Speech Communication 83 (2016) 94–103

Contents lists available at ScienceDirect

Speech Communication

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

Frequency importance function of the speech intelligibility index for Mandarin Chinese

Jing Chen, Qiang Huang, Xihong Wu*

Department of Machine Intelligence, Speech and Hearing Research Center, and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, 100871, China

ARTICLE INFO

Article history:
Received 10 August 2015
Revised 22 July 2016
Accepted 26 July 2016
Available online 27 July 2016

Keywords:
Speech intelligibility index
Frequency importance function
Mandarin Chinese
Speaker gender

ABSTRACT

The speech intelligibility index (SII) is a widely used objective method of predicting speech intelligibility, in which the frequency importance function (FIF) is a key component. The FIF characterizes the relative contribution of different frequency bands to speech recognition. In this work, FIFs for Mandarin Chinese were derived for monosyllabic words spoken by male and female speakers. These words were phoneme balanced and selected from the word lists of a national standard, which have been used for measuring the articulation index in China since 1995. A pilot experiment was conducted to determine suitable signal-to-noise ratios (SNR) for measuring speech intelligibility. The main experiment was conducted to derive the FIFs using 288 test conditions (4 SNRs × 36 filtering conditions × 2 speaker genders). The noise was speech-spectrum shaped and it was generated separately for the male and female speech materials. The results show that, using 1/3 octave analysis bands: (1) The FIF averaged across genders has a peak in the frequency range between 1000 and 2500 Hz, which is consistent with the FIF for English monosyllabic words; (2) The frequency bands centered at 160, 1600, and 2000 Hz are slightly more important for Mandarin Chinese than for English; (3) Male speech is more intelligible than female speech, and the band centered at 160 Hz is more important for female than male speech. The FIF differences between Mandarin and English and the effect of speaker gender are analyzed and discussed.

1. Introduction

1.1. Speech intelligibility index

The Articulation Index (AI) or its revised version, the Speech Intelligibility Index (SII), is a useful tool for predicting speech intelligibility under a variety of adverse listening conditions caused by, for example, background noise, filtering and reverberation, and it has been applied clinically in the fitting of hearing aids. The AI/SII predict speech intelligibility from the intensities of the speech and noise received by the ear, both as a function of frequency (French and Steinberg, 1947). Essentially, the predictions are based on the proportion of the speech spectrum that is audible, with each frequency band weighted according to the typical contribution of that band (i.e. its importance) to intelligibility. French and Steinberg (1947), Fletcher and Galt (1950), and later Kryter (1962) developed the AI method. In 1980s, researchers started to re-examine the AI calculation scheme (Pavlovic, 1987; Pavlovic et al., 1986; Pavlovic and Studebaker, 1984; Studebaker et al., 1987), and their work led to a new index, the SII, accepted as ANSI-S3.5 (ANSI, 1997).

The calculation of the SII can be expressed by the equation:

$$SII = \sum_{i=1}^{n} I_i A_i \tag{1}$$

where n is the number of frequency bands, and I_i and A_i are the values of the importance function, and the audibility function at the ith band, respectively. The audibility function is derived from the signal-to-noise ratio (SNR) within each frequency band, which represents the extent to which the speech is audible. The frequency-band importance function (FIF) is a series of weighting factors with values from 0.0 to 1.0, representing the relative importance of different frequency bands to intelligibility (ANSI, 1997). The sum of the weighting factors is 1, and their distribution across frequency bands depends on the speech material.

The SII can be used to predict speech intelligibility via a transfer function originally recommended by Fletcher and Galt (1950):

$$S = (1 - 10^{-AP/Q})^N \tag{2}$$

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

* Correspondence Author. Fax: 86 10 6275 9989.
E-mail address: wxh@cis.pku.edu.cn (X. Wu).

http://dx.doi.org/10.1016/j.specom.2016.07.009
0167-6393/

talker and listener, and Q and N are fitting constants that depending on the characteristics of the speech (e.g. nonsense syllables versus meaningful sentences). Q is a correction factor "to compensate for changes in proficiency"; N represents "the number of independent sounds in a test item" or "a constant that controls the shape of the line S" (Studebaker and Sherbecoe, 1991).

1.2. Frequency importance function

The FIF characterizes the relative contribution of different frequency bands to intelligibility. Previous studies have shown that the FIF depends on the speech material. The initial FIF was based on nonsense syllables, including the functions of French and Steinberg (1947), and Fletcher and Galt (1950). The only importance function used in the ANSI-S3.5 (1969) standard is a function for nonsense syllables. The rationale for using nonsense syllables was to ensure that speech intelligibility was determined mainly by the acoustical characteristics of the stimuli, rather than by cognitive or other factors. For example, a meaningful syllable might be recognized by guessing based on prior knowledge, even when some phonemes of the syllable were not heard.

Black (1959) developed an FIF for phonetically balanced (PB) words, as he considered that the FIF for nonsense syllables was not sufficiently representative of everyday speech. The results revealed greater weighting of low-frequency bands for PB words than for nonsense syllables. Studebaker et al. (1987) estimated the FIF for continuous discourse. The results revealed greater importance of low-frequency bands than for PB words. It was concluded that as context in the test materials increases, the low-frequency bands make a progressively greater relative contribution (Theodore and Donald, 1992). In ANSI S3.5, six FIFs are provided for six types of speech, including nonsense syllables, PB words, and short messages (ANSI, 1997).

FIF have not been established for Mandarin Chinese. As a tonal language, Mandarin Chinese has four typical tones that carry the lexical meaning of each word. Acoustically, the mean fundamental frequency (F0) and the F0 contour shapes determine tone perception (Chao, 1968). For example, varying the F0 glide in the syllable /bi/ from flat, to rising, to falling and rising, and to falling, changes the meaning of the syllable from "compel" to "nose" to "compare" and to "close". Morphemes of Mandarin Chinese are monosyllabic and monosyllables are combined to form polysyllabic words. Chinese syllables have more voiceless consonants and fewer voiced consonants than English syllables (Lin and Wang, 1992). All these characteristics of Mandarin Chinese suggest its FIF may be different from that for English. There is a national standard for calculating the AI of Mandarin Chinese (GB-T15485, 1995), but the FIF in this standard was derived from early work in English studies (Zhang and Ma, 1965), and it was set to 0.05 for each of the twenty frequency bands.

The present study was aimed at deriving an FIF to be used in calculating the SII for Mandarin Chinese, based on a PB monosyllabic word list. The words were selected from the national standard of China, "Acoustics-Speech articulation testing method" (GB-T15508, 1995). A method similar to that described by Studebaker and Sherbecoe (1991) was adopted to estimate the FIF.

2. Methods

2.1. Participants

Six young university students (22–24 years old, 3 females) with audiograms in the normal range (< 25 dB HL at 125, 250, 500, 1000, 2000, 4000, and 8000 Hz) and with less than a 15 dB difference in thresholds between the two ears at all test frequencies) participated in this study. Their first language was Mandarin Chinese and all of them were raised or educated in Beijing, Tianjin or Hebei province, indicating they could understand standard Mandarin Chinese well. All participants were compensated for their participation.

Table 1
The pinyin form of the fifty monosyllabic word used in the experiments. The number at the end of each syllable represents the Chinese tone.

Chuang1	Sun1	Guan1	Liang2	Qing4
Jin1	Bian4	Tian1	Dong4	Gong4
Zheng4	Sheng4	Dang1	Men2	Ren2
Fan4	Zhan4	Que1	Dui4	Wo3
Shuo1	You2	Xiao4	Ye4	Jia3
Tou4	Pao3	Hao4	Mei4	Lai2
Er4	Zi4	Shi2	Ci4	Xu3
Li2	Di4	Ni3	Yi2	De2
Ke1	Ge2	He1	Zhe4	She2
Dang1	Ta1	Bu4	Zhu4	Wu4

2.2. Apparatus

Participants were seated in a chair at the center of an anechoic chamber (Beijing CA Acoustics), which was 560 cm in length, 400 cm in width, and 193 cm in height. All acoustic signals were digitized at a sample rate of 16 kHz using a 24-bit Creative Sound Blaster PCI128 sound card (which had a built-in anti-aliasing filter) and audio editing software (Cooledit Pro 2.0), under the control of a computer with a Pentium IV processor. The signals were delivered to a loudspeaker (Dynaudio Acoustics, BM6 A), which was in the frontal azimuthal plane at 0° (with respect to the median plane). The loudspeaker height was 106 cm, which was approximately ear level for a seated participant with average body height. The distance between the loudspeaker and the center of the participant's head was 200 cm.

2.3. Stimuli

Fifty monosyllabic words were selected based on the word lists of the KXY1-10 articulation test (see below for details). The word lists were part of the national standard of China "Acoustics-Speech articulation testing method" (GB-T15508, 1995), and proposed by National Technical Committee on Acoustics of Standardization Administration of China. There were 10 word lists with 75 monosyllabic words in each list. These words were selected from daily conversation and phonetically balanced. Speech intelligibility was approximately the same for all lists. To select 50 monosyllabic words from these lists, the rules were followed: (1) the proportion of each consonant/vowel appearing in the output list was consistent with its proportion averaged across all Chinese syllables. For example, the proportion of consonant /b/ is 5.2% according to the Acoustic Handbook (Ma and Shen, 2004), so the number of appearances in the 50-word list for this consonant is 3 (the closest integer to $50 \times 5.2\%$); (2) the combination probability of syllable-initial consonants and syllable-final vowels (Ma and Shen, 2004) was as large as possible so that the selected syllables were the ones that occur most frequently. The final word list is shown in Table 1. There are 23 syllable-initial consonants, and they cover all initial consonants in Mandarin Chinese. The greatest number of appearances among all initial consonants is 6 for /d/, which corresponds to its frequency of appearance in the language (12%). One syllable in the list is without initial consonant (/er4/) and it corresponds to the situation of non-initial words. Thirty-one final phonemes among all 38 in the language are included in the list. The most frequent is the vowel /a/, whose frequency of appearance is the highest (12.4%). The frequency of appearance for the seven absent final phonemes