speeches in the babble of different masker numbers. As found by Brungart et al. [24], with a given masker number N, the identification score was generally the lowest in the condition that all the maskers were of the same gender as the target talker, highest in the condition that all the maskers were of the same gender different from the target talker, and some place between in the condition with a fixed gender ratio of maskers. This study followed Hoen et al. [10] to use half number of male and female for all the conditions of finite masker number $1<N<\infty$. The only exception was the condition of N=1, where the masker corresponding to the male or female target talker was of the same gender as the talker. However, although the masking performance of one of the maskers in the condition of N=2, whose gender was different from the target talker, was deemed as comparatively weaker than the other masker, it shows the increasing of masker number of

N=2 over *N*=1 still significantly enhanced the total masking, observing the changes of score between the two conditions to be almost the largest ones among those score changes between any two successive conditions of *N* in all the tests under study.

4 Summary

Chinese Mandarin speech identification in multi-talker babble in a simulated reverberant environment with spatially separated talkers was studied through subjective speech identification tests of word and sentence, with diotic and dichotic listening respectively. In both the tests, the speech identification scores changed non-monotonically with the masker number *N* increasing from 1 to infinity, which first declined gradually until reaching their minimums and then rose up. Statistical difference was found between the scores of diotic and dichotic listening. Dichotic listening achieved higher scores than diotic listening for all the values of *N* checked in both the tests, showing the binaural effect has an advantage for reducing babble masking. With both diotic and dichotic listening, the scores of sentence test were significantly higher than the word test for all the values of *N*, showing the linguistic connection in sentence could help the listener achieve a better perception of the target speech in babble masking. The results could be applied in fields such as telephone/video conference and virtual reality. For example, the speech identification could be enhanced with considering the spatial factors of the speeches, for instance, to record the speeches using artificial head or sound-field microphone.

However, this study was performed under a single reverberation condition and one signal to noise ratio. Further researches will include more reverberation conditions and signal to noise ratios.

The authors thank the students who participated in subjective tests of speech identification as listeners, and those who helped in recording the speech signals as talkers. Also the authors pay deep gratitude to the reviewers for their valuable suggestions and questions which substantially helped improve the scientific quality of the manuscript. This work was supported by the National Natural Science Foundation of China (10774048, 51078326), Science Foundation of Zhejiang Province, China (Y5090138) and Science and Technology Planning Project of Guangdong Province, China (2011B061300066).

1 Cherry E C. Some experiments on the recognition of speech, with one and two ears. J Acoust Soc Am, 1953, 25: 975–979
2 Yost W A, Dye R H, Sheft S. A simulated cocktail party with up to three sound sources. Percept Psychophys, 1996, 58: 1026–1036
3 Hawley M L, Litovsky R Y, Colburn H S. Intelligibility and localization of speech signals in a multisource environment. J Acoust Soc Am, 1999, 105: 3436–3448
4 Arbogast T L, Mason C R, Kidd G J. The effect of spatial separation on informational and energetic masking of speech. J Acoust Soc Am, 2002, 112: 2086–2098
5 Freyman R L, Balakrishnan U, Helfer K S. Effect of number of masking talkers and auditory priming on informational masking in speech recognition. J Acoust Soc Am, 2004, 115: 2246–2256
6 Agus T R, Akeroyd M A, Noble W, et al. An analysis of the masking of speech by competing speech using self-report data. J Acoust Soc Am, 2009, 125: 23–26
7 Carhart R, Johnson C, Goodman J. Perceptual masking of spondees by combinations of talkers. J Acoust Soc Am, 1975, 58: S35
8 Brungart D. Informational and energetic masking effects in the perception of two simultaneous talkers. J Acoust Soc Am, 2001, 109: 1101–1109
9 Simpson S, Cooke M P. Consonant identification in *N*-talker babble is a nonmonotonic function of *N*. J Acoust Soc Am, 2005, 118: 2775–2778
10 Hoen M, Meunier F, Grataloup C. Phonetic and lexical interferences in informational masking during speech-in-speech comprehension. Speech Comm, 2007, 49: 905–916
11 Brungart D S, Simpson B D. The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. J Acoust Soc Am, 2002, 112: 664–676
12 Culling J F, Hodder K I, Toh C Y. Effects of reverberation on perceptual segregation of competing voices. J Acoust Soc Am, 2003, 114: 2871–2876
13 Marrone N, Mason C R, Kidd G J. The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. J Acoust Soc Am, 2008, 124: 3064–3075
14 Kidd G J, Mason C R, Brughera A, et al. The role of reverberation in release from masking due to spatial separation of sources for speech identification. Acta Acust United Ac, 2005, 91: 526–536
15 Yang W, Hodgson M. Validation of the auralization technique: Comparative speech intelligibility tests in real and virtual classrooms. Acta Acust United Ac, 2007, 93: 991–999
16 Peng J X. Feasibility of subjective speech intelligibility assessment based on auralization. Appl Acoust, 2005, 66: 591–601
17 GB/T 15508. Acoustics—Speech articulation testing method. Standard of P. R. China, 1995
18 Wong L L, Soli S D, Liu S, et al. Development of the Mandarin hearing in noise test (MHINT). Ear Hearing, 2007, 28: 70–74
19 Zhang H, Chen J, Wang S, et al. Edit and evaluation of Mandarin sentence materials for Chinese speech audiometry. Chin J Otorhinolaryngol Head Neck Surg, 2005, 40: 774–778
20 Steeneken H J M, Houtgast T. Phoneme-group specific octave-band weights in predicting speech intelligibility. Speech Comm, 2002, 38: 399–411
21 Steeneken H J M, Houtgast T. Validation of the revised STIr method. Speech Comm, 2002, 38: 413–425
22 Danhauer J L, Leppler J G. Effects of four noise competitors on the California Consonant Test. J Speech Hear Disord, 1979, 44: 354–362
23 Cullington H E, Zeng F G. Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects. J Acoust Soc Am, 2008, 123: 450–461
24 Brungart D S, Simpson B D, Ericson M A, et al. Informational and energetic masking effects in the perception of multiple simultaneous talkers. J Acoust Soc Am, 2001, 110: 2527–2538