Time-compression thresholds for Mandarin sentences in normal-hearing and cochlear implant listeners[1#]

Qinglin Meng[1,2 *], Xianren Wang[3], Yuexin Cai[4], Fanhui Kong[3], Alexa Nadezhda Buck[2], Guangzheng Yu[1 *], Nengheng Zheng[5 *], Jan W. H. Schnupp[2 *]

1. Acoustics Lab of School of Physics and Optoelectronics and State Key Laboratory of Subtropical Building Science, South China University of Technology, China
2. Hearing Research Group, Department of Biomedical Sciences, City University of Hong Kong, Hong Kong SAR of China
3. Department of Otorhinolaryngology, the First Affiliated Hospital, Sun Yat-Sen University and Institute of Otorhinolaryngology, Sun Yat-Sen University, China
4. Department of Otolaryngology, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University and Department of Hearing and Speech Science, Xin Hua College of Sun Yat-Sen University, China
5. College of Information Engineering, Shenzhen University, China

* Corresponding Authors: Q. Meng (mengqinglin@scut.edu.cn), G. Yu (scgzyu@scut.edu.cn), N. Zheng (nhzheng@szu.edu.cn), J. Schnupp (wschnupp@cityu.edu.hk)

Running Title: Fast Mandarin Sentence Perception

[1]Abbreviations: time-compression thresholds (TCTs), Mandarin speech perception (MSP), Mandarin hearing in noise test (MHINT)

Abstract:

Faster speech may facilitate more efficient communication, but if speech is too fast it becomes unintelligible. The maximum speeds at which Mandarin words were intelligible in a sentence context were quantified for normal hearing (NH) and cochlear implant (CI) listeners by measuring time-compression thresholds (TCTs) in an adaptive staircase procedure. In Experiment 1, both original and CI-vocoded time-compressed speech from the Mandarin speech perception (MSP) and Mandarin hearing in noise test (MHINT) corpora was presented to 10 NH subjects over headphones. In Experiment 2, original time-compressed speech was presented to 10 CI subjects and another 10 NH subjects through a loudspeaker in a soundproof room. Sentences were time-compressed without changing their spectral profile, and were presented up to three times within a single trial. At the end of each trial, the number of correctly identified words in the sentence was scored. A 50%-word recognition threshold was tracked in the psychophysical procedure. The observed median TCTs were very similar for MSP and MHINT speech. For NH listeners, median TCTs were around 16.7 syllables/s for normal speech, and 11.8 and 8.6 syllables/s respectively for 8 and 4 channel tone-carrier vocoded speech. For CI listeners, TCTs were only around 6.8 syllables/s. Speech reception thresholds in noise were also measured in Experiment 2, and were found to be strongly correlated with TCTs for CI listeners. In conclusion, the Mandarin sentence TCTs were around 16.7 syllables/s for most NH subjects, but rarely faster than 10.0 syllables/s for CI listeners, which quantitatively illustrated upper limits of fast speech information processing with CIs.

Highlights:

1. Young normal hearing subjects' Mandarin TCTs were around 16.7 syllables/s.
2. Vocoded TCTs were around 11.8 (8-channel) and 8.6 (4-channel) syllables/s.
3. In comparison, TCTs with cochlear implants were only around 6.8 syllables/s.
4. TCTs were strongly correlated with SRTs for cochlear implant subjects.

Keywords: cochlear implant; speech rate; time compression; speech reception threshold

1. Introduction

Speech rates can be quite variable in daily communication, and the effect of varying speech rate, or of related acoustic parameters such as phoneme or temporal gap durations, on speech perception, have been the subject of numerous previous investigations (Bosker, 2017; Janse, 2003; Klumpp et al., 1961; Koch et al., 2016; Liberman et al., 1967; Shen et al., 2017; Thomas et al., 1970). Previous research has documented the normal variation in the speed of communication (Garvey, 1953), investigated the brain's speech decoding mechanisms at variable speeds (Pefkou et al., 2017), or quantified perceptual differences between individuals with different acoustic hearing conditions for clinical applications (Versfeld et al., 2002).

Many previous experiments used fixed-rate speech material to measure speech recognition scores. Psychoacoustic tests presented with fixed parameters may suffer from ceiling or floor effects, i.e., many subjects have very high or very low scores, and the real limits of their ability are not fully resolved (Nilsson et al., 1994). By measuring the 50% speech reception threshold (SRT), that is, the point at which half the words or syllables in sentence are correctly identified, tracks a point which is arguably "too difficult" for full speech comprehension, but it avoids ceiling effects and can therefore be measured and compared accurately across different conditions. Adaptive staircase procedures can be used to measure these thresholds efficiently. Thus, tracking the 50% correct threshold is a useful measure of acoustic constraints, which has precedents in the literature (Hagerman and Kinnefors, 1995; Brand and Kollmeier, 2002; Meng, et al., 2016).

In recent years, several studies have used adaptive methods to measure SRTs as a function of speech rate, also known as the time-compression threshold (TCT) (Kocinski et al., 2016; Schlueter et al., 2015; Versfeld et al., 2002). In these studies, the duration of speech stimuli was adaptively compressed by time-compression algorithms, and adaptive psychophysical procedures were used to measure TCTs in acoustic hearing subjects for Dutch (Versfeld and Dreschler 2002), German (Schlueter et al., 2015), and Polish (Kocinski et al., 2016) subjects. The tonal Mandarin Chinese (*Putonghua*) is a significantly different language from the western languages. Mandarin sentences are composed of mostly monosyllabic and disyllabic words. Almost all Mandarin syllables are structured with an "initial consonant + final vowel" and a lexical tone. (A small number of Mandarin syllables can also have a consonant coda, either [n] or [ŋ], or in some northern accents [ɚ]). To the best of our knowledge, the measurement of TCTs with Mandarin sentences in normal hearing (NH) listeners, has not previously been documented in the literature.

Furthermore, it is potentially useful to know whether, and if so, how much, TCTs differ between NH and cochlear implant (CI) listeners. Modern CIs can provide very useful open-set speech recognition abilities to most users in a quiet environment and at normal, conversational speed. However, because of much-reduced transmission of spectral and temporal fine structure in these artificial electric stimuli, CI users cannot make full use of the redundant information in original speech signals, making it harder for them to compensate for losses of information if the signals are degraded or greatly accelerated. Fu and colleagues have carried out several measurements on the effects of speaking rate on speech intelligibility for both English (Fu et al., 2001; Ji et al., 2013) and Mandarin-speaking (Li et al., 2011; Su et al., 2016) CI users. Their general conclusion was that, compared with NH listeners, CI users have significantly more difficulty in understanding fast speech. In these experiments, they used fixed speaking rates, rather than adaptive procedures, and TCTs have not been quantified.

The TCTs reported in current study measured the 50% recognition rate by tracking the point where 50% of the words are correctly recognized as reported, in keeping with the adaptive procedures of some previous SRT studies (Hagerman and Kinnefors, 1995; Brand and Kollmeier, 2002; Meng, et al., 2016). Some researchers, such as Versfeld and Dreschler (2002) dealt with that problem by measuring 50% sentence recognition rate for tracking TCTs at which all words in the test sentence were correctly identified. That is a sensible thing to do if the objective is to measure the parameters that demonstrably permit highly efficient verbal communication. However, in order to score 50% of sentences correctly, the proportion of words that needs to be recognized correctly is substantially greater than 50%, and at these higher word recognition rates, linguistically skilled subjects are likely to be able in many cases to guess some of the missing words from context. Thresholds measured in this way therefore measure a combination of high-level cognitive linguistic performance, as well as acoustic or auditory factors. Here we were interested in comparing normally hearing cohorts with CI patients who vary greatly in the quality and quantity of their linguistic experience prior to testing, and sought to focus on quantifying the effects of deficits in relatively low-level auditory processing provoked by CI processing or simulated CI processing through vocoding on speech reception. We therefore wanted to minimize the advantages that NH listeners may have had purely from substantially greater linguistic experience. Unlike Versfeld and Dreschler (2002), we therefore decided to track a 50% correct syllable recognition threshold, rather than the threshold at which "50% of sentences are recognized 100% correctly". Arguably, 50% syllable or word recognition TCTs are "too demanding", in that they identify TCTs that are quite a bit faster than the maximum speed at which effective communication is possible. But tracking higher % correct syllable recognition scores can be problematic. If a subject's acoustic performance was good enough to correctly identify a larger proportion, perhaps 80% or so, of the syllables, and if their linguistic expertise in skill in the language tested is high would very likely be able to guess the remaining 20% correctly from context. This makes word recognition thresholds much higher than 50% impossible to track in a manner that is independent of linguistic competence. Another difference between our method and that used by Versfeld and Dreschler (2002) is that we presented each test sentence up to three times to each subject. This reduces the chance that cognitive factors such as brief lapses of attention or memory load are significant limiting factors in the measured performance.

The first aim of this study was to adaptively measure young adult NH native Mandarin speakers' TCTs, quantified as the number of syllables per second at which the listener recognizes 50% of the syllables correctly, in an "ideal" acoustic environment, i.e. presented diotically over high-quality headphones in the absence of background noise. According to Versfeld and Dreschler (2002), TCT in young NH subjects was about 12.5 syllables/s. Schlueter et al. (2015) used the same procedure as Versfeld and Dreschler (2002) and got a median TCT around 11.8 syllables/s in young NH subjects. Because we used a 50% correct threshold rather than the more difficult 100% threshold of Versfeld and Dreschler (2002) and Schlueter et al. (2015), we expected to find thresholds higher than the 12.5 syllables/s reported by these authors. The second aim was to measure the TCTs with Mandarin sentences in CI users. According to Su *et al*. (2016), most CI subjects obtained high scores (>50%) even under a "fast" condition with a 5.67 syllables/s as mean rate. Therefore, we expected TCTs of at least some CI users to be faster than 5.67 syllables/s. The third aim was to investigate whether TCTs and SRTs in noise are correlated for CI users, as one might expect given that a suboptimal transmission of speech cues through a CI could easily lead to a reduced performance in both tasks. If TCTs and SRTs correlate highly, as Versfeld and Dreschler (2002) observed, then TCT measurement might be useful for audiological testing due to its time efficiency.

To achieve these aims, two experiments were carried out in this study. Experiment 1 tested TCTs in NH subjects listening to original (i.e. non-vocoded) and vocoder-CI-simulated sentences with varying degrees of time compression via headphones. Experiment 2 tested TCTs and SRTs in NH and CI subjects listening to original sentences via a loudspeaker. The main contribution of this work is, for the first time, to publish 50% TCT data with CI listeners and with NH listeners who are native Mandarin speakers.

2. Experiment 1: Time-compression Thresholds for Original and Vocoded Speech in Normal-hearing Subjects

2.1 Subjects

Ten young NH native Mandarin-speaking students (N1-10; 20-26 years old) from South China University of Technology (SCUT) were recruited. Their pure-tone thresholds were 20 dB HL or better in both ears at octave frequencies between 125 and 8000 Hz (same for Experiment 2). None of them had ever heard or read the speech material used in the experiments prior to participation. All subjects were compensated for their time. Participation was voluntary, and all procedures were approved by the local institution's ethical review board.

2.2 Speech material

We used sentence material from two published speech databases for Mandarin as it is spoken in mainland China: the Mandarin speech perception (MSP) corpus (Fu et al., 2011) and the Mandarin hearing in noise test (MHINT) corpus (Wong et al., 2007). Both corpora were developed for evaluation of speech perception ability of hearing-impaired people, with careful consideration of the phonetic balance in the sets. MSP consists of 10 lists of 10 sentences each. Each MSP sentence contains 7 monosyllabic words. Here we combined neighboring lists pairwise (i.e., 1&2, 3&4, 5&6, 7&8, 9&10) to obtain five lists of 20 sentences each. MHINT comprises 12 lists, each with 20 sentences. Each MHINT sentence includes 10 monosyllabic words. MSP and MHINT were recorded by a single female speaker and a single male speaker respectively. There were typically no gaps between subsequent words, but when short, silent gaps did occur naturally between words, these were considered to be part of the sentence duration. The average speaking rate of the recorded MSP and MHINT sentences were 3.5 and 4.5 syllables/s respectively.

2.3 Algorithms: Time-scale compression and vocoder simulation

The 'synchronized overlap-add, fixed synthesis' (SOLAFS) algorithm (Henja et al., 1991) allows compressing or stretching audio signals at arbitrary rates without an accompanying change in relative spectral distribution. The MATLAB implementation of SOLAFS used here was downloaded from Dan Ellis's homepage (https://www.ee.columbia.edu/~dpwe/resources/matlab/solafs-matlab.html) in August 2017. This method for time-compressing is very similar to the PSOLA method used in the previous TCT study by Versfeld and Dreschler (2002). It produces very natural sounding time-compressed speech, and none of our normally hearing subjects reported hearing any distortions in the time compressed material. Readers interested in further background on the time compression algorithms may wish to consult Dorran David's thesis: "Audio time-scale modification." at https://arrow.dit.ie/cgi/viewcontent.cgi?article=1002&context=engdoc.

Sine-carrier vocoders were used for CI simulation with NH subjects. In the vocoder processing, 8 or 4 sixth-order Butterworth band-pass filters were implemented to split the speech signal in the frequency range of 80 to 7999 Hz into 8 or 4 bandpass signals. The cutoff frequencies of the filters were defined by equally dividing the basilar membrane according to the Greenwood function (Greenwood, 1990). They were [80.0, 214.9, 424.0, 748.0, 1250.1, 2028.2, 3234.1, 5102.9, 7999] Hz for the 8ch vocoders, and [80.0, 424.0, 1250.1, 3234.1, 7999] Hz for the 4ch vocoders. The temporal envelopes of the band-passed signals were extracted by full-wave rectification and an eighth-order Butterworth low-pass filter with a cutoff frequency of 250 Hz. Then each envelope was multiplied by a sinusoidal signal, whose frequency was at the center of the corresponding channel and its initial phase was a random value. Finally, the modulated signals from all sub-bands were summed to synthesize a vocoded stimulus.

Figure 1 shows spectrograms for one sentence from each database for illustration. Figure 1 (a) and (d) show the original sentences (with rates of 3.4 and 5.1 syllables/s), while (b) and (e) show versions that are time-compressed to a rate of 17.0 syllables/s. We can see that much of the spectral information (e.g., the harmonic structure) is preserved under time compression. Note that this algorithm compresses time uniformly, and thus shortens vowels or consonants by the same scale. In contrast, natural speakers speaking at a fast rate tend to shorten vowels more than consonants. However, Schlueter et al., (2014) formally compared uniform versus non-uniform time-compression algorithms with NH subjects and concluded that uniform compression provides clear advantages for speech perception studies. Figure 1 (c), (f), and (g) show three vocoded speech examples, which were time-compressed to 11.5 and 8.5 syllables/s prior to either 8ch or 4ch vocoder processing for the MSP and MHINT sentences, or 4ch vocoder processing for the MHINT sentence. These examples illustrate the characteristic of coarse spectral resolution following vocoder processing which has been widely used for the simulation of similar spectral resolution as a model for CI stimulus perception (Dorman et al., 1997). The presentation rates of 17.0 syllables/s for the normal and 11.5 and 8.5 syllables/s for the vocoded examples shown in Fig. 1 were chosen because they are close to the median TCT rates observed in the NH experiments for these conditions.

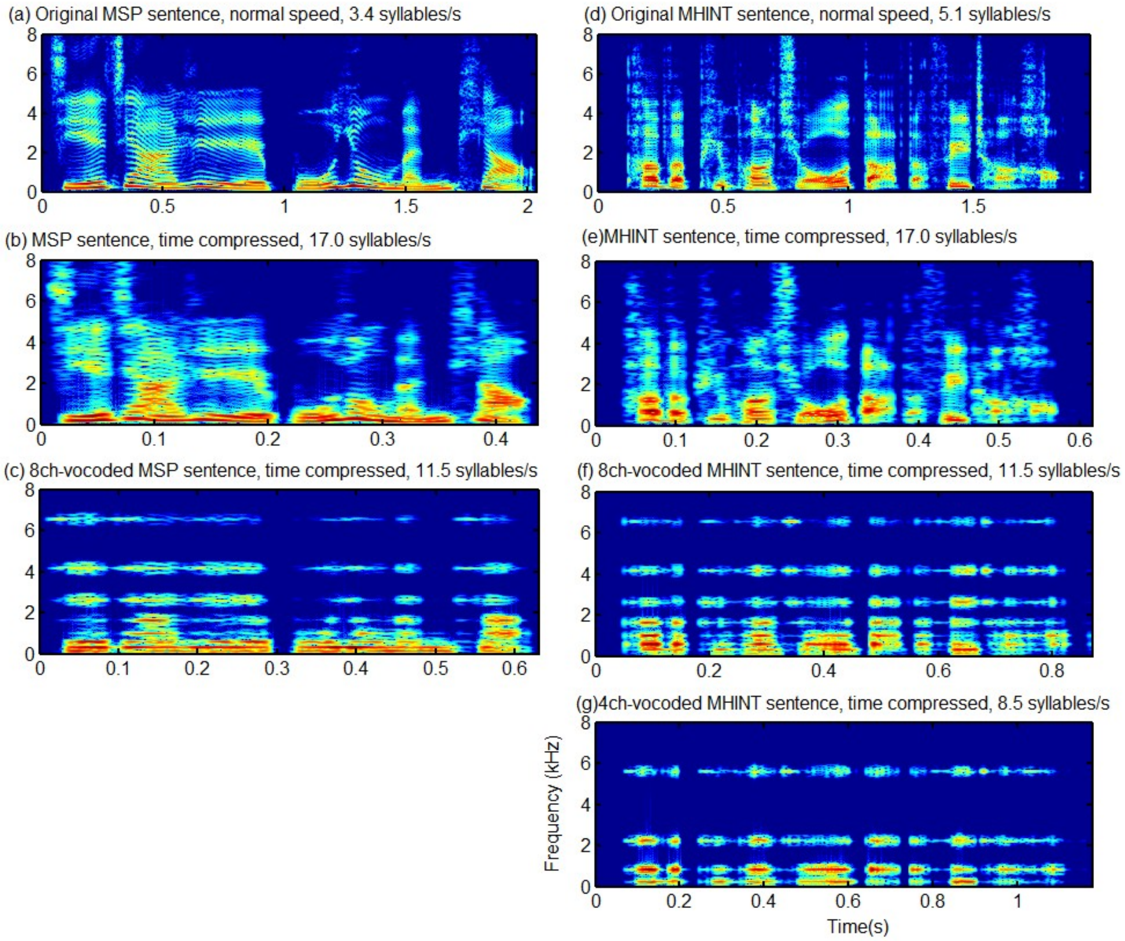

Fig. 1 Spectrograms of sentences. (a), (d). One original sentence from each of MSP and MHINT, respectively. The content of (a) is 节假日不用门票 (/jié jià rì bú yòng mén piào/, meaning ‘on holidays no need to buy tickets’). The content of (d) is 她的乒乓球打得非常好 (/tā de pīng pāng qiú dă de fēi cháng hăo/, meaning ‘she is very good at playing table tennis’). (b), (e). Compressed sounds (17.0 syllables/s) of the sentences in (a) and (d) respectively. (c), (f). Eight-channel sine-wave vocoded sounds of the compressed signals (11.5 syllables/s) of the sentences in (a) and (d) respectively. (g). Four-channel sine-wave vocoded sound of the compressed signal (8.5 syllables/s) of the sentence in (d). Their audio files are provided in supplementary materials (1)-(7). (Color online)

2.4 Procedure

Using original speech signals from the database or their corresponding vocoder simulated signals we measured TCTs using an adaptive staircase procedure. For the vocoder simulation (See Section 2.3), eight-channel (8ch) and four-channel (4ch) sine-wave carrier vocoders were used to process the MHINT speech, and only eight-channel (8ch) sine-wave carrier vocoder was used to process the MSP speech as it consists of fewer (only five) lists. This yielded the stimulus material for five blocks for each subject, i.e., original-MSP, 8ch-vocoder-MSP, original-MHINT, 8ch-vocoder-MHINT, and 4ch-vocoder-MHINT. In each original block, two sentence lists were used and the mean TCT between the two was recorded as the final result. In each vocoder block, three lists were used. The average TCT between the last two lists was recorded as the final result; the first list was considered as training. For each subject, the order of the testing materials and the order of blocks using MSP or MHINT material were all randomized. In the vocoder conditions, the compressed sounds were processed by vocoders before presentation as described below. All sounds were sampled at 16 kHz and presented diotically at a comfortable level (approximately 70 dBA) through an audio interface (Focusrite Scarlett 2i4) and headphones (Sennheiser HD650) in a soundproof room. Sound levels were measured for the MSP-SSN levels using a sound level meter coupled to the headphones via an artificial ear.

In each adaptive staircase procedure, 20 sentences from one list were used in a random order. For each subject, none of the sentences was presented in more than one trial. The selected sentences were time compressed according to a variable ratio R which was initially set to be 0.5 for the first sentence and then adaptively adjusted for each trial. Each sentence was compressed to $1+R$ multiplied by the original speaking rate (in syllables/s) using the uniform time-scaling algorithm introduced in Section 2.3.

After each presentation, the subjects were instructed to repeat as much of the last sentence as possible. The experimenter recorded for each syllable whether it was correctly repeated on a custom graphics user interface. When the subjects were not able to repeat every syllable in the sentence correctly, they were given up to two further presentations and attempts with the same sentence, so the final correct score for each sentence represented the performance after up to three attempts. A sentence was scored as "intelligible" when more than half of the syllables were repeated correctly. If a sentence was rated intelligible by this criterion, R increased for the next sentence, otherwise R decreased. R changed by a factor of 1.5 until the second reversal occurred, followed by a factor of 1.25 until the sixth reversal, and a factor of 1.1 thereafter until the end of this list. The TCT was estimated by taking the arithmetic mean of the speaking rate (in syllables/s) of the last eight out of the twenty sentences presented. No other feedback was given to subjects during the tests (other than that sentences which scored 100% on the first attempt were not repeated).

3. Experiment 2: Time-compression Thresholds and Speech-reception Thresholds in CI users and NH listeners

3.1 Subjects

Ten Mandarin-speaking CI users (ages 10-40 years) participated in this experiment. They were recruited through contacts at ENT departments at hospitals in Guangzhou, China. Further details about the CI users are shown in Table 1. Another 10 NH native Mandarin-speaking students (N11-20; ages 20-24 years) from South China University of Technology were recruited. No subject had previously heard the presented stimuli or read the text of the experimental materials before participation. As we wanted our CI test population to include samples across the diverse patient population, we did not define narrow, formal exclusion criteria. The only inclusion criterions for CI subjects were an absence of functional residual acoustic hearing, that they had been fitted with a device according to current clinical standards for treatment of their severe to profound hearing loss, and that they self-reported an ability to communicate effectively in a quiet environment by only using their CI device(s). Participation was compensated and all subjects gave informed consent in accordance with the local institution's review board.

TABLE 1. CI user demographic information, hearing history, and device information

Subject	Gender	Age(yr)	CI Experience (yr)	CI Processor	Etiology
C1	F	21	17	Right: Cochlear Nucleus 5	Congenital
C2	M	24	15	Left: Cochlear Freedom	Drug-induced
C18	M	25	21	Right: Cochlear Nucleus 5	Congenital
C19	F	40	1.5	Both: Cochlear Freedom	Sudden deafness
C20	M	10	8	Right: Cochlear Freedom	Congenital
C21	F	34	7	Right: Cochlear CP900	Drug-induced
C22	M	37	1.25	Right: Nurotron NSP60B	Sudden deafness
C23	F	29	3	Right: Nurotron NSP60B	Sudden deafness
C25	F	38	6	Right: Cochlear CP802	Sudden deafness
C26	M	31	1.5	Right: Cochlear Nucleus 6	Drug-induced

3.2 Speech stimuli and tasks

The MSP and MHINT databases used for experiment 1 also provided the stimuli for the experimental blocks with the CI group and the NH control group in Experiment 2. In a first block with the CI group, we confirmed that the subjects did not have abnormally low speech recognition abilities by measuring their word recognition scores for one list from each database. In the second block with the CI patients, three lists from each database were selected for adaptive measurements of TCTs, using the same methodology as for Experiment 1, and we also measured their SRTs in speech-shaped noise (SSN) and in babble noise for comparison.

The SSN was generated by imposing the long-term average amplitude spectrum of all sentences in corresponding databases onto white noise using Fourier methods:

$$Y[m]=\left|FFT\left(x[n]\right)\right|e^{j2\pi\varphi[n]} \text{ and} \tag{1}$$

$$N[n]=R\left(IFFT\left(Y\right)\right) \tag{2}$$

where *x* was a vector generated by concatenating all speech signals in corresponding database, *φ* was a vector whose values were randomly and uniformly distributed in the interval of [0, 1] and length was the same as x, Y was a long-term average spectrum of the SSN to be generated, N was the SNN, m and n were sampling point, and FFT and $IFFT$ represented the fast Fourier transform and inverse fast Fourier transform.

The MSP babble noise was generated by summation of sentences No. 1, No. 21, time-reversed No. 41, No. 61, and time reversed No. 81 from MSP database. The MHINT babble noise was generated by summation of sentence No. 1, No. 8, time-reversed No. 16, No. 24, and time-reversed No. 32 of the MHINT practice lists. The sentences for babble noise generation were selected pseudo-randomly from corresponding database in order to form a babble with the same long-term average spectral-temporal pattern as the target sentences. The babble speakers are identical to the target speaker, so no speaker difference cues could be used for target streaming. This babble noise condition was expected to be more challenging than the SSN condition, because five talker babble noise was not sparse enough to allow time-glimpsing of target speech, and it might introduce more informational masking. The spectrograms of babble noise and SSN for MHINT are illustrated in Fig.2.

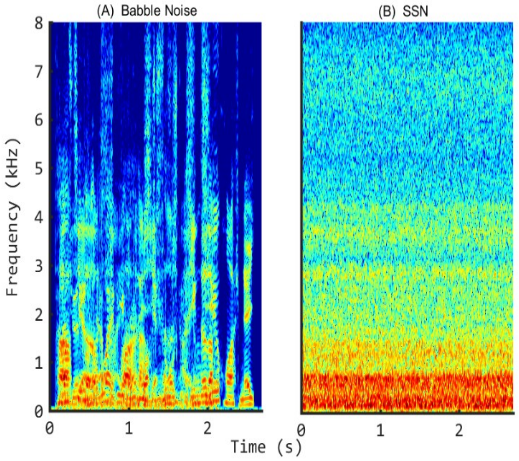

Fig.2 Spectrogram demos of babble noise (A) and SSN (B) for MHINT database. (Color online)

For each subject, the order of presentation for lists and sentences within lists was randomized for each block. All sounds were sampled at 16 kHz and presented at a comfortable level (approximately 70 dBA), passed through an audio interface (Focusrite Scarlett 2i4) and a loudspeaker (Genelec 8010A) in a soundproof room. Sound levels were measured using a sound level meter positioned roughly at the location of the center of the listener's head.

NH control subjects were tested in the same manner as that used in block 2 of the CI subjects. No other feedback was given to subjects during the tests (other than that sentences which scored 100% on the first attempt were not repeated).

3.3 Algorithm: CI signal processing strategy

During this experiment, all CI subjects' processors were set to their habitual, day-to-day algorithms. In particular, these strategies included the advanced combination encoders (ACE) for Cochlear devices (n=8) (Vandali et al., 2000) and the advanced peak selection (APS) strategy for the Nurotron devices (n=2) (Zeng et al., 2015). ACE and APS are interleaved sampling *n*-of-*m* strategies, which extract temporal envelopes from outputs of *m* bandpass filters and sequentially select *n* channels having the largest energy to stimulate the nerves (Meng et al., 2017; Zeng et al., 2008). None of the CI subjects reported using noise reduction algorithms in their default settings.

In figure 3 we show the electrodograms of the same original sound stimuli shown in Fig 1. Fig 3 (a) shows the electrodogram for the MSP sentence, Fig 3. (c) for the MHINT sentence, at their original speeds. Figure 3. (b) and (d) show corresponding electrodograms for the compressed sentences at rate of 7.0 syllables/s. This value was chosen as it reflects the median TCT results measured across CI subjects. The electrograms shown here are for illustration and were generated according to the mapping from one subject using a Cochlear device. Given that each subject used their own devices and settings, the precise stimulus patterns received by each participant will be subject to some degree of individual variation.

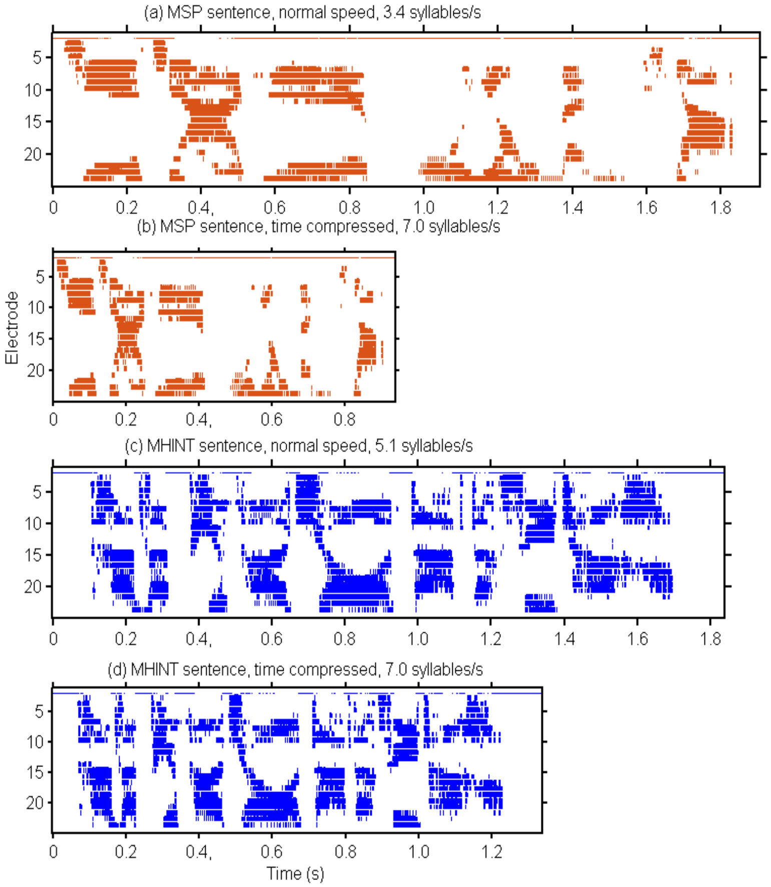

Fig.3 Electrodogram examples. (a), (c). MSP and MHINT sentences (same as those used in Fig. 1) at original speeds, in electrodogram form. (b), (d). Time compressed versions (7.0 syllables/s) of the sentences in (a) and (c) respectively. Resynthesized audio signals based on these examples, using an electrodogram to acoustic sound vocoder (Meng et al., 2018), are provided in supplementary materials (8)-(11). (Color online)

3.4 Psychophysics procedure

In the first block, the original sentence signals from selected lists were played unaltered. The number of correct syllables as a percentage of each list was taken as the final score.

In the second block, TCTs were measured by the same procedure as described in Section 2.4 and SRTs were also measured by using an adaptive procedure (Meng et al., 2016). In the SRT procedure, 20 sentences from one list were presented in a random order. The SNR for each trial was adaptively adjusted by changing the level of target speech with background noise unchanged. A one-down one-up adaptive method was used to adjust the SNR. A sentence was scored as "intelligible" when more than half of its syllables were repeated correctly. The step size before the second reversal was 8 dB, followed by a 4-dB step size before the fourth reversal, and a 2-dB step size until the end of the current list. The arithmetic mean of the SNRs for the last eight sentences was recorded as the SRT. For both blocks, the subjects were instructed to repeat as much of the sentence as possible. Within each trial, the sentence could be presented one to three times based on the response from the subjects. For each subject, no sentence was used for more than one trial.

3.5 Statistical Methods

For TCT and SRT results from both experiments, a Wilcoxon signed rank test was used to compare within-subject conditions; a Wilcoxon rank sum test was used to compare between-subject conditions; a Holm-Bonferroni correction was used for multi-pair comparison; and a linear correlation analysis was used to observe the correlation between TCTs and SRTs.

4. Results

4.1 Experiment 1 (TCT in NH and Simulated CI cohorts)

The results of experiment 1 are shown in Fig. 4 as a boxplot. Among these 10 NH subjects (N1-10), subject N3 had much higher TCTs than the other subjects under most conditions (see the light purple line which goes across the two outlier points).

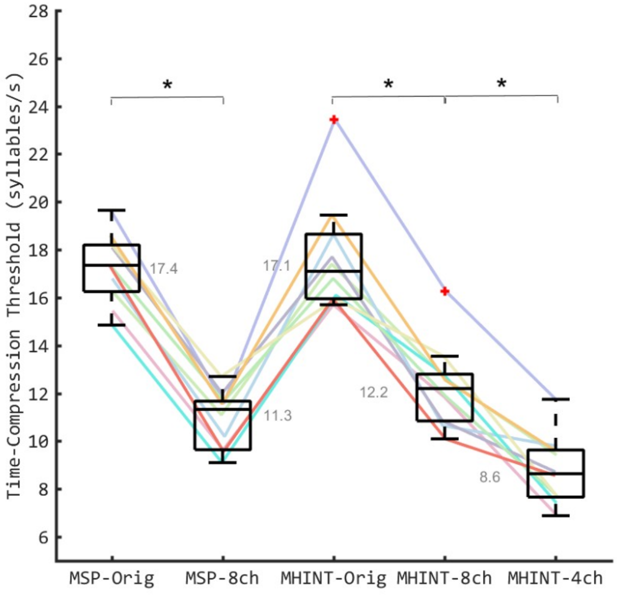

Fig. 4 Time-compression thresholds (TCTs) measured with 10 NH subjects in Experiment 1. Boxplots summarize the data over the subjects for a given stimulus condition. Left: original (i.e., non-vocoded) and 8ch sine-wave vocoded results for MSP sentences. Right: original and 8ch and 4ch sine-wave vocoded results for MHINT sentences. The significant differences between different conditions within each database are illustrated by the asterisks (*p < 0.05; Wilcoxon signed rank test). The boxplots are plotted on top of colored lines showing the performance of individual subjects. For color code, see legend of Fig. 5

Using the MSP sentences, the median TCTs and interquartile ranges (IQR) were 17.4 (2.0) syllables/s for the original, non-vocoded stimuli, and 11.3 (2.0) syllables/s in 8ch vocoder simulation. The corresponding results for the MHINT sentence material were similar, with a median TCT of 17.1 (2.7) syllables/s with the original sentences, 12.2 (2.0) syllables/s in 8ch vocoder simulation, and 8.6 (2.0) syllables/s in 4ch vocoder simulation set. (The speech rates of the example spectrograms shown in Fig. 1 b, c, e, f, and g were matched to these median TCT thresholds).

In all cases, the median TCTs for the 8ch vocoder simulations were significantly lower than those for the original sentence material ($p < 0.01$, $n = 10$) and similarly for MHINT, the median TCTs for the 4ch vocoder simulation were significantly lower still ($p < 0.01$, $n = 10$). The observed difference between median TCTs for the original MSP and MHINT material were very small and not statistically significant ($p = 0.56$, $n = 10$). Meanwhile, the observed difference in median TCT for 8ch vocoder simulations from the MHINT and the MSP sentence sets was small (0.9 syllables/s) but just reached statistical significance ($p < 0.05$, $n = 10$). The average median for 8ch vocoder was 11.8 syllables/s.

The super-performing subject N3 completed the MSP test first followed by the MHINT test. His TCTs for MSP were 20.0 and 11.6 syllables/s under the original and 8ch-vocoder condition respectively, and for MHINT 23.5, 16.3, and 11.8 syllables/s under the original, 8ch-vocoder, and 4ch-vocoder conditions respectively. This shows that his TCTs for 8ch-vocoder and 4ch-vocoder conditions are comparable to the median performance of other subjects under the original and 8ch-vocoder conditions respectively. To further verify the validity of these results of N3, two more MHINT sentence lists, unfamiliar to the subject, with fixed rates of 23 and 25 syllables/s were used to test the word recognition rate. N3 obtained 52 and 25 % respectively, but 5 other NH people (including the first author) attempted this task and could not repeat a single word at such speeds. Three super-fast sentence signals which are either vocoded or not but intelligible for N3 are provided in supplementary materials (12)-(14).

4.2 Experiment 2 (TCT and SRT in CI and NH cohorts)

The word recognition results of the CI subjects in the first block are shown in Figure 5. The median scores, and interquartile ranges (IQR) were 93.6 (10.7) % correct for MSP and 90.0 (27.0) % for MHINT. The median MHINT score was significantly lower than the mean MSP score (*p* < 0.05, *n* = 10 subjects, Wilcoxon signed rank test). All CI subjects scored above 50%, demonstrating they could effectively use speech communication, as self-reported before the experiment. The word recognition results of the NH control subjects in the first block are not shown in the figure, because they all got perfect scores with only two syllable mistakes in total.

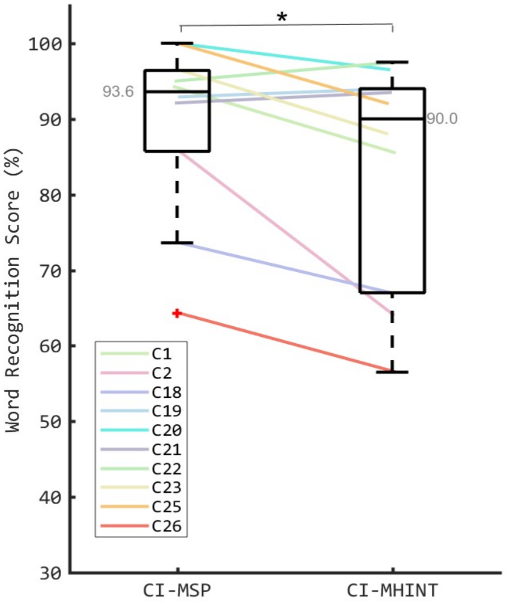

Fig.5 Word recognition scores measured with 10 CI subjects using two databases in Experiment 2. Colored solid lines represent individual subjects, and boxplots summarize the data over the subjects for a given stimulus condition. The significant differences between databases within each subject group are illustrated by the asterisk (*p < 0.05; Wilcoxon signed rank test).

The adaptive results of the second block are shown in Figure 6 (TCTs) and 7 (SRTs).

The TCTs for CI-MSP were median 6.6 (IQR 1.3); for CI-MHINT: 6.9 (1.9); for NH-MSP: 16.6 (3.0); for NH-MHINT: 16.2 (1.1) syllables/s. (We selected one sentence with a speaking rate of 7.0 syllables/s, approximately the median TCT for each database, for the illustrative electrodogram example in Fig. 3 b and d). The median TCTs with CIs were significantly lower than those in NH subjects (*p* <0.001, *n* = 20, Holm-Bonferroni corrected). There was no significant difference between databases for either NH (*p* = 0.70, *n* = 10) or CI (*p* = 0.32, *n* = 10) listeners. Note also that while the IQRs for the CI and NH subjects may appear similar in absolute terms, when expressed as a percentage of the median values, the IQRs for the CI subjects are roughly 3 times larger than those seen in the NH cohort, which is not unexpected given that CI cohorts tend to be a highly heterogeneous.

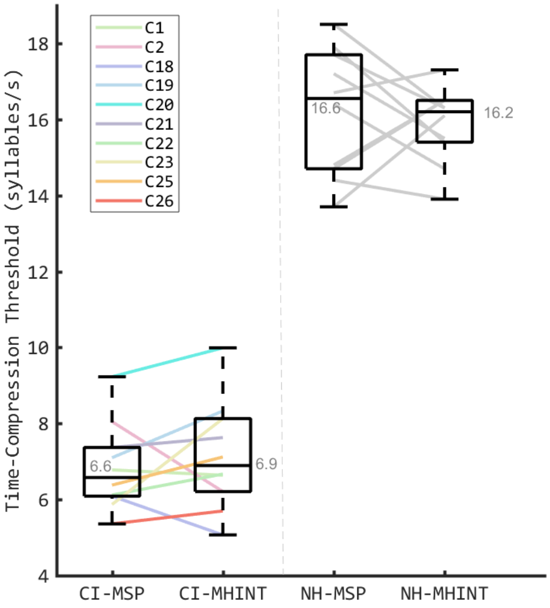

Fig. 6 Time-compression thresholds (TCTs) measured with 10 CI subjects (colored lines) and 10 NH control subjects (gray lines) using two databases in Experiment 2. Solid lines represent individual subjects, and boxplots summarize the data over the subjects for a given stimulus condition. No significant difference between databases within each subject group was observed. No significant difference between databases within either subject group was found (Wilcoxon signed rank test).

The median TCTs of the two NH groups in Exp. 1 (using headphones) and 2 (using a free-field loudspeaker) had no significant difference for either MSP or MHINT (*p* > 0.05, *n* = 20). The median TCTs (and interquartile range) for all the 20 NH subjects were 17.0 (2.9) syllables/s and 16.3 (1.5) syllables/s for MSP and MHINT respectively. The average median for both databases is 16.7 syllables/s.

Figure 7 shows that, comparing the TCTs with MHINT of the CIs in Exp. 2 and the simulated CIs in Exp. 1, the median TCTs with CIs were significantly lower than that of the 8ch-vocoded CIs (by 5.3 syllables/s, p < 0.001, n = 20) and even that of the 4ch-vocoded CIs (by 1.7 syllables/s, *p* < 0.05, *n* = 20).

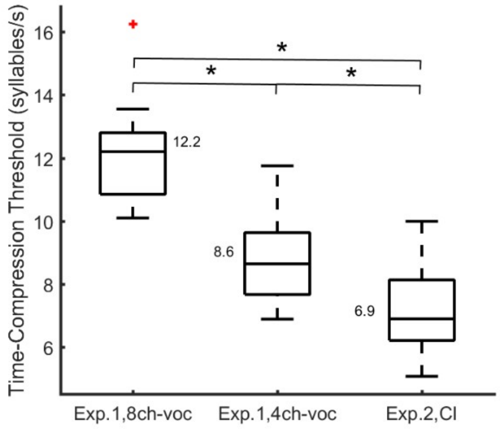

Fig. 7 Time-compression thresholds (TCTs) measured with 10 CI subjects in Experiment 2 and 10 (8ch and 4ch) simulated CI subjects in Experiment 1 with MHINT sentences. The significant differences are illustrated by the asterisks (*p < 0.05; Wilcoxon signed rank test and Wilcoxon rank sum test).

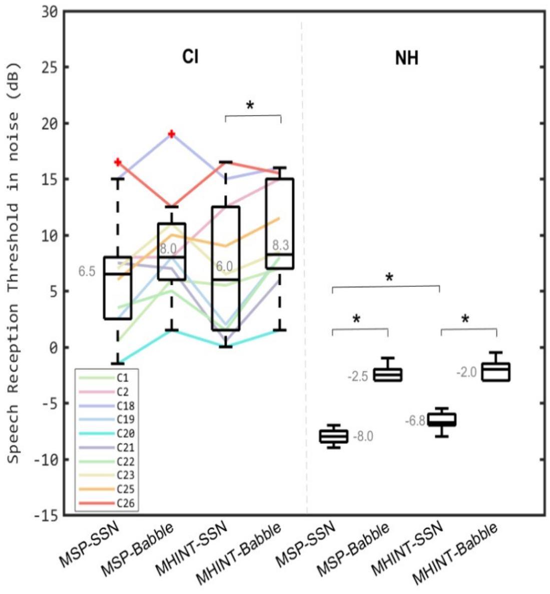

Fig. 8 Speech reception thresholds (SRTs) in speech-shaped noise (SSN) and babble noise measured with 10 CI subjects and 10 NH control subjects using two databases in Experiment 2. Solid lines represent individual subjects, and boxplots summarize the data over the subjects for a given stimulus condition. The significant differences between databases or between noise types within each subject group are illustrated by the asterisk (* p < 0.05; Wilcoxon signed rank test).

Figure 8 shows the median SRTs and interquartile range (IQR) which are 6.5 (5.5) dB for CI-MSP-SSN; 8.0 (5.0) dB for CI-MSP-Babble; 6.0 (11.5) dB for CI-MHINT-SSN; 8.3 (8.0) dB for CI-MHINT-Babble;

−8.0 (1.0) dB for NH-MSP-SSN; −2.5 (1.0) dB for NH-MSP-Babble; −6.8 (1.0) dB for NH-MHINT-SSN; and −2.0 (1.5) dB for NH-MHINT-Babble. The median SRTs of CI subjects were always significantly higher than those of NH subjects ($p < 0.001$, n = 20).

For CI listeners, database type had no significant effect ($p > 0.05$, n = 10) for either noise type; and the median SRTs for MSP-SSN and MSP-Babble were not significantly different ($p > 0.05$, n = 10). However, the median SRT for MHINT-SSN was significantly lower than that for MHINT-Babble by 2.3 dB ($p < 0.01$, n = 10).

For NH listeners, the median SRT with MSP-SSN was significantly lower than that for MHINT-SSN by 1.2 dB; the median SRTs with MSP-Babble and MHINT-Babble had no significant difference ($p > 0.05$, n = 10); median SRTs for SSN were significantly lower than those for babble noise for both databases by more than 4.8 dB ($p < 0.001$, n = 10). Thus, SRTs for SSN were always lower than the corresponding SRTs for babble, a trend as expected, and significantly so in 3 out of 4 cases.

Figure 9 A shows the correlation between speech recognition scores in quiet and TCTs for the CI subjects. Figure 9 B shows the correlations between SRTs in SSN and TCTs for the NH and the CI groups, and Figure 9 C shows the correlation between SRTs in babble noise and TCTs. For the NH group, the correlations in Figure 9 B and C were not significant ($p > 0.2$). For the CI group, all three comparisons shown in Figure 9 showed significant correlations (Speech recognition in quiet vs TCT: $r^2 = 0.424$, $p = 0.042$; SRT-SSN vs TCT: $r^2 = 0.597$, $p = 0.009$; SRT-Babble vs TCT: $r^2 = 0.672$, $p = 0.004$;).

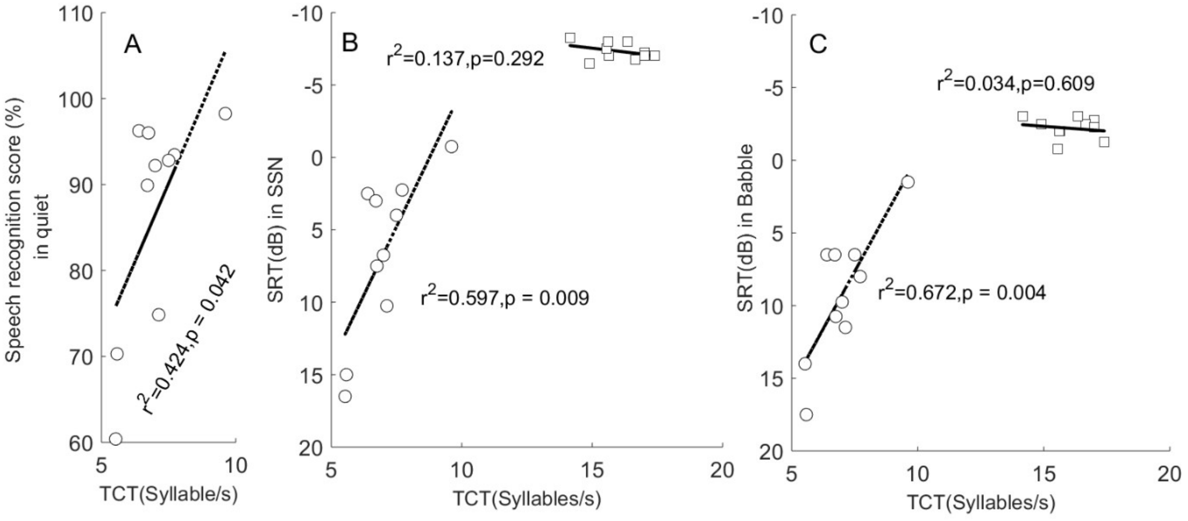

Fig.9 Correlations between TCTs and speech recognition scores in quiet (A), TCTs and SRTs in SSN (B), and TCTs and SRTs in babble noise (C) for the CI group (circles) and NH group (squares). Linear regressions within group are shown by lines.

5. Discussion

Human brains perceive and encode speech in a complex and powerful way, which involves both acoustic cue extraction and a hierarchy of successive analyses and processing strategies which are aided by topdown, anticipatory influences (Ding et al., 2016; Moore, 2012; Schnupp et al., 2011). CI processors, which mimic aspects of the peripheral auditory system, provide the patients with good perceptual ability for normal speech in a quiet environment, but they cannot efficiently represent degraded speech or speech variability because of their coarse temporal-spectral-intensity resolution at the electrode interface as well as potentially regressive or underdeveloped central systems. People with NH can easily handle fast speech, but the TCTs for Mandarin had not been reported previously for NH and also for CI subjects, although several experiments had demonstrated that CI users experienced difficulties when listening to spoken Mandarin at fast speeds (Li et al., 2011; Su et al., 2016). In this study, we measured the upper speech speed limits (i.e., TCTs) for both NH and CI listeners, mainly in the peripheral acoustic level, using a 50% threshold adaptive staircase procedure.

The median TCT of the 20 NH subjects was around 16.7 syllables/s. This tells us that the fastest speed of an "intelligible" Mandarin sentence for young NH native Mandarin-speaking listeners is about 4 times of the original speed of the sentences. However, note that for the meaning of entire sentences to be understood, correct identification of half the words on average is insufficient, so these results should not be interpreted to mean that spoken material can be presented at 4 times normal speed without a substantial drop in overall comprehension. Note also that the 16.7 syllables/s result is much higher than previous TCT results in NH subjects speaking nontonal languages, e.g., 12.5 syllables/s with Dutch in Versfeld and Dreschler (2002) and 11.8 syllables/s with German in Schlueter et al., (2015), but the reasons for this difference are likely to be methodological. Versfeld and Dreschler (2002) and Schlueter et al., (2015) used 50% sentence recognition thresholds (where 100% of the words in a sentence had to be repeated correctly) and a single repeat. In contrast, we tracked a 50% correct syllable recognition threshold which did not require subjects to be able to repeat the entire sentence correctly, and up to three repeats, as discussed in the introduction. Our methodological choices were motivated by the desire to track a speech rate threshold which reflects limitations of sensory reception which cannot be compensated for by the use of linguistic knowledge to "fill in the blanks". These choices seemed particularly appropriate for testing CI patients, who are a highly variable cohort, both with respect to the quality of the auditory input they receive and in the amount of linguistic competence they have been able to develop, given their individual histories and experience. In this study, no training session was given to our NH subjects listening to normal speech, but training might further improve their TCT scores, as shown in NH subjects in Schlueter et al., (2015).

In comparison to the normal TCT scores with a median of 16.7 syllables/s, the median TCTs for the NH subjects in Exp.1 under conditions of 8 and 4 channel vocoders were noticeably slower, at about 11.8 and 8.6 syllables/s. In as far as vocoding mimics the effect of CI speech processors, this result implies that CI processing reduces the ability to understand speech at faster speaking rates, presumably because it fails to resolve some of the acoustic cues that contribute to the informational redundancy of speech.

The median TCT of the CI subjects in Exp.2 was about 6.8 syllables/s. In two fixed speaking rate study with CIs (Li et al., 2011; Su et al., 2016), it was reported that most of their CI subjects had high scores with a 5.67 syllables/s mean rate. They are consistent with the lower limit of TCTs with CIs found in our study. Our results also clearly demonstrate the degraded performance in CI listeners compared with NH with the median TCT for CI listeners in Experiment 2 being significantly lower than that for the 4-channel tone-carrier vocoder simulation in NH listeners from Exp.1 with a difference of 1.7 syllables/s. This provides evidence of one overestimation problem of the classical vocoder methods for CI simulation task.

As for the two databases, MHINT with longer and faster speed was found to be more difficult for CI listeners to recognize than MSP (e.g., Su et al., 2016). However, no significant difference was found in most cases of TCT measurements in current study, because the speed was adaptively changed and tracked in TCT measurements. This suggests that TCTs can be consistently measured using different speech materials (at least for MSP and MHINT). However, considering variances of the variability in human speech and presentation quality, more materials could be used to compare their effects on fast Mandarin speech perception. For instance, a recent Mandarin Matrix database (Hu et al., 2018) might be a good choice.

As for the individual variability among TCTs in the 20 NH subjects (ten in Exp. 1 and ten in Exp.2) and the ten CI subjects (in Exp. 2) for original speech, the interquartile range for each TCT test condition in both experiments was in the range from 1.1 to 3.0 syllables/s. For NH listeners, the difference may come from the cue extraction and learning rate on the natural neural responses of the super-fast speech. For CIs, another important factor is the quality of acoustic-to-neuron transmission. The supernormal N3 subject from Exp. 1 had a mean TCT of 23.5 syllables/s with two MHINT lists. This subject's TCTs with the 8ch-vocoder and 4ch-vocoder were even comparable to the mean TCTs with original and 8ch-vocoder stimuli for all other NH subjects. This subject's results suggest that there are people with the extraordinary talent of extracting information from super-fast speech but the underlying brain mechanism remains a mystery.

Furthermore, for CI users there were strong correlations between TCTs and normal speed speech recognition including both speech recognition scores in quiet and SRTs in noise (SSN and babble). These consistent correlations can be explained by the fact of the wide variance of overall speech performance abilities among subjects. As Fig.5 shows, the CI subjects' mean speech recognition scores were between 60% and 100%. However, one weak trend can be found that the correlations were stronger for more difficult tests with normal speed speech, i.e., SRT in babble > SRT in SSN > speech recognition scores in quiet. This implies that TCT measurements may indicate a speaking rate related speech perception ability which is limited by some common factors which are also important for speech-in-noise perception, for example, the coarse temporal and spectral fine structure representation. Even the audiometric application of TCT is not the purpose of current study, the results here together with the Versfeld and Dreschler (2002) suggest that TCT is promising to be used as an audiometric test which may reflect the speech perception, especially in fluctuating noise, with different hearing-impaired conditions. An advantage of TCT measurement is that a TCT test list should cost less time than a normal speed SRT test because of the time-compression processing.

6. Conclusion

1. For NH listeners, the fastest intelligible Mandarin sentence (with a 50%-syllable recognition threshold, i.e., the TCT) had a median rate around 16.7 syllables/s.
2. Both simulated and actual CI processing degrade the ability to perceive rapid speech. The fastest intelligible Mandarin sentence for 8 and 4 channel tone-carrier vocoder simulation in NH listeners was 5.5 and 8.7 syllables/s lower respectively and the fastest intelligible Mandarin sentence for actual CI listeners had a median rate 9.6 syllables/s lower when compared with normal stimuli in NH listeners in the same acoustic environments.

3. CI listeners' TCTs had a strong correlation with their normal speed speech recognition especially SRT in noise suggesting some common underlying mechanisms and a potential application for TCT in audiometry that is worth exploring.

Acknowledgments

We thank all the subjects who participated in these experiments and the three students Wang Liliang, Lin Chuxiang, and Zhao Tong who assisted with data capturing. The authors also thank Wang Qianran and Tu Wenhe for helping with CI subject recruitment. This work is jointly supported by National Natural Science Foundation of China [Grant No. 11704129, 61771320, and 11574090], the Fundamental Research Funds for the Central Universities (SCUT), State Key Laboratory of Subtropical Building Science [SCUT, Grant No. 2018ZB23], and Shenzhen Science and Innovation Funds [Grant No. JCYJ 20170302145906843].

References:

Bosker, H.R. 2017. Accounting for rate-dependent category boundary shifts in speech perception. Atten Percept Psychophys 79, 333-343.

Brand, Thomas, and Birger Kollmeier. "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests." The Journal of the Acoustical Society of America 111.6 (2002): 2801-2810.

Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. NAT NEUROSCI 19, 158-64.

Dorman, M.F., Loizou, P.C., Rainey, D. 1997. Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. J ACOUST SOC AM 102, 2403-11.

Fu, Q.J., Galvin, J.R., Wang, X. 2001. Recognition of time-distorted sentences by normal-hearing and cochlear-implant listeners. J ACOUST SOC AM 109, 379-84.

Fu, Q.J., Zhu, M., Wang, X. 2011. Development and validation of the Mandarin speech perception test. J ACOUST SOC AM 129, EL267-73.

Garvey, W.D. 1953. The intelligibility of abbreviated speech patterns. Quarterly Journal of speech 39, 296-306.

Greenwood, D.D. 1990. A cochlear frequency-position function for several species--29 years later. J ACOUST SOC AM 87, 2592-605.

Hagerman, Björn, and Catharina Kinnefors. "Efficient adaptive methods for measuring speech reception threshold in quiet and in noise." Scandinavian audiology 24.1 (1995): 71-77.

Henja, D., Musicus, B. 1991. The solafs time-scale modification algorithm. Bolt, Beranek and Newman (BBN) Technical Report.

Hu, H., Xi, X., Wong, L. L., Hochmuth, S., Warzybok, A., Kollmeier, B. 2018. Construction and evaluation of the Mandarin Chinese matrix (CMNmatrix) sentence test for the assessment of speech recognition in noise. International journal of audiology, 57(11), 838-850.

Janse, E. 2003. Production and perception of fast speech Netherlands Graduate School of Linguistics.

Ji, C., Galvin, J.R., Xu, A., Fu, Q.J. 2013. Effect of speaking rate on recognition of synthetic and natural speech by normal-hearing and cochlear implant listeners. Ear Hear 34, 313-23.

Klumpp, R.G., Webster, J.C. 1961. Intelligibility of Time-Compressed Speech. The Journal of the Acoustical Society of America 33, 265-267.

Koch, X., Janse, E. 2016. Speech rate effects on the processing of conversational speech across the adult life span. J ACOUST SOC AM 139, 1618.

Kocinski, J., Niemiec, D. 2016. Time-compressed speech intelligibility in different reverberant conditions. APPL ACOUST 113, 58-63.

Li, Y., Zhang, G., Kang, H.Y., Liu, S., Han, D., Fu, Q.J. 2011. Effects of speaking style on speech intelligibility for Mandarin-speaking cochlear implant users. J ACOUST SOC AM 129, EL242-7.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M. 1967. Perception of the speech code. PSYCHOL REV 74, 431-61.

Meng, Q., Zheng, N., Li, X. 2016. Mandarin speech-in-noise and tone recognition using vocoder simulations of the temporal limits encoder for cochlear implants. J ACOUST SOC AM 139, 301-10.

Meng, Q., Zheng, N., Li, X. 2017. Loudness Contour Can Influence Mandarin Tone Recognition: Vocoder Simulation and Cochlear Implants. IEEE Trans Neural Syst Rehabil Eng 25, 641-649.

Meng, Q., Yu, G., Wan, Y., Kong, F., Wang, X., and Zheng, N. 2018. Effects of Vocoder Processing on Speech Perception in Reverberant Classrooms. APSIPA ASC 2018, Nov., Hawaii, USA. pp 761-765.

Moore, B.C. 2012. An introduction to the psychology of hearing Brill.

Nilsson, M., Soli, S.D., Sullivan, J.A. 1994. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. J ACOUST SOC AM 95, 1085-99.

Pefkou, M., Arnal, L.H., Fontolan, L., Giraud, A. 2017. θ-Band and β-Band Neural Activity Reflects Independent Syllable Tracking and Comprehension of Time-Compressed Speech. The Journal of Neuroscience 37, 7930-7938.

Schlueter, Anne, Lemke, U., Kollmeier, B., & Holube, I. "Intelligibility of time-compressed speech: The effect of uniform versus non-uniform time-compression algorithms." The Journal of the Acoustical Society of America 135.3 (2014): 1541-1555.

Schlueter, A., Brand, T., Lemke, U., Nitzschner, S., Kollmeier, B., Holube, I. 2015. Speech perception at positive signal-to-noise ratios using adaptive adjustment of time compression. J ACOUST SOC AM 138, 3320-31.

Schnupp, J., Nelken, I., King, A. 2011. Auditory neuroscience: Making sense of sound MIT press.

Shen, Y., Pearson, D.V. 2017. Recognition of synthesized vowel sequences in steady-state and sinusoidally amplitude-modulated noises. J ACOUST SOC AM 141, 1835.

Su, Q., Galvin, J.J., Zhang, G., Li, Y., Fu, Q.J. 2016. Effects of Within-Talker Variability on Speech Intelligibility in Mandarin-Speaking Adult and Pediatric Cochlear Implant Patients. TRENDS HEAR 20.

Thomas, I.B., Hill, P.B., Carroll, F.S., Garcia, B. 1970. Temporal order in the perception of vowels. J ACOUST SOC AM 48, 1010-3.

Vandali, A.E., Whitford, L.A., Plant, K.L., Clark, G.M. 2000. Speech perception as a function of electrical stimulation rate: using the Nucleus 24 cochlear implant system. Ear Hear 21, 608-24.

Versfeld, N.J., Dreschler, W.A. 2002. The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners. The Journal of the Acoustical Society of America 111, 401-408.

Wong, L.L., Soli, S.D., Liu, S., Han, N., Huang, M.W. 2007. Development of the Mandarin Hearing in Noise Test (MHINT). Ear Hear 28, 70S-74S.

Zeng, F.G., Rebscher, S., Harrison, W., Sun, X., Feng, H. 2008. Cochlear implants: system design, integration, and evaluation. IEEE Rev Biomed Eng 1, 115-42.

Zeng, F.G., Rebscher, S.J., Fu, Q.J., Chen, H., Sun, X., Yin, L., Ping, L., Feng, H., Yang, S., Gong, S., Yang, B., Kang, H.Y., Gao, N., Chi, F. 2015. Development and evaluation of the Nurotron 26-electrode cochlear implant system. Hear Res 322, 188-99.