Front. Electr. Electron. Eng. 2012, 7(1): 107–115
DOI 10.1007/s11460-012-0189-8

RESEARCH ARTICLE

Xihong WU, Jing CHEN

A computational model for assessment of speech intelligibility in informational masking

Abstract The existing auditory computational models for evaluating speech intelligibility can only account for energetic masking, and the effect of informational masking is rarely described in these models. This study was aimed to make a computational model considering the mechanism of informational masking. Several psychoacoustic experiments were conducted to test the effect of informational masking on speech intelligibility by manipulating the number of masking talker, speech rate, and the similarity of F0 contour between target and masker. The results showed that the speech reception threshold for the target increased as the F0 contours of the masker became more similar to that of the target, suggesting that the difficulty in segregating the target harmonics from the masker harmonics may underlie the informational masking effect. Based on these studies, a new auditory computational model was made by inducing the auditory function of harmonic extraction to the traditional model of speech intelligibility index (SII), named as harmonic extraction (HF) model. The predictions of the HF model are highly consistent with the experimental results.

Keywords auditory computational model, speech intelligibility, informational masking, F0 contour, harmonic extraction

1 Introduction

Speech talking is a main way of communication for human's everyday life. There are usually several sound sources existing in the real communication environment except the target speech, such as reflections, noise from machines, or interfering speech from other talkers, all of which can affect speech communication. Speech intelligibility is a measure of the effectiveness of speech communication [1], which is an integrated performance from auditory analysis to linguistic analysis in human brain. Studies on the related theories and computational models have been an important branch in fields of hearing research and speech perception.

Two main factors are thought to contribute to the reduced intelligibility when target speech is masked by interfering sounds: 1) energetic masking (EM), which occurs when peripheral neural activity elicited by a signal is overwhelmed by that elicited by the masker, leading to a degraded or noisy neural representation of the signal; 2) informational masking (IM), which is also called "non-energetic masking" and conceptualized as anything that reduces intelligibility once energetic masking has been accounted for, including effects, such as difficulty in determining how to assign acoustic elements in the mixture to the target and masker [2–7]. The effect of purely energetic masking on speech intelligibility has been well documented and can be evaluated by auditory computational models. The effects of informational masking on speech intelligibility are more complicated, involving multiple levels of processing, and are rarely described by current computational models.

Effects of IM on speech perception have been studied by manipulating the stimulus characteristics. Brungart et al. [4] found that the recognition of speech in multitalker environments generally decreased when the target and masking talkers had similar voice characteristics: the target was more intelligible when the masker and the target were spoken by different-sex talkers than when they were spoken by same-sex talkers or the same talker. When the number of masking talkers was manipulated, the results showed that speech recognition was a non-monotonic function of the number of masking talkers [8,9]. It was also reported that a native-language speech masker produced more IM than a non-native

Received October 10, 2011; accepted December 15, 2011

Xihong WU (✉), Jing CHEN (✉)
Speech and Hearing Research Center, Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China
E-mail: wxh@cis.pku.edu.cn, chenj@cis.pku.edu.cn

speech masker [3]. Similarly, time-reversed speech produced less IM than normal speech, but performance with a time-reversed speech masker was poorer than for nonnative speech masker, perhaps due to the increased forward masking for the former [10]. Although these studies revealed the phenomena for the effect of IM on speech intelligibility, the mechanism remains unclear, and it is difficult to produce a computational model based on current studies.

The first computational model to describe the masking effect on speech intelligibility was based on the theory of articulation index [11,12] by Bell Lab, which was developed as an ANSI standard in 1969 [13]. In the following research, the concept of articulation index (AI) was replaced by speech intelligibility index (SII), and the ANSI standard was renewed as Ref. [14]. In China, a national standard of SII was also developed based on Zhang and Ma's research [15–17]. Several computational models apart from the SII have been developed for predicting speech intelligibility. The speech transmission index (STI) is based on the generation and analysis of an artificial test signal that replaces the speech signal and is specifically applied in room acoustics [18]. The spectro-temporal modulation index (STMI) is based on a model of how the auditory cortex analyzes the spectrotemporal modulations presented in speech [19,20]. These two models account correctly for several situations, such as band-limiting, noise, reverberation, and so on, but they are not effective for evaluating speech intelligibility when the target speech is masked by maskers containing informational masking, such as speech or speech-like sounds [21].

To produce a computational model for evaluating speech intelligibility in IM, we studied both the mechanism of IM and speech intelligibility models. In this paper, human behavior experiments are introduced first in Sect. 2 as studies on mechanism. An auditory computational model called the "harmonic feature model" is developed and described in Sect. 3, which can account for part of the experimental results. Finally, a discussion section based on these work is given.

2 Studies on mechanism of informational masking

When target speech is masked by a normal speech, both EM and IM occur, and their relative contribution cannot be easily divided directly according to the speech intelligibility scores. To evaluate the effect of IM, a psychoacoustic paradigm has been used in several studies by introducing a perceived spatial separation between the target speech and the masker, via the precedence effect [3,6,22,23]. For example, when the target and masker are both presented by a loudspeaker to the listener's right and a loudspeaker to the listener's left, and the sound from the right loudspeaker leads that from the left loudspeaker by 3 ms for both the target and masker, both the target and masker are perceived as coming from the right loudspeaker, due to the precedence effect [24]. In other words, the target and masker are perceived as being co-located. However, if the delay between the two loudspeakers is reversed only for the masker, the target is still perceived as coming from the right loudspeaker, but the masker is perceived as coming from the left loudspeaker. Thus, the relative perceived locations of the target and masker can be manipulated without substantially changing sound levels or spectra at the ears. It has been confirmed that for both Chinese and English speech materials, when the masker is speech, a perceived spatial separation between the target speech and masker can lead to a 3–8 dB release from masking, but when the masker is speech-spectrum noise, the release from masking is only about 1 dB [3,6].

Although several studies revealed that speech recognition was a non-monotonic function of the number of masking talkers, as mentioned earlier, this conclusion is based on English speech, and it is hard to induce it on Chinese speech due to the language difference between Chinese and English [6]. A study was developed to test the effect of the number of masking talkers on speech-on-speech masking in Chinese. Targets were nonsense sentences spoken by a Chinese female, and maskers were nonsense sentences spoken by other 1, 2, 3, or 4 Chinese females. All stimuli were presented by two spatially separated loudspeakers. Using the precedence effect, manipulation of the delay between the two loudspeakers for the masker determined whether the target and masker were perceived as coming from the same or different locations. The results show that the masking effect remarkably increased with the number of masking talkers increased progressively from 1 to 4, which is also confirmed by the calculation of the speech intelligibility index. However, the perceived spatial separation, which predominantly reduced informational masking, caused the largest improvement in speech identification with the two-talker masker, indicating that two-voice speech had the highest informational masking impact. Details of this work can be found in Ref. [25].

Target speech can be better recognized under speech-on-speech masking conditions if certain differences between target and masker (e.g., in loudness, pitch, and location) can be used as cues for streaming. However, it is rarely reported in previous studies whether the speech rate can be a cue for releasing IM. In this experiment, the rate difference between target and masking speech was manipulated by changing the rate of masking speech using the synchronized overlap-add fixed synthesis (SOLAFS) algorithm [26], which can change speech rate without pitch shifting. The speech rate ratio of target

speech to masking speech (the speech rate ratio, SRR) was quantified. Both target and masker speech were Chinese nonsense sentences, and they were co-presented with the signal-to-masker ratio (SMR) of –7 dB. The results show that speech recognition was significantly increased with the SRR increased from 1 to 1.5 or the SRR decreased from 1 to 0.5. Moreover, the unmasking effect of precedence-induced perceived spatial separation on target-speech recognition was increased monotonically with the increase of the SRR from 0.5 to 1.5. These results suggest that the speech rate is one of the factors influencing both EM and IM of Chinese speech.

The large effect of perceived spatial separation for the speech masker but not the noise masker is caused by the large IM contained in the former but not in the latter. Phonemes, words, and syllables from the masking speech may be confused with those from the target speech. Potentially, these components of IM can be reduced by using a speech-like synthesized masker without linguistic content. The target stimuli were nonsense Chinese Mandarin sentences spoken by a young female. The masking stimuli were synthesized signals with four types of F0 contours and steady speech-spectrum noise. The intention was to synthesize signals with similar acoustic characteristics to speech, such as the harmonic structure with fluctuating F0 contour during voiced parts, and noise-like spectra during unvoiced parts, except that there was no formant structure. Formants supply essential cues for phoneme identification, so the synthesized signals were completely unintelligible and so should not activate "knowledge-driven" forms of IM. However, they should lead to stimulus-driven IM. We hypothesized that the most stimulus-driven IM would occur when the target and masker had identical F0 contours. The similarity between the target speech and masking synthesized signals was manipulated by using maskers with different F0 contours. Based on the original F0 of the target speech, these F0 contours were modified using the following formula:

$$F'0(t) = \overline{F0} \times \exp\left(m \times \ln \frac{F0(t)}{\overline{F0}}\right), \qquad (1)$$

where $F'0(t)$ represents the modified F0 contour, $F0(t)$ represents the original F0 contour, and $\overline{F0}$ represents the mean F0 of the sentence [27]. Four manipulation coefficients (m) of the F0 contour were applied (1, 0, –1, and 2) corresponding to the four conditions (original, flat, reversed, and amplified).

The one-third octave spectra for one sentence of the target and for each type of masker are presented in Fig. 1(a). Although the F0 contours differed among the four synthesized maskers, the energy distribution across frequency-bands was consistent, so their effect on speech intelligibility as predicted by the SII would have been identical. Hence, if four maskers have different effects on speech intelligibility, the difference is probably caused by IM. Figure 1(b) shows average percent correct word identification as a function of SMR for the five maskers. The results indicate that maskers with FM harmonics produced more masking than the steady speech-spectrum noise, and synthesized maskers whose F0 contour was similar to that of the target produced the greatest masking effect. More details about this work can be found in Ref. [28].

3 Harmonic feature model

3.1 Basic theory

The fact that the speech-like synthesized maskers produced masking effects greater than that produced by steady speech-shaped noise suggests that the harmonic structures of an interfering sound can contribute to IM and probably plays a role in the IM that occurs in

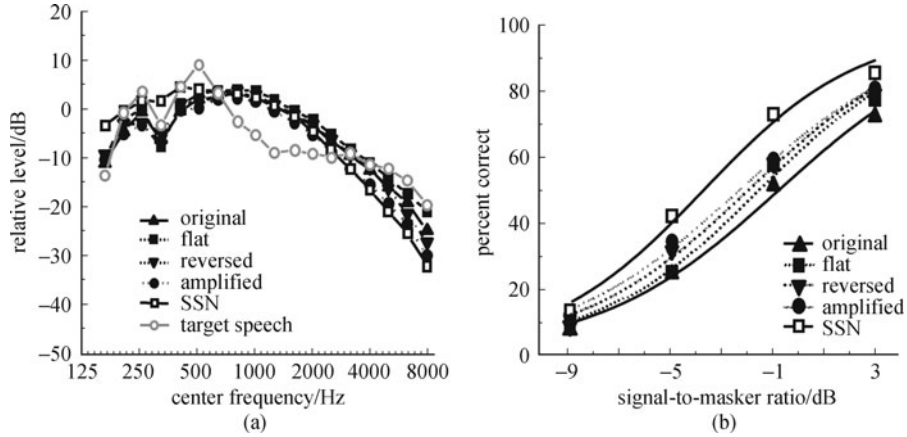

Fig. 1 (a) Averaged one-third octave band spectra for target (open circles) and five types of maskers; (b) symbols show the mean percent correct identification of key words across 10 listeners as a function of SMR for the five masking conditions (adapted from Ref. [25])

speech-on-speech masking situations. The more similar the pitch contour of masker to that of the target speech, the greater was the masking effect, suggesting that the difficulty in segregating the target harmonics from the masker harmonics may underlie the IM effect. Source separation based on harmonic structure has been modeled in several ways, using the concept of a "harmonic sieve" [29], auto-correlation models [30–32], and the idea of harmonic cancellation of the masker [33]. These models have been successfully used to predict results of studies on the identification of concurrent vowels. With an auto-correlation model, the pitch contour of a target speech signal in quiet can be extracted accurately [34], and then, the autocorrelation value of the time-lag corresponding to the pitch in every allocated frequency channel, which contains one harmonic determined as a vector of harmonic features, called HF_{clear}. A similar vector of harmonic features can be extracted from the mixture of target and background; this is called HF_{mixed}. The difficulty of extracting the target harmonics from the mixture can be measured by computing the distance between the two vectors: HF_{clear} and HF_{mixed}. Figure 2 gives examples for the harmonic feature in a clear speech signal and the harmonic feature in a mixed signal. In each example, the autocorrelation function was computed in every frequency channel and shown channel by channel in the main panel as the autocorrelogram, and the pitch value can be extracted from the summary normalized correlogram at the bottom panel. In frequency channels containing harmonics, the autocorrelation value of the lag corresponding to the pitch were signed with stars in the autocorrelogram, and then, harmonic feature was given with these allocated autocorrelation value and shown in Fig. 2(b).

The spectra of unvoiced speech sounds are different from those of vowels in a number of ways [35]. However, given that the focus of this study is on the role of the harmonic structure of speech and masking sounds, and the unvoiced segments in our maskers were bursts of speech-shaped noise, the masking effect of the unvoiced segments is treated here as being solely energetic masking and is computed with a traditional SII model.

3.2 Model description

Figure 3 shows a schematic diagram of the model. The input to the model was the clear target speech as the reference signal and the mixed speech and masker as the test signal. Both signals were processed using a model of the auditory periphery based on the AIM-MAT toolbox [36], in which the stages were: a filter simulating the outer/middle ear transfer function [37]; a 64-channel gammatone filter bank with filter bandwidths as specified by Glasberg and Moore [38]; conversion of the output of each filter to a simulated neural activity pattern (NAP) by log compression, half-wave rectification and low-pass filtering, where the low-pass cutoff frequency was 1200 Hz for channels with center frequency less than 4000 Hz and 300 Hz for channels with center frequency is higher than 4000 Hz. The lower low-pass cutoff frequency for higher frequency channels is to extract the envelope, which reflects the periodicity information in higher frequency channels [39]. And then, an autocorrelation function was applied to the NAP using a frame length of 35 ms and an overlap between frames of 10 ms. The pitch contour of the clear target speech (reference signal) can be extracted from the output of autocorrelation calculation and determine the voiced segments and

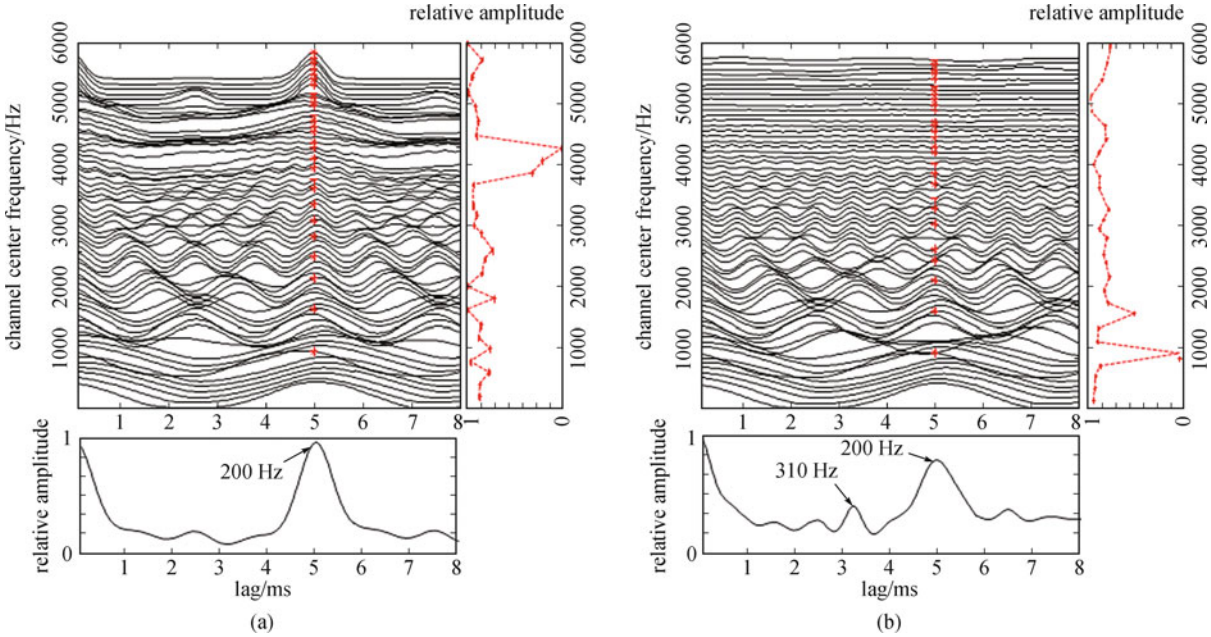

Fig. 2 Extraction of harmonic features from autocorrelogram of a synthetic vowel /u/ with an F0 of 200 Hz (a) and a synthetic vowel pair with the same /u/ and an additional /a/ with an F0 of 310 Hz (b)

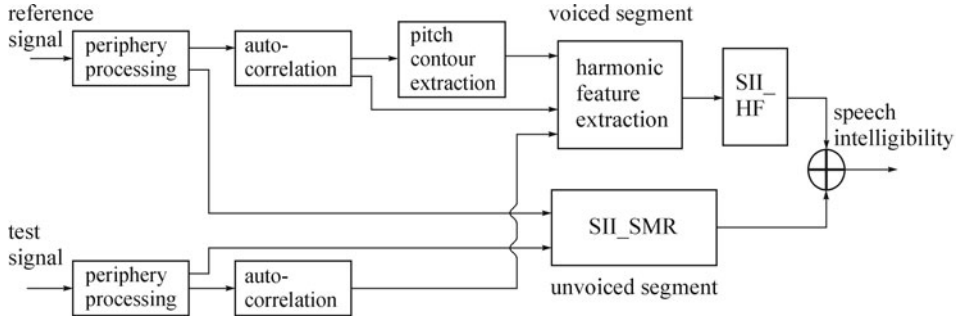

Fig. 3 Schematic diagram of HF model

the unvoiced segments. For the voiced segments, HF_{clear} and HF_{mixed} were extracted respectively from the autocorrelograms of the reference signal and of the test signal. The Euclidean distance between HF_{clear} and HF_{mixed} for a given frame was calculated using the following formula:

$$d(HF_{\text{mixed}}, HF_{\text{clear}}) = \sqrt{\sum_{i=1}^{n} \left(HF_{\text{mixed}}(i) - HF_{\text{clear}}(i)\right)^2}, \quad (2)$$

where d is the distance between these two harmonic feature vectors, i is the index of the vector, and n is the dimensional number of the vector and determined by dividing the value of the maximum harmonic frequency by the F0 value. To use this distance value in the SII model, the distance value was transformed to a speech intelligibility index using Eq. (5) below; this index is called the SII_HF, and the detailed calculation method will be introduced in the third paragraph. For the unvoiced segments, the target speech-to-masker ratio was computed and converted to a traditional speech intelligibility index, which is called here as the SII_SMR. Finally, SII_HF and SII_SMR were combined as a weighted function, as described by Eq. (3):

$$SR = \sum_{j=1}^{n} w(j) S_j, \quad (3)$$

where SR is the output of the model, S_j is the speech intelligibility of segment j, $w(j)$ is a weighting function and is defined by the ratio of the temporal length of segment j to the whole length of the input signal.

In a traditional SII model, the SII can be used to predict speech intelligibility via a transfer function, such as that recommended by Fletcher and Galt [12]:

$$S = (1 - 10^{-AP/Q})^N, \quad (4)$$

where S is the percent correct intelligibility score, A is the SII value, P is a proficiency factor that accounts for the proficiency of the talker and listener and is usually assumed to be 1, and Q and N are fitting constants depending on the speech characteristics [40]. Here, A corresponds to SII_SMR and was determined by multiplying the value of the SMR within each frequency band with the frequency-important function [40,41] and summing across bands. In a previous study, a 1/3 octave frequency importance function and parameter values of Q (0.525) and N (2.58) based on Chinese Mandarin were found to give good fits to the data [42].

The SII_HF is an index corresponding to the distance between HF_{clear} and HF_{mixed}, where a larger value corresponds to lower speech intelligibility. Hence, the transfer function was modified as defined by Eq. (5):

$$S = (1 - 10^{-1/(SII\_HF \times Q')})^{N'}, \quad (5)$$

where *SII_HF* is the normalized value of the distance from Eq. (2), and Q' and N' are parameters that are adjusted to fit the experimental results. The experimental data used for fitting were the speech recognition scores from five subjects for the masking condition "original" at four SMRs, which is taken from the experiment shown in Fig. 1. There were 20 data points, and correspondingly, 20 values of speech intelligibility were computed using the HF model using five test sentences with each of the synthesizes maskers. Fitting these data through the Levenberg-Marquardt method gave parameter values $Q' = 1.12$ and $N' = 50$.

3.3 Model evaluation

The HF model was used to predict the data of F0 contour experiment for all 20 conditions (5 masker types × 4 SMRs), and the predictions were compared with the obtained results. In the experiment, 15 sentences were used in each condition, but five of these for masking condition "original" were used in the parameter fitting, so only the remaining 10 sentences were used for evaluation.

As shown in Fig. 4(a), speech recognition performance predicted by the model is similar to that of the experiment manipulating pitch contour mentioned earlier. Predicted speech recognition for all five masking conditions systemically increased as the SMR was increased from –9 dB to 3 dB. Predicted recognition for the steady noise masker was clearly higher than for the four synthesized maskers. Predicted speech recognition for the

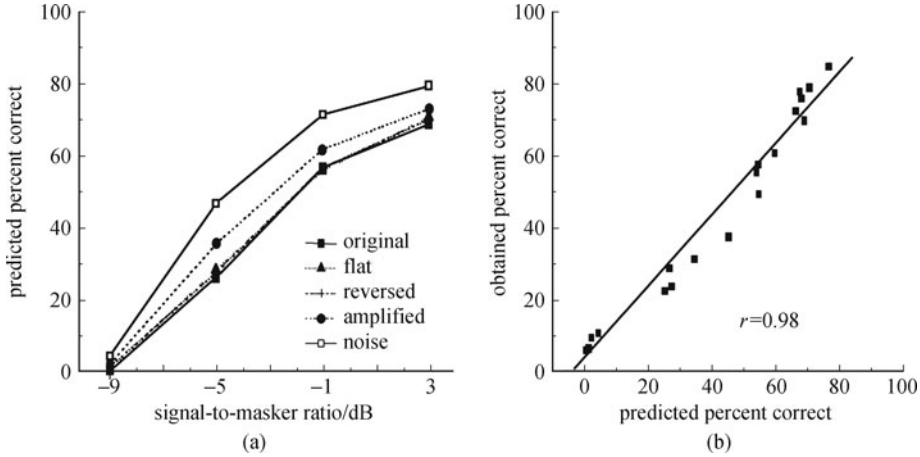

Fig. 4 (a) Percent correct speech recognition predicted by HF model for F0 contour experiment; (b) a scatter plot of predicted versus obtained percent correct

"amplified" masker was higher than that for condition "original", "flat", and "reversed". The predicted results were similar but not identical to the averaged scores shown in Fig. 1 and were consistent with the obtained scores. The correlation between the predicted and obtained scores is shown in Fig. 4(b). The correspondence is very good, with a correlation coefficient of 0.98.

To compare the performance of HF model to the existing models, such as SII, STI, and STMI mentioned earlier, the same test material was used to get the predicted percent correct of these models, respectively. For the SII model, the calculation method of ANSI 1997 [14] was adapted, and the frequency important was from Ref. [42], which is specific for Chinese speech. For the STI model, the method introduced in Ref. [18] was used, and the Matlab program was from the official website. For the STMI model, the calculation tool was introduced in Ref. [20], and given by the corresponding author of this reference. Figure 5 shows the results for each model, where the subplot represents SII, STI, and STMI from left to right, respectively. The correlation value between the human performance and model scores are also signed in the figure, and they are 0.84 for SII, 0.90 for both STI and STMI. The relatively higher scores of STI and STMI are probably because maskers used in this study do not cause distortions like reverberation, phase-jitter or band-noise, which are mainly accounted in STI and STMI. Apparently, the correlation value for HF model, 0.98, is consistently higher than those of the existing models, confirming that the HF model is effective for evaluating speech intelligibility in informational masking.

4 Discussion and conclusions

4.1 Effects of F0 contour

It is well known that differences in F0 contour between a target talker and competing talkers can facilitate tracking of the target talker [31,43,44]. The F0 contour of "original" masker was almost identical to that of the target. As a result, the only cue could be used to discriminate the target speech from the masker was the short-term spectral envelope and changes in spectral envelope over time. When the F0 contour was manipulated

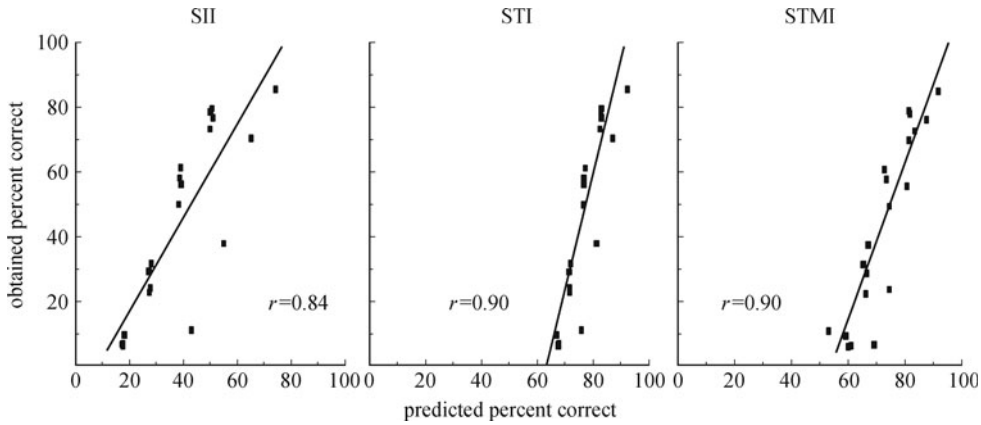

Fig. 5 Each subplot shows a scatter plot of predicted versus obtained percent correct. The predicted percent corrects are computed by the models of SII (left), STI (middle), STMI (right), respectively. The obtained percent correct is consistent with the result of F0 contour experiment

by reversal, amplification, or flattening, this decreases the similarity of FM of the target and masker providing additional perceptual cues. The variation of the threshold for speech recognition across the five maskers confirmed that the similarity of F0 contour played a significant role. Moreover, since the spectra of the four types of synthesized maskers were almost identical, these four maskers would have produced similar energetic masking, at least over the frequency range where the harmonics were not resolved. The changes in threshold with synthetic masker type are probably therefore mainly due to changes in IM.

4.2 Informational masking

Of the many factors that can affect IM, this study focused on the acoustic characteristic of the masking sound using a speech-like signal that conveyed no phonetic information, thus making it likely that any IM that was found would depend mainly on signal-driven factors (not knowledge-driven factors that related language content). Perceptual similarities between the target and the masker are considered to be one of the main causes of IM. IM produced by a synthetic speech-like sound was greatest when its F0 contour was identical to that of the target speech. The results suggested that IM can occur due to the similarity of F0 contour of the target and masker, even when the masker is unintelligible.

The effect of informational masking on speech intelligibility in this study was hypothesized and modeled in term of the difficulty of extracting the harmonics of the target from the mixed signals, which is consistent with the conventional thought that IM occurs because the listener has difficulty attending to the target and perceptually disentangle it from the interference [3,4,7]. It might be argued that speech intelligibility under the "original" condition is worse than other synthetic masking conditions due to the more energetic masking induced by its identical F0 contour as that of the target; however, it has been discussed in Durlach's review paper [5] that even though the EM is defined as "occurring due to the overlap between target and masker at the periphery", there remains the problem of specifying what is meant by "periphery" and "overlap", and according to this definitional structure, what appears as IM at one level can appear as EM at a higher level, and all masking is EM if examined at a sufficiently high level. In a traditional model of auditory periphery, which is identified with the auditory nerve, the output of the four synthetic maskers used this research are identical and could not reflect their different effect on speech intelligibility (see Fig. 1), but when a higher level processor, auto-correlation computing for harmonic extraction in this paper, is induced, these four maskers can be differentiated. Based on the rationale above, the effect of F0 contour on speech intelligibility is considered as IM.

4.3 HF model

The SII model is limited in that it can only predict speech intelligibility for situations dominated by energetic masking. Following the basic framework of the SII model, the HF model additionally characterizes the difficulty of extracting harmonic feature of the target signal from the mixture. The predictions of the HF model are consistent with the experimental results, and its performance is better than the existing models.

Because the HF model can account for the effects of both energetic masking and IM arising from signal-driven processes, it could be employed in other psychoacoustic studies based on speech-on-speech masking to analyze the relative contribution of signal-driven processes and knowledge-driven processes, such as studies of masking by native versus non-native languages. However, further work is needed to assess how well the HF model works for such data.

Acknowledgements The work was supported in part by the National Natural Science Foundation of China (Grant Nos. 90920302 and 91120001), the HGJ Grant (No. 2011ZX01042-001-001), the project from the Ministry of Science and Technology of China (No. 2010DFA31520), and a Newton International Fellowship Alumni Follow-on Funding from the Royal Society, UK.

References

1. Geneva: International Organization for Standardization. ISO 9921, Ergonomics — Assessment of speech communication. 2003
2. Watson C S. Uncertainty, informational masking, and the capacity of immediate auditory memory. In: Yost W A, Watson C S, Eds. Auditory Processing of Complex Sounds. NJ: Lawrence Erlbaum Associates, 1987, 267–277
3. Freyman R L, Balakrishnan U, Helfer K S. Spatial release from informational masking in speech recognition. Journal of the Acoustical Society of America, 2001, 109(5): 2112–2122
4. Brungart D S, Simpson B D, Ericson M A, Scott K R. Informational and energetic masking effects in the perception of multiple simultaneous talkers. Journal of the Acoustical Society of America, 2001, 110(5): 2527–2538
5. Durlach N I, Mason C R, Kidd G Jr, Arbogast T L, Colburn H S, Shinn-Cunningham B G. Note on informational masking. Journal of the Acoustical Society of America, 2003, 113(6): 2984–2987
6. Wu X H, Wang C, Chen J, Qu H W, Li W R, Wu Y H, Schneider B A, Li L. The effect of perceived spatial separation on informational masking of Chinese speech. Hearing Research, 2005, 199(1–2): 1–10
7. Mattys S L, Brooks J, Cooke M. Recognizing speech under

a processing load: Dissociating energetic from informational factors. Cognitive Psychology, 2009, 59(3): 203–243

8. Freyman R L, Balakrishnan U, Helfer K S. Effect of number of masking talkers and auditory priming on informational masking in speech recognition. Journal of the Acoustical Society of America, 2004, 115(5): 2246–2256
9. Simpson S A, Cooke M. Consonant identification in N-talker babble is a nonmonotonic function of N. Journal of the Acoustical Society of America, 2005, 118(5): 2775–2778
10. Rhebergen K S, Versfeld N J, Dreschler W A. Release from informational masking by time reversal of native and non-native interfering speech. Journal of the Acoustical Society of America, 2005, 118(3): 1274–1277
11. French N R, Steinberg J C. Factors governing the intelligibility of speech sounds. Journal of the Acoustical Society of America, 1947, 19(1): 90–119
12. Fletcher H, Galt R H. The perception of speech and its relation to telephony. Journal of the Acoustical Society of America, 1950, 22(2): 89–151
13. ANSI. ANSI S3.5, Methods for the calculation of the articulation index. New York: American National Standards Institute, 1969
14. ANSI. ANSI S3.5-1997, Methods for the calculation of the speech intelligibility index. New York: American National Standards Institute, 1997
15. Zhang J L. Statistic relations on articulation index across different speech test materials. Acoustics, 1964, 1: 90–94 (in Chinese)
16. Zhang J L, Ma D Y. A new method for calculating articulation index. Acoustics, 1965, 2: 80–84 (in Chinese)
17. Zhang J L. The statistic relation on articulation index between syllable and phoneme. Physics, 1974, 23: 315–320 (in Chinese)
18. Houtgast T, Steeneken H J. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. Journal of the Acoustical Society of America, 1985, 77(3): 1069–1077
19. Chi T, Gao Y, Guyton M C, Ru P, Shamma S. Spectro-temporal modulation transfer functions and speech intelligibility. Journal of the Acoustical Society of America, 1999, 106(5): 2719–2732
20. Elhilali M, Chi T, Shamma S A. A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. Speech Communication, 2003, 41(2–3): 331–348
21. Chen J. Mechanism of informational masking and computational model for evaluating speech intelligibility. Dissertation for the Doctoral Degree. Beijing: Peking University, 2009 (in Chinese)
22. Li L, Daneman M, Qi J G, Schneider B A. Does the information content of an irrelevant source differentially affect spoken word recognition in younger and older adults? Journal of Experimental Psychology: Human Perception and Performance, 2004, 30(6): 1077–1091
23. Huang Y, Huang Q, Chen X, Qu T S, Wu X H, Li L. Perceptual integration between target speech and target-speech reflection reduces masking for target-speech recognition in younger adults and older adults. Hearing Research, 2008, 244(1–2): 51–65
24. Litovsky R Y, Colburn H S, Yost W A, Guzman S J. The precedence effect. Journal of the Acoustical Society of America, 1999, 106(4): 1633–1654
25. Wu X H, Chen J, Yang Z G, Huang Q, Wang M Y, Li L. Effect of number of masking talkers on speech-on-speech masking in Chinese. In: Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech2007). 2007, 390–393
26. Henja D, Musicus B. The solafs time-scale modification algorithm. Technical report. Bolt Beranek & Newman, 1991
27. Binns C, Culling J F. The role of fundamental frequency contours in the perception of speech against interfering speech. Journal of the Acoustical Society of America, 2007, 122(3): 1765–1776
28. Chen J, Li H H, Li L, Moore C J B, Wu X H. Informational masking of speech produced by speech-like sounds without linguistic content. Journal of Acoustic Society of America, 2011 (conditionally accepted)
29. Scheffers M T M. Sifting vowels: Auditory pitch analysis and sound segregation. Dissertation for the Doctoral Degree. Groningen, Netherlands: University of Groningen, 1983
30. Licklider J C R. A duplex theory of pitch perception. Experientia, 1951, 7(4): 128–134
31. Assmann P F, Summerfield Q. Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency. Journal of the Acoustical Society of America, 1989, 85(1): 327–338
32. Meddis R, Hewitt M J. Modeling the identification of concurrent vowels with different fundamental frequencies. Journal of the Acoustical Society of America, 1992, 91(1): 233–245
33. de Cheveigné A. Concurrent vowel identification. III: A neural model of harmonic interference cancellation. Journal of the Acoustical Society of America, 1997, 101(5): 2857–2865
34. Cooke M, Ellis D P W. The auditory organization of speech and other sources in listeners and computational models. Speech Communication, 2001, 35(3–4): 141–177
35. Greenberg S, Ainsworth W. Speech processing in the auditory system: An overview. In: Greenberg S et al., Eds. Speech Processing in the Auditory System. Springer: Berlin, 2004, 20–22
36. Patterson R D, Allerhand M H, Giguère C. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. Journal of the Acoustical Society of America, 1995, 98(4): 1890–1894
37. Glasberg B R, Moore B C J. A model of loudness applicable to time-varying sounds. Journal of the Audio Engineering Society, 2002, 50(5): 331–342
38. Glasberg B R, Moore B C J. Derivation of auditory filter shapes from notched-noise data. Hearing Research, 1990, 47(1–2): 103–138
39. Wu M, Wang D, Brown G J. A multipitch tracking algorithm for noisy speech. IEEE Transactions on Speech and Audio Processing, 2003, 11(3): 229–241

40. Studebaker G A, Sherbecoe R L, Gilmore C. Frequency-importance and transfer functions for the Auditec of St. Louis recordings of the NU-6 word test. Journal of Speech and Hearing Research, 1993, 36(4): 799–807
41. Rhebergen K S, Versfeld N J, Dreschler W A. Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. Journal of the Acoustical Society of America, 2006, 120(6): 3988–3997
42. Huang Q. Frequency importance function of Mandarin Chinese speech and models for speech intelligibility evaluation. Dissertation for the Master's Degree. Beijing: Peking University, 2007 (in Chinese)
43. Brokx J P L, Nooteboom S G. Intonation and the perceptual separation of simultaneous voices. Journal of Phonetics, 1982, 10(1): 23–36
44. Darwin C J, Brungart D S, Simpson B D. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. Journal of the Acoustical Society of America, 2003, 114(5): 2913–2922

Xihong WU received his B.S. degree from Jilin University in 1989, and Ph.D. degree from the Department of Radio-Electronics of Peking University (PKU) in 1995. He began his postdoctoral research in 1995, became an assistant professor in 1997, an associate professor in 1999 and a full professor in 2004 in the National Laboratory on Machine Perception at PKU. He is currently the director of the Speech and Hearing Research Center, and the deputy dean of Key Laboratory of Machine Perception (Ministry of Education) at PKU. His research interests are computational auditory models, auditory scene analysis, auditory psychophysics, speech signal processing, and natural language processing. In these areas he has published over 120 papers in leading international journals and conference proceedings. He is an Associate Editor of *Neural Network*, *Chinese Scientific Journal of Hearing and Speech Rehabilitation*, *Journal of Audiology and Speech Pathology* (*Chinese*), and *Chinese Journal of Electronics*. He is the vice director of Chinese Academy of Audiological Rehabilitation (CAAR) and a senior member of IEEE.

Jing CHEN received her Ph.D. degree in signal and information processing at Peking University in 2009. Following this, she began working as a postdoctoral research associate at the Hearing Lab of Department of Experimental Psychology, University of Cambridge. Her research interests include auditory psychoacoustics, auditory computational models, and speech enhancement for the hearing impaired.