

Perception of Native English Reduced Forms in Adverse Environments by Chinese Undergraduate Students

Simpson W. L. Wong¹  · Jenny K. Y. Tsui¹ · Bonnie Wing-Yin Chow² · Vina W. H. Leung¹ · Peggy Mok³ · Kevin Kien-Hoa Chung^{4,5}

Published online: 1 April 2017
© Springer Science+Business Media New York 2017

Abstract Previous research has shown that learners of English-as-a-second-language (ESL) have difficulties in understanding connected speech spoken by native English speakers. Extending from past research limited to quiet listening condition, this study examined the perception of English connected speech presented under five adverse conditions, namely multi-talker babble noise, speech-shaped noise, factory noise, whispering and sad emotional tones. We tested a total of 64 Chinese ESL undergraduate students, using a battery of listening tasks. Results confirmed that the recognition of English native speech was more challenging for Chinese ESL learners under unfavorable listening conditions, in comparison to a noise-free listening condition. These findings carry significant implications for the importance of training and assessments on connected speech perception across various listening environments.

Keywords Reduced forms dictation · Speech-in-noise perception · Connected speech · Chinese learners of English-as-second-language · Noise masking

Introduction

Casual speech, embedded with pronunciation variations and ambiguities, poses a greater listening challenge to non-native listeners than carefully pronounced speech due to the reduction

✉ Simpson W. L. Wong
wls Wong@gmail.com

¹ Department of Psychology, The Education University of Hong Kong, Tai Po, Hong Kong

² Department of Applied Social Sciences, City University of Hong Kong, Kowloon, Hong Kong

³ Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Shatin, Hong Kong

⁴ Department of Early Childhood Education, The Education University of Hong Kong, Tai Po, Hong Kong

⁵ Department of Special Education and Counselling, The Education University of Hong Kong, Tai Po, Hong Kong

processes that occur frequently (Broersma and Scharenborg 2010). According to Ladefoged (2000), native English speakers tend to follow an ease of articulation principle in which differences between segments (consonants or vowels) are kept to a minimum. As such, connected speech consists highly of reduced forms, in which sounds are deleted or changed. Some common examples of connected speech phonological processes include assimilation (the changing of sounds to suit neighboring sounds e.g. *bad cat* -> *bag cat*), contraction (the shortening of word forms, e.g., *do not* -> *don't*) and elision (the deletion of sounds, e.g., the deletion of /h/ sound such that *could have* -> *could've*). These reduced forms pose a challenge to non-native learners, especially when their first and second languages do not share similar properties in reduced variants of pronunciation (Mitterer and Tuinman 2012). This difficulty has been demonstrated mostly in a noise-free laboratory environment (e.g., Gaskell and Snoeren 2008; Henrichsen 1984; Ito 2006; Mitterer and Tuinman 2012; Shockey 2003) and has been little considered in less optimal listening conditions such as multi-talker noise. As such, the present study aimed to further the present understanding of speech perception in non-native listeners by examining their perception of English casual speech across various adverse listening conditions. Furthermore, the present study also aimed to examine whether listening performances in these conditions can predict general listening comprehension. Such research will provide insight into the consistency of listening performances across various listening conditions, and suggest if similar or different skills are required for decoding connected speech across conditions.

Perception of English Connected Speech Under Adverse Listening Conditions

In everyday settings, listeners often face adverse listening conditions whereby auditory, acoustic or linguistic signals relevant to speech recognition are degraded or entirely lost. For instance, conversations in a bar are hard to follow when there are multiple background talkers creating competing speech signals. Additionally, speech in the form of whispering is hard to identify even in a quiet room (Ito et al. 2005). These settings cause deterioration in speech perception even in native English speakers, but are even more challenging for non-native speakers of English (Bradlow and Bent 2002; Crandell and Smaldino 1996; Van Engen and Bradlow 2007). Characterizing the influence of these unfavorable listening conditions on speech perception is essential for enhancing the ecological validity of research findings of second language connected speech perception (Vlasenko et al. 2012).

As shown in previous research, noise, emotion and whispering degrade clear speech in different manners. What is less known are two major issues: (1) the level of difficulties that various adverse listening conditions represent to the listeners, and (2) which kinds of listening conditions for reduced forms perception best predict general listening comprehension skills, after controlling for reduced forms perception in a quiet environment. These two issues are both theoretically and practically important, therefore forming the investigative foci of the present study. From an acquisition perspective, the level of difficulties will inform whether one source of degradation is more detrimental than another to connected speech perception. In other words, existing knowledge about the types of signal degradation across various types of adverse listening conditions suggest only the qualitative differences; whereas the data contributed by the present study will reveal quantitative differences. The latter information will be useful for theorists to further identify if there are additional important component skills that should be included in L2 listening comprehension models. Furthermore, to the best of our knowledge, there is only one study to date examining the amount of variances that con-

nected speech perception can contribute to general listening comprehension, among all other linguistic variables such as vocabulary knowledge (Wong et al. 2017). Such findings suggest the possibility that additional components such as the compensatory strategies employed to decode degraded connected speech under particular listening conditions may contribute additional variances to general listening comprehension, after controlling for reduced forms perception in a quiet setting. Thus, we will test a range of performances in comprehending degraded connected speech and examine their predictive values to general listening comprehension. For educators, a better understanding of the effect of different adverse conditions on connected speech will enable them to design better-informed English listening training that equips students with the listening skills for overcoming the listening challenges across various adverse listening conditions encountered in their daily lives. Below, we review the linguistic properties of various sources of degradation.

Speech Perception in Noise

Existing literature on non-native perception of native English speech under adverse listening conditions has mostly focused on additive noise, whereby energy from other sound sources is present (Bradlow and Bent 2002; Garcia Lecumberri and Cooke 2006; Shi 2009; Van Engen 2010; Van Engen and Bradlow 2007). In a noisy environment, listeners must discern between competing noise signals to identify speech targets (Van Engen 2010).

Energetic masking, a form of masking believed to occur at an auditory peripheral level, results when the acoustic cues from the target signal are rendered inaudible. Specifically, energetic masking results in the loss of cues necessary for the identification of speech segments as well as prosodic cues relevant to the identification of segment boundaries. Generally, energetic masking tends to be more effective on sounds with lower energy, such as dental fricatives which are pronounced with the tip of tongue against the teeth (Garcia Lecumberri et al. 2010). Only speech sounds that are high in energy (e.g., strong fricatives such as the /s/ or /z/ sound, formants, i.e. amplitude peaks in sound frequencies) are likely to be audible under energetic masking. A typical example of an energetic masker is speech-shaped noise, a type of spectro-temporally static noise whose shape approximates the average long-term spectrum of speech.

In contrast, informational masking refers to the interference with central or higher levels of auditory processing. Specifically, informational masking occurs when both target speech cues and noise masker are audible, hence affecting speech intelligibility by reducing the listener's ability to discern speech signals from the noise masker (Drullman and Bronkhorst 2004; Kidd et al. 2007). Brungart (2001) proposed that informational masking is most effective when the noise masker is spectro-temporally dynamic in nature (thus allowing glimpses at the target speech). On this note, the strongest informational maskers are considered to be single-talker babble and multi-talker babble, although Garcia Lecumberri et al. (2010) point out that all noise maskers have the potential to provide informational masking. The informational masking effect has been shown to be particularly effective on ESL learners (Cutler et al. 2004; Hazan and Simpson 2000; Mayo et al. 1997). As informational masking is believed to affect central processing, this particular susceptibility in ESL learners may indicate a strong reliance on central processing for speech identification. For example, in administering the Hearing-In-Noise Test (HINT) on English, Chinese and Korean native listeners, Jin and Liu (2012) reported better English sentence recognition under multi-talker babble than speech-shaped masker for English native listeners, and a reversed performance pattern in Chinese

and Korean non-native listeners (significantly higher sentence identification in speech-shaped noise than multi-talker babble).

It should be noted that this finding may not be completely attributable to differences in masking type. Spectro-temporal differences between the two types of noise may have also contributed to differences in speech identification performances. While multi-talker babble is a modulatory noise with some spectro-temporal variation in masker energy, speech-shaped noise is a stationary noise with little or no spectro-temporal variation in masker energy. Temporal modulations in noise consist of dips in spectro-temporal energy, thus allowing listeners glimpses of the target speech (Garcia Lecumberri et al. 2010). In contrast, stationary noise induces a constant acoustic energy to mask the target speech at all times. Given the difference between the two types of noise, it may be that native listeners are more efficient at extracting speech cues in the temporal gaps present in multi-talker babble than their non-native counterparts (Jin and Liu 2012; Mi et al. 2013). Chinese ESL learners have in fact been proposed to be particularly susceptible to modulatory noise, due to influence from Chinese as their first language, a language dependent on tonal perception (Jin and Liu 2012). The current study will thus aim to explore the differential effects of informational and energetic masking, as well as spectro-temporally dynamic and static masking by employing three different noise types with different properties.

Perception of Whispered Speech and Sad Emotional Speech

Whispering is a socially significant form of communication and is adopted in a variety of situations such as private communication, to avoid disturbing others, or simply as a form of playful interaction (Cirillo 2004). Because speech in the form of whispering is produced by exhalation and does not involve vocal cord vibrations, whispered speech poses another challenge to second language reduced forms perception (Fujimura and Lindqvist 1971). Several generation mechanisms were found to cause the acoustic characteristics of whispered and normal speech to be distinct from one another. Firstly, the lack of vocal cord vibrations in whispered speech means a lack of fundamental frequency (f_0) (Mansell 1973). Secondly, whispered speech is lower than normal speech by at least 20 dB in magnitude (Jovicic and Dordevic 1996). Finally, a higher formant shift has been recorded for vowel frequencies in whispered speech relative to normally voiced speech, increasing the difficulty of vowel discrimination (Ito et al. 2005; Konno et al. 1996; Tartter 1991). The detrimental effects of whispering on speech recognition have been illustrated in various studies (Cirillo 2004; Ito et al. 2005; Morris 2003). However, considerably few studies have examined the effect of whispered speech on how non-native listeners perceive whispered speech in their second language. As such, the present study aims to fill in this important research gap.

The conveyance and recognition of emotions in speech are significant to human survival in terms of collecting feedback about environmental events (Buck 1985). Studies examining the production of emotional speech have reported changes in fundamental frequency (Scherer et al. 2001; Williams and Stevens 1972), speech rate (Pell 2001; Scherer et al. 2001), pitch (Bao et al. 2007) and vowel formants (Maekawa 2004) in speech streams that carry various emotions, thus causing difficulty in speech recognition. Despite the importance of emotion in communication, existing research on emotional speech have mostly focused on the extent to which second language learners are able to identify the emotional state underlying the speech rather than the content of speech (Bao et al. 2007; Clavel et al. 2004; Nogueiras et al. 2001; Ververidis and Kotropoulos 2006). In one of the few studies that have examined the accuracy of speech identification under emotional tones, Vlasenko et al. (2012) found

a decreased word recognition accuracy of about 50% in emotive German speech using the Vera am Mittag corpus (VAM), a database that contains unscripted and authentic audio–visual speech recorded from a German talk show. In another study, Polzin and Waibel (1998) showed a mean percentage decrease of about 25% in accuracy of word identification under the Happiness, Sadness and Fear conditions, relative to a neutral condition. Both studies highlight the challenge of emotive speech perception in native listeners. The issue of emotive speech perception in non-native listeners, however, remains unclear. As such, we contribute to the understanding of this issue in the current study.

The Present Study

The bulk of research on ESL learners' perception of English connected speech has been largely conducted in a quiet or optimal listening environment, with minimal research focusing on unfavorable listening conditions. It is thus in the interest of the present study to examine the influence of such conditions on the recognition of native English connected speech. The current study will address this by systematically presenting sentences or phrases embedded with nine types of connected speech processes (assimilation, contraction, elision, flapping, glottalization, intrusion, juncture, palatalization and vowel weakening) in four main types of listening conditions: (a) noiseless, (b) noise, (c) whispered tones, and (d) emotional tones. Under the noise condition, three types of noise maskers will be employed: speech-shaped noise, factory noise and multi-talker babble. These noise maskers were selected for their different properties. In particular, speech-shaped noise is spectro-temporally static and induces energetic masking, factory noise also induces energetic masking but is spectro-temporally dynamic, while multi-talker babble is dynamic in nature and induces informational masking. Based on previous studies, it is hypothesized that Chinese ESL listeners' native English speech perception will be significantly impaired under noise (Garcia Lecumberri et al. 2010; Jin and Liu 2012; Mi et al. 2013), whispered speech (Ito et al. 2005) and emotional speech (Vlasenko et al. 2012) as compared to a noiseless condition. Additionally, we expect to replicate findings from Jin and Liu (2012) such that informational masking (multi-talker babble) causes a bigger masking effect than energetic masking (factory and speech-shaped noise). Given the belief that Chinese ESL listeners are more susceptible to modulatory noise, factory noise is expected to create a bigger masking effect than speech-shaped noise.

Lastly, we aim to examine whether listening comprehension under the above conditions would be able to predict general listening comprehension after the ability to comprehend connected speech in a noiseless condition is controlled. We propose that listening comprehension under each of the various listening conditions (Noiseless, Speech-shaped noise, Factory noise, Multi-talker babble, Whispered and Emotional tones) would be able to predict overall general listening comprehension, as speech perception in everyday listening conditions is often confounded by additive noise, the use of a lower, whispered tone, or the expression of emotion in speech.

Methods

Participants

Sixty-four Chinese undergraduate students from two universities in Hong Kong were recruited for this study, of which 14 were males, and mean age was 19.66 years ($SD = 1.21$). Both universities adopt English as the main medium-of-instruction. The participants,

recruited via university intranets and emails, comprised of students from various fields including arts, business, engineering, science and social science. According to our questionnaire, 9.38% of them have resided in English speaking countries, of which the mean duration of residence abroad is 1.61 (SD = 8.43) months.

All participants in this study were native speakers of Hong Kong Cantonese and second language (L2) learners of English. In Hong Kong, English is implemented as a core subject in the educational curriculum. Participants reported a mean formal English language acquisition onset age of 3.59 (SD = 1.94) years old, and a mean overall score of 3.05 (SD = 1.73) for the English language in the Hong Kong Diploma of Secondary Examination (HKDSE). The standard of HKDSE was benchmarked with the GCE A level by the Hong Kong Examinations and Assessment Authority (HKEAA) and Universities and Colleges Admissions Service (UCAS) in the United Kingdom (HKEAA 2013a). By allocating UCAS Tariff points to each level, the standard of level 3 to level 5* of the HKDSE was found comparable to grades E to A* of the GCE A level. Participants' mean overall English score of 3.05 in the HKDSE was thus equivalent to a grade E or the lowest pass grade in the GCE A level. This score as an indicator of general English proficiency of our participants suggested an overall acceptable English proficiency in our participants that met the minimum requirements of most undergraduate programs in higher education institutions. As revealed in a benchmarking study, this standard was equivalent to International English Language Testing (IELTS) band score of 5.58 to 5.68 (HKEAA 2013b), indicating that the participants attained a standard between those of *modest users* and *competent users* of English who could communicate effectively with native English speakers.

Procedure

Before the commencement of the study, participants were given an information sheet and a consent form for the study. Following the obtainment of their consent, participants were administered the following tasks in a fixed order. All audio materials were played from a laptop computer and were delivered through Audio-Technica ATH-SJ11 headphones. All testing sessions took place in either a research laboratory or an empty classroom, and lasted for approximately 1 hour. Following the completion of the experiment, participants received either 1 course credit or HKD 50 book coupon for participation.

Materials

Recording of Materials: Configuration and Speakers. The speech tokens for the reduced forms dictation task and speech gating task were recorded in a sound proof booth using a high quality recorder (Roland R-09HR) and digitalized at a sample rate of 44.1 kHz with a 16-bit amplitude resolution. All recorded materials were processed using the open-source audio editing software Audacity (version 2.0.2, Audacity Team 2012) for noise reduction. The speech tokens were recorded by 3 native English female speakers, two with a British Received Pronunciation (RP) accent and one with a General American (GA) accent. The two British speakers were of ages 51 and 22, and had been residing in Britain for 37 and 21 years respectively. The American speaker was of age 32, and had been residing in the United States for 28 years. Speech production speed of the two RP speakers and the GA speaker were approximately 186 (syllable per minute), 277 and 222 spm, respectively. The majority of the audio stimuli was recorded with the British (RP) speakers as British English is believed to be the most common variety of English adopted in the primary and secondary curriculum in schools. The American speaker was recruited for the production of speech embedded with

Table 1 Examples of reduced speech processes

Reduced forms	Examples	Transcription
Contraction	I <u>won't</u> call my sister	/wənt/
Juncture	Please <u>hand it</u> over	/hændɪt/
Elision	<u>Left</u> behind by the adults	/lɛf/
Vowel weakening	It is the law <u>of</u> the jungle	/əv/
Assimilation	The driver takes <u>them</u> home	/teɪksəm/
Intrusion	Is that your <u>idea of</u> a joke?	/aɪdiəʊv/
Flapping	<u>Writing</u> a letter	/laɪrɪŋ ə lɛtɚ/
Glottalization	The movie is about a <u>meat</u> eater	/miʔ/
Palatalization	It costs <u>you</u> too much	/kɔʃju/

reduced forms that are not commonly observed in British English, e.g., flapping. To ensure the intelligibility of the recorded speech stimuli, we presented all speech tokens to a volunteer American native speaker, who accurately identified all recorded sentences except one, which as a result was excluded from the study.

Reduced Forms Dictation Test

This test, adapted from Henrichsen's (1984) sandhi-variation exercise, assesses participants' abilities in identifying English words in their reduced forms. In this task, participants were presented with 41 aural sentences or phrases embedded with nine types of connected speech phonological processes, namely assimilation, contraction, elision, flapping, glottalization, intrusion, juncture, palatalization and vowel weakening (see Table 1 for examples used in the study). Following the presentation of each item, participants were instructed to type the whole item on the computer. Although this open-ended response format demands spelling and typing skills, it does not provide any clues to the listeners. To minimize the confounding effect, the spell check function was enabled to assist spelling. Participants were allowed to listen to the sentence up to two times. Scoring was based on the correct dictation of certain target features instead of the whole sentence or phrase. Omissions were counted as errors. As such, the maximum score in this test was 47. The speech materials used in this test were adapted from a variety of teaching materials, literature and studies on connected speech (Henrichsen 1984; Hewings 2007; Ito 2006; Matsuzawa 2006; Mok et al. 2011; Shockey 2003; Wang 2005).

Speech-in-Noise Comprehension Test

This test assesses participants' perception of fluent English connected speech in the presence of additive noise interference. Thirty sentences or phrases containing nine types of connected speech processes: assimilation, contraction, elision, flapping, glottalization, intrusion, juncture, palatalization and vowel weakening were presented aurally to participants. Additionally, three types of noise maskers were employed such that 10 speech items were presented in factory noise, 10 in multi-talker (32-talker) babble and 10 in speech-shaped noise. Following

the presentation of each item, participants were asked to type the entire item onto a laptop. Each item could be played twice. The maximum score in this test was 57 (Factory noise = 18; Multi-talker babble = 18; Speech-shaped noise = 21). All speech items were presented at 70 dB SPL (± 1) and masked by noise at 65 dB SPL. A signal-to-noise ratio of 5 dB was employed in the study. Speech materials were adapted from Hit Parade Listening (Kumai and Timson 2010) and English Pronunciation in Use Advanced (Hewings 2007), both of which are educational materials with a focus on reductions in spoken English. The noise samples were obtained from the Perception and Neurodynamics Lab in the Ohio State University (Narayanan 2012). The multi-talker babble and factory noise were obtained from the NOISEX-92 corpus (Varga and Steeneken 1993), while the speech-shaped noise was created by shaping white noise using the average spectra of utterances from the wall-street journal corpus (Paul and Baker 1992). The software Audacity (version 2.0.2, Audacity Team 2012) was used to mix speech and noise. Output levels of recordings were unified using the audio normalization software MP3Gain (version 1.2.5, Sawyer 2010).

Whispered Speech Comprehension Test

This test assesses participants' comprehension of native English connected speech in whispered tones. Twenty whispered speech items containing nine connected speech processes: assimilation, contraction, elision, flapping, glottalization, intrusion, juncture, palatalization and vowel weakening were presented aurally to participants, who were then instructed to type the entire item onto the computer. Each item could be played twice. The maximum score in this test was 25. All whispered speech materials were adopted from the corpus of Characterizing Individual Speakers Speech Project (CHAINS; Grimaldi and Cummins 2008). In the recording of these materials, speakers were instructed to read the entire set of stimuli in a whisper, and were made sure not to lapse into breathy voice or modal voicing. Whisper phonation is distinguished from other phonation types by a constricted glottis resulting in turbulent airflow and "a characteristic hissing quality" (Laver 1994).

Emotional Speech Comprehension Test

This test aimed to assess participants' comprehension of emotional speech. In this task, 22 sentences or phrases spoken in an emotionally intense manner and containing nine types of connected speech processes: assimilation, contraction, elision, flapping, glottalization, intrusion, juncture, palatalization and vowel weakening were presented to participants, following which they were required to type the entire item onto a laptop. Participants were allowed to listen to the same sentence up to two times. The maximum score in this task was 26. Speech tokens included in the test contain speech signals degraded to different extents due to the manifestation of emotion. As sadness produces the most changes of acoustic features (pitch, intensity and timing) among the basic emotion states (Ververidis and Kotropoulos 2006), participants were only examined on their comprehension of emotional speech in the sad condition. All stimuli were recorded by a British and an American native speaker, who were presented with the speech materials and instructed to read the sentences with commensurate emotional intensity. The speakers were asked to mimic the speaking style of Sylvester Stallone occurs in one of the scene in the movie *Rambo* that is consistent with crying. Speech

materials in this test were adapted from the MOCHA-TIMIT corpus, which was designed to include the main connected speech processes in English (Wrench and Hardcastle 2000).

General Listening Comprehension Test

This listening test aimed to assess participants' proficiency in native English speech comprehension in real life. As such, participants were required to listen to 4 brief audio clips of conversations or BBC/ CNN news reports delivered by English native speakers, and subsequently answer a total of 21 multiple-choice questions evaluating their understanding of the content. All audio clips were 1.5–2 min in length. The approximate average speed of speakers in the audio clips was at 280 spm. The listening materials and questions were adopted from a list of commercially available English listening training materials (Berwick et al. 2008; Chan 2003; Collins 2007; Qin 2003). Materials with themes closely related to everyday life such as personal relationships, culture and environment were selected for use rather than those with technical or field-specific content, in order to avoid the effect of difference in specific field knowledge on participants' performance. The purpose of including four audio clips was to minimize the bias favoring participants who have more knowledge of particular topics.

Results

The present study aimed to examine Chinese ESL learners' perception of native English connected speech in various listening conditions (noiseless, speech-shaped noise, factory noise, multi-talker babble, whispered and emotional tones). As such, the following statistical analyses were conducted to test this research question.

Mean Performance Across Listening Conditions

Firstly, the internal reliabilities as well as skewness and kurtosis values indicating normality of distribution in all tests used in the current study are presented in Table 2. Performances on reduced forms dictation in quiet, speech-shaped noise, factory noise, multi-talker babble and whispered tones appear to be normally distributed, while a non-normal distribution is observed for speech perception under sad emotional tones.

Following the violation of normality in speech perception under sad emotional tones, the Friedman test was conducted to examine the differences in reduced forms dictation percentage correct across the listening conditions: noiseless, speech-shaped noise, factory noise, multi-talker babble, emotional tones and whispered tones. Results revealed a significant main effect of listening condition on the accuracy of speech perception, $\chi^2(5) = 192.36$, $p < .001$. Post hoc analysis with Wilcoxon signed-rank test was conducted with Bonferroni correction, resulting in a significance level set at $p < 0.003$. In comparison to noiseless reduced forms dictation, a significant reduction in accuracy of speech recognition was observed under the following adverse listening conditions: sad emotional speech ($Z = -6.704$, $p < .001$), whispered speech ($Z = -5.762$, $p < .001$), multi-talker babble ($Z = -6.875$, $p < .001$), and speech-shaped noise ($Z = -3.650$, $p < .001$). No significant difference was observed in the accuracy of speech recognition between noiseless reduced forms dictation and factory noise ($Z = -.391$, $p = .696$).

Table 2 Normality distribution and internal reliabilities

Tasks	Skewness (SE)	Kurtosis (SE)	Cronbach's alpha/ KR-20
<i>Reduced forms dictation</i>			
a. Noiseless	-0.03 (0.30)	-0.88 (0.59)	.85
b. Speech-shaped Noise	0.44 (0.30)	-0.08 (0.59)	.69
c. Factory Noise	0.24 (0.30)	-0.27 (0.59)	.55
d. Multi-talker Babble	0.43 (0.30)	-0.22 (0.59)	.71
e. Whispered speech	0.14 (0.30)	-0.29 (0.59)	.71
f. Emotional speech	1.05 (0.30)	1.45 (0.59)	.82
General listening comprehension	-0.26 (0.30)	-0.39 (0.59)	.64

In comparing listening comprehension performance among the adverse listening conditions, significantly higher speech perception was observed under whispered speech ($Z = -4.628, p < .001$), factory noise ($Z = -6.658, p < .001$), and speech-shaped noise ($Z = -6.264, p < .001$), in comparison to emotional speech. No significant difference was observed in accuracy of speech recognition between emotional speech and multi-talker babble ($Z = -2.672, p < .05$). Additionally, whispered speech comprehension differed significantly from that of noise, such that it was lower under speech-shaped noise ($Z = -4.301, p < .001$) and factory noise ($Z = -6.080, p < .001$), but superior to that of multi-talker babble ($Z = -6.113, p < .001$). Finally, recognition of speech differed significantly among the three types of noise. Among the three noise maskers, results revealed significantly lower speech perception accuracy under speech-shaped ($Z = -4.616, p < .001$) and multi-talker babble ($Z = -6.787, p < .001$) noise, in comparison to factory noise. Multi-talker babble as a noise masker caused a greater impairment in speech perception in comparison to speech-shaped noise ($Z = -6.785, p < .001$). Participants' speech perception in terms of percentage accuracy across all listening conditions is illustrated in Fig. 1.

Taken together, these results illustrated a general impairment of English fluent speech perception in Chinese ESL learners under all adverse listening conditions, aside from factory noise. In the noise condition, Chinese ESL learners' recognition of English reduced forms was found to be affected most by multi-talker babble as a noise masker than speech-shaped and factory noise.

Predicting General Listening Comprehension

A bivariate correlations analysis was conducted to examine the correlations between listening comprehension under various conditions and general listening comprehension, as illustrated in Table 3. Significant moderate correlations are observed between reduced forms dictation under all conditions and general listening comprehension ($r_s = .38 - .57$).

Finally, a regression analysis was computed to identify the various listening conditions that would predict general listening comprehension in Chinese ESL learners, the results of which are illustrated in Table 4. As reduced forms dictation has been previously identified as a significant predictor of general listening comprehension (Wong et al. 2017), it was entered as an independent variable in the first step. All other listening conditions (whispered, emotional, speech-shaped noise, factory noise and multi-talker babble) were entered in the

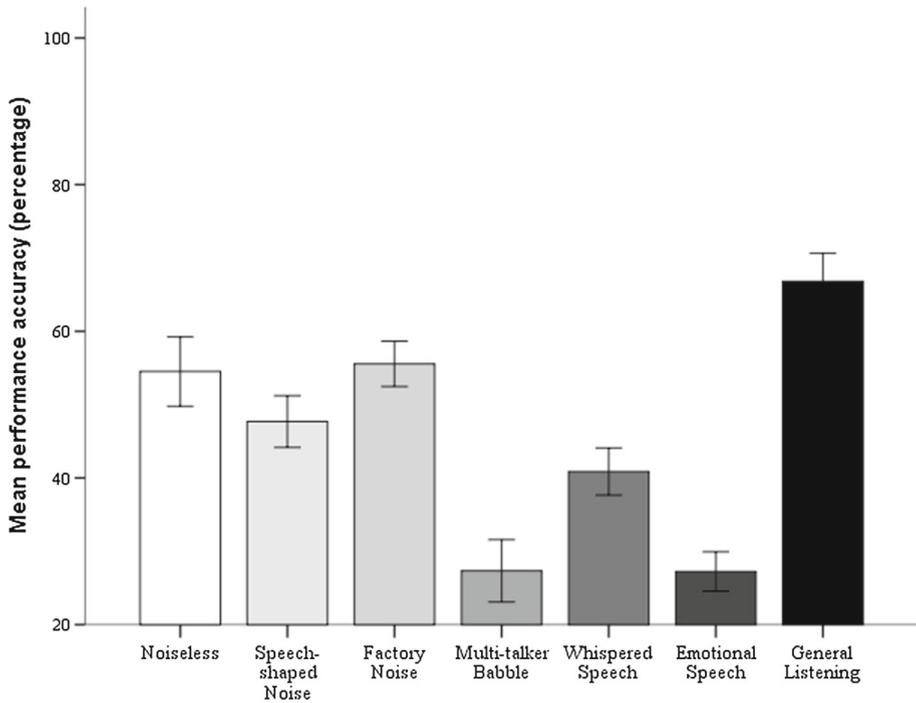


Fig. 1 Mean performance accuracy (percentage) and error bars (95% CI) of perception of target phrases across the listening conditions (N = 64)

Table 3 Zero-order Pearson’s correlations between general listening comprehension and reduced forms dictation in six listening conditions

	1	2	3	4	5	6	7
1. General listening comprehension	–						
2. Noiseless reduced forms perception	.52**	–					
3. Reduced forms perception in speech shaped noise	.49**	.71**	–				
4. Reduced forms perception in factory noise	.57**	.54**	.60**	–			
5. Reduced forms perception in multi-talker babble	.38**	.69**	.69**	.53**	–		
6. Whispered reduced forms perception	.47**	.68**	.66**	.46**	.68**	–	
7. Emotional reduced forms perception	.53**	.68**	.72**	.51**	.64**	.68**	–

** $p < .01$ (2-tailed)

second step. Reduced forms dictation in a quiet environment was identified as a significant predictor of general listening comprehension in the first step, accounting for 28% of variance in general listening comprehension. Following the addition of reduced forms dictation under other conditions, these listening tasks accounted for 44% of variance in general listening comprehension. Only factory noise was identified as a significant predictor of general listening comprehension, a finding that was inconsistent with the predicted hypothesis that all listening conditions (Noise, Whispered and Emotional tones) would be able to predict general listening after controlling for reduced forms perception in a noiseless setting.

Table 4 Hierarchical regression analysis predicting general listening comprehension

Listening conditions	β	t	p value	R	R ²	ΔR^2	ΔF
Step 1				.52	.28	.28	23.48
1. Noiseless reduced forms perception	.52	4.85	.00**				
Step 2				.67	.44	.16	3.42
1. Reduced forms perception in speech shaped noise	-.02	-0.11	.91				
2. Reduced forms perception in factory noise	.40	3.11	.00**				
3. Reduced forms perception in multi-talker babble	-.21	-1.32	.19				
4. Whispered reduced forms perception	.12	0.79	.43				
5. Emotional reduced forms perception	.25	0.16	.12				

** $p < .01$

General Discussion

The present study aimed to examine Chinese ESL learners' perception of English connected speech in various adverse listening conditions. Connected speech perception was found to be significantly impaired under all conditions, except for factory noise. These results confirmed that unfavorable listening conditions posed a challenge to reduced forms speech perception in Chinese ESL learners, and also affect listeners differently.

Listening Comprehension Across Adverse Listening Conditions

The finding of a more severe impairment of speech perception under multi-talker babble relative to speech-shaped and factory noise was in line with findings from [Jin and Liu \(2012\)](#) as well as previous proposals that informational masking (multi-talker babble) causes a greater masking effect than energetic masking (speech-shaped and factory noise) in ESL learners.

Unexpectedly, accuracy of speech perception under factory noise condition was significantly higher than that in speech-shaped noise, and was not significantly reduced from that under a noiseless condition. Given that factory noise is modulatory in nature and speech-shaped noise static in nature, this finding contradicts previous proposals that ESLs are overall less efficient at extracting information from temporal gaps in modulatory noise ([Garcia Lecumberri and Cooke 2006](#); [Hazan and Simpson 2000](#); [Jin and Liu 2012](#)). Chinese ESL learners, in particular, are believed to be particularly susceptible to modulatory noise, due to influence from Chinese as their first language, a language dependent on tonal perception ([Jin and Liu 2012](#)). One possible explanation for these unexpected findings is the relatively 'easy' speech-to-noise ratio (SNR) employed in the present study. Specifically, the present study had employed a SNR of 5dB across all noise masking conditions, whereas previous studies such as [Jin and Liu \(2012\)](#) have employed noticeably SNRs with higher noise levels (-10 to 5 dB for speech-shaped noise; -15 to 0 dB for multi-talker babble). SNR impacts directly upon the availability of low-level cues for speech perception. As such, it is likely that the 'easier' SNR employed in the present study may have caused a minimal loss of cues for speech perception, thus causing a minimal impairment of speech perception. Furthermore, the content of the noise may play a key role in speech perception in Chinese ESL learners. The notion that the content of the background noise can influence target speech perception has been illustrated in several studies, in which a greater impairment in sentence recognition was observed in English babble than Mandarin babble in both native English listeners and

Chinese non-native listeners (Van Engen 2010; Van Engen and Bradlow 2007). Van Engen (2010) suggests that this effect may be driven by a linguistic similarity between target speech signal and noise. Following on from this proposal, the content of the background noise in the present study may modulate Chinese ESL learners' susceptibility to modulatory noise such that the susceptibility is greater when the modulatory noise is comprised of speech materials than when the background noise does not contain any speech. Therefore, a minimal loss of speech recognition cues at a SNR of 5 dB, coupled with its modulatory nature and lack of linguistic similarity to speech, may have caused factory noise to induce minimal masking in the present study.

The finding of impaired whispered speech perception relative to noiseless reduced forms dictation is consistent with findings from Ito et al.'s (2005) study, in which a similar magnitude in the decrease in accuracy of sentence identification in their participants was observed. While the present study observed a 22.99% decrease in whispered speech recognition relative to normally voiced speech (reduced forms dictation), Ito et al. (2005) observed a 20% decrease in accuracy of sentence identification. This finding is especially interesting given that Ito et al. (2005) had recruited native listeners for their study, in contrast to the non-native listeners in the present study. However, caution should be exercised in the comparison of findings from both studies due to obvious differences in methodology. Additionally, it may also be worth noting that Ito et al.'s (2005) study was conducted in Japanese, using speech materials adapted from a Japanese corpus. Future studies are needed to further examine how whispering affect the comprehension of reduced forms in non-native languages.

Emotional speech was also found to be a challenge to speech recognition, as evidenced by significantly lower accuracy rates in emotional speech recognition relative to reduced forms dictation. This finding is in line with that of Vlasenko et al.'s (2012) study, who observed a similar reduction in accuracy of emotive German native speech among native German listeners. Taken together, these findings suggest that the effect of emotional speech on speech perception may also occur across languages. However, further studies are needed to replicate this effect across various types of languages to confirm this proposal.

The injection of emotion into speech has been documented to cause a variety of changes to speech, all of which may impact on speech perception. For instance, pitch has been known to vary according to emotion (Bao et al. 2007). A high glottal volume velocity, which represents the air velocity through the glottis during a vocal fold vibration, suggests joy or surprise, while a low glottal volume velocity indicates a harsher sound such as anger or disgust (Nogueiras et al. 2001). Other prosodic features important to speech recognition have also been reported to be affected, including pausing and stress (Litman et al. 2000). Interestingly, emotion has also been observed to cause other changes in speech such as a shift towards a centralized position in formant values for all vowels. However, the present study can only speculate on the changes underlying the emotional speech stimuli in the present study, as the study did not conduct linguistic analyses of speech segments. Nevertheless, we speculate that the participants in the present study may have experienced difficulty in speech processing following likely changes in pitch, owing to their first language experience. Previous literature has suggested that Chinese non-native listeners attend differently to suprasegmental information in speech perception, owing to experience with their first language (Klein et al. 2001; Lee and Nusbaum 1993) Following on from this notion, we speculate that Chinese non-native listeners may rely particularly on their ability to differentiate among tones in speech perception. As such, any changes in pitch may be particularly detrimental to speech perception due to difficulty in detecting acoustic differences among tones. Although there is a growing interest in the area of emotional speech recognition, there is still much work to be done, given the complexity of emotion. Aside from a large spectrum of emotions, each emotion can also vary in intensity.

Yet, this challenge deserves to be addressed given the necessary role of emotion in human communication.

Finally, reduced forms dictation in a quiet environment was identified as a predictor of general listening comprehension, an unsurprising finding given that everyday native speech consists highly of reduced forms. Following the addition of reduced forms dictation in other conditions, only reduced forms dictation in factory noise was found to contribute unique variance to general listening comprehension, inconsistent with our hypothesis that reduced forms dictation under each condition would be able to predict general listening comprehension. A plausible explanation for these findings is that the compensatory listening strategies employed for comprehension of factory-noise modulated connected speech is critical for decoding running speech. This strategy may also have been crucial in recovering some amount of degraded speech signals masked by modulatory factory noise in the present study. However, future research is needed to examine the underlying mechanisms of speech perception under various conditions.

Limitations and Future Directions

It should be noted that this study is not without its limitations, the first of which concerns realism. The bulk of our existing knowledge on speech perception stems from controlled laboratory dictation tasks in which a carefully recorded speech segment is presented to participants via a computer, following which participants are to identify the sentence. Communication, however, rarely occurs in such a manner. Rather, it takes place in a host of everyday situations, such as in a restaurant or on a sidewalk next to a busy road. Also future studies may wish to take the issue of realism into consideration, perhaps employing conversational tasks outside of a laboratory setting. Secondly, the present study did not employ a control group, which may have provided additional insight into differences in speech processing between native English speakers and Chinese non-native speakers of English. Nevertheless, the present study still makes a valuable contribution to the understanding of the effects of various adverse listening conditions on non-native speech perception. Thirdly, the present study provided a look into speech perception under imperfect listening conditions, but did not look at the role of cognitive-linguistic factors in these tasks. Given that everyday communication will often take place in listening conditions that are less than optimal, the importance of identifying speech-linguistic skills that can help to overcome the mismatch between such conditions and an optimal listening condition should not be overlooked.

Conclusion

Results from the present study carry practical implications. From an educational point of view, the deleterious effect of additive noise/ emotion- or whispering-modulation on students' perception of English connected speech requires more attention, especially on the speech-linguistic skills that can overcome this effect. Additionally, knowing how speaking styles (i.e. whispered or emotional speech) affect speech perception in non-native listeners can inform educators on the design of ESL classroom materials as well as their delivery.

In conclusion, this study has contributed to the overall understanding of English connected speech recognition under various listening conditions in Chinese ESL learners. Overall, speech perception under additive noise, whispered and emotional tones is significantly

impaired in Chinese ESL learners. Among the three types of noise maskers, multi-talker babble was found to cause the greatest impairment in speech perception, providing support for the proposal that ESL learners are more susceptible to informational masking than energetic masking. Findings from this study have important implications for the design and delivery of materials in an ESL classroom.

Acknowledgements We thank all the students who participated in this study. We are also grateful to Lauren Couillard, Marnie Evans, and Marianne Katherine Hewitt, who helped with the preparation of speech stimuli.

Compliance with ethical standards

Funding This study was funded by Research Grants Council (RGC) of the University Grants Committee (UGC), Hong Kong (ECS 846212) and Internal Research Grant of the Education University of Hong Kong (RG72/2015-2016).

Conflict of interest The authors declare that they have no conflict of interest.

References

- Audacity Team. (2012). Audacity (Version 2.0.2) [Computer software]. <http://audacity.sourceforge.net/>.
- Bao, H., Xu, M. X., & Zheng, T. F. (2007). *Emotion attribute projection for speaker recognition on emotional speech*. Paper presented at the 8th annual conference of the international speech communication association, Antwerp, Belgium.
- Berwick, G., Hardy-Gould, J., Southern, A., Thorne, S., & Wallwork, A. (2008). *BBC World News English Arts and Entertainment*. BBC Worldwide Ltd.
- Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America*, 112(1), 272–284. doi:10.1121/1.1487837.
- Broersma, M., & Scharenborg, O. (2010). Native and non-native listeners' perception of English consonants in different types of noise. *Speech Communication*, 52(11), 980–995. doi:10.1016/j.specom.2010.08.010.
- Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109(3), 1101–1109. doi:10.1121/1.1345696.
- Buck, R. (1985). Prime theory: An integrated view of motivation and emotion. *Psychological Review*, 92(3), 389. doi:10.1037/0033-295X.92.3.389.
- Chan, Y. H. (Ed.). (2003). *CNN Interactive English*. Taipei: Hebron Soft Ltd.
- Cirillo, J. (2004). Communication by unvoiced speech: The role of whispering. *Anais da Academia Brasileira de Ciências*, 76(2), 413–423. doi:10.1590/S0001-37652004000200034.
- Clavel, C., Vasilescu, I., Devillers, L., & Ehrette, T. (2004). *Fiction database for emotion detection in abnormal situations*. Paper presented in the 8th International Conference on Spoken Language Processing, Jeju Island, Korea.
- Collins, S. (2007). *Practical everyday English*. Barcelona: Montserrat Publishing.
- Crandell, C. C., & Smaldino, J. J. (1996). Speech perception in noise by children for whom English is a second language. *American Journal of Audiology*, 5(3), 47–51. doi:10.1044/1059-0889.0503.47.
- Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, 116(6), 3668–3678. doi:10.1121/1.1810292.
- Drullman, R., & Bronkhorst, A. W. (2004). Speech perception and talker segregation: Effects of level, pitch, and tactile support with multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, 116(5), 3090–3098. doi:10.1121/1.1802535.
- Fujimura, O., & Lindqvist, J. (1971). Sweep-tone measurements of vocal-tract characteristics. *The Journal of the Acoustical Society of America*, 49(2B), 541–558. doi:10.1121/1.1912385.
- Garcia Lecumberri, M. L., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America*, 119(4), 2445–2454. doi:10.1121/1.2180210.
- Garcia Lecumberri, M. L., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, 52(11), 864–886. doi:10.1016/j.specom.2010.08.014.

- Gaskell, M. G., & Snoeren, N. D. (2008). The impact of strong assimilation on the perception of connected speech. *Journal of Experimental Psychology: Human Perception and Performance*, 34, 1632–1647. doi:10.1037/a0011977.
- Grimaldi, M., & Cummins, F. (2008). Speaker identification using instantaneous frequencies. *IEEE Transactions on Audio, Speech and Language Processing*, 16(6), 1097–1111. doi:10.1109/TASL.2008.2001109.
- Hazan, V., & Simpson, A. (2000). The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects. *Language and Speech*, 43(3), 273–294. doi:10.1177/00238309000430030301.
- Henrichsen, L. E. (1984). Sandhi-variation: A filter of input for learners of ESL. *Language Learning*, 34(3), 103–123. doi:10.1111/j.1467-1770.1984.tb00343.x.
- Hewings, M. (2007). *English pronunciation in use advanced*. Cambridge: Cambridge University Press.
- HKEAA. (2013a). Press Release: HKDSE level 5** awarded with UCAS Tariff Points. Retrieved from http://www.hkeaa.edu.hk/DocLibrary/MainNews/PR_20121218_eng.pdf.
- HKEAA. (2013b). Press Release: Results of the benchmarking study between IELTS and HKDSE English Language Examination. Retrieved from http://www.hkeaa.edu.hk/DocLibrary/MainNews/press_20130430_eng.pdf.
- Ito, Y. (2006). Effect of reduced forms on ESL learners' input-intake process. In J. D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 51–58). Honolulu: University of Hawaii, National Foreign Language Resource Center.
- Ito, T., Takeda, K., & Itakura, F. (2005). Analysis and recognition of whispered speech. *Speech Communication*, 45(2), 139–152. doi:10.1016/j.specom.2003.10.005.
- Jin, S. H., & Liu, C. (2012). English sentence recognition in speech-shaped noise and multi-talker babble for English-, Chinese-, and Korean-native listeners. *The Journal of the Acoustical Society of America*, 132(5), EL391–EL397. doi:10.1121/1.4757730.
- Jovicic, S. T., & Dordevic, M. M. (1996). Acoustic features of whispered speech. *Acustica*, 82, S228. Retrieved from <https://getinfo.de/app/Acoustic-features-of-whispered-speech/id/BLSE%3ARN003953233>.
- Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2007). Informational masking. In W. A. Yost, A. N. Popper, & R. R. Fay (Eds.), *Auditory perception of sound sources* (pp. 143–189). New York: Springer.
- Klein, D., Zatorre, R. J., Milner, B., & Zhao, V. (2001). A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers. *Neuroimage*, 13(4), 646–653. doi:10.1006/nimg.2000.0738.
- Konno, H., Toyama, J., Shimbo, M., & Murata, K. (1996). *The effect of formant frequency and spectral tilt of unvoiced vowels on their perceived pitch and phonemic quality*. IEICE Technical Report, 39–45.
- Kumai, N., & Timson, S. (2010). *Hit parade listening* (3rd ed.). Tokyo: Macmillan Language House.
- Ladefoged, P. (2000). *A course in phonetics* (4th ed.). Fort Worth, TX: Harcourt Brace Jovanovich.
- Laver, J. (1994). *Principles of phonetics*. Cambridge: Cambridge University Press.
- Lee, L., & Nusbaum, H. C. (1993). Processing interactions between segmental and suprasegmental information in native speakers of English and Mandarin Chinese. *Perception and Psychophysics*, 53(2), 157–165. doi:10.3758/BF03211726.
- Litman, D. J., Hirschberg, J. B., & Swerts, M. (2000). Predicting automatic speech recognition performance using prosodic cues. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference* (pp. 218–225). Association for Computational Linguistics.
- Maekawa, K. (2004). Production and perception of 'paralinguistic' information. In *Speech Prosody 2004, international conference*.
- Mansell, P. (1973). An experimental investigation of articulatory reorganisation in whispered speech. *Forschungsberichte des Instituts für Phonetik und sprachliche Kommunikation der Universität München*, 2, 201–253. Retrieved from <http://www.ibrarian.net/navon/page.jsp?paperid=13127527>.
- Matsuzawa, T. (2006). Comprehension of English reduced forms by Japanese business people and the effectiveness of instruction. In J. D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 59–66). Honolulu: University of Hawaii, National Foreign Language Resource Center.
- Mayo, L. H., Florentine, M., & Buus, S. (1997). Age of second-language acquisition and perception of speech in noise. *Journal of Speech, Language, and Hearing Research*, 40(3), 686–693. doi:10.1044/jslhr.4003.686.
- Mi, L., Tao, S., Wang, W., Dong, Q., Jin, S. H., & Liu, C. (2013). English vowel identification in long-term speech-shaped noise and multi-talker babble for English and Chinese listeners. *The Journal of the Acoustical Society of America*, 133(5), EL391–EL397. doi:10.1121/1.4800191.
- Mitterer, H., & Tuinman, A. (2012). The role of native-language knowledge in the perception of casual speech in a second language. *Frontiers in Psychology*, doi:10.3389/fpsyg.2012.00249.

- Mok, P., Setter, J. & Low, E. L. (2011). The perception of word juncture characteristics in three varieties of English. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS)* (pp. 1410–1413). Hong Kong.
- Morris R. W. (2003). *Enhancement and recognition of whispered speech*. Doctoral dissertation, Georgia Institute of Technology, Atlanta.
- Narayanan, A. (2012). Sound demo for IBM-masked noise. Retrieved from <http://web.cse.ohio-state.edu/pnl/demo/IBM.html>.
- Nogueiras, A., Moreno, A., Bonafonte, A., & Mariño, J. B. (2001). *Speech emotion recognition using hidden Markov models*. Paper presented in EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, Aalborg, Denmark.
- Paul, D. B., & Baker, J. M. (1992). The Design for the Wall Street Journal-based CSR Corpus. ICSLP-92.
- Pell, M. D. (2001). Influence of emotion and focus location on prosody in matched statements and questions. *The Journal of the Acoustical Society of America*, 109(4), 1668–1680. doi:10.1121/1.1352088.
- Polzin, T. S., & Waibel, A. (1998). *Detecting emotions in speech*. Paper presented in cooperative multimodal communication: Second international conference, Tilburg, Netherlands.
- Qin, Y. Y. (Ed.). (2003). *Crazy English* (Vol. 42). Guangzhou: Renzhen Enterprise Co., Limited.
- Sawyer, G. (2010). MP3 Gain. <http://mp3gain.sourceforge.net/>.
- Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, 32(1), 76–92. doi:10.1177/0022022101032001009.
- Shi, L. F. (2009). Normal-hearing English-as-a-second-language listeners' recognition of English words in competing signals. *International Journal of Audiology*, 48(5), 260–270. doi:10.1080/14992020802607431.
- Shockey, L. (2003). *Sound patterns of spoken English*. Cornwall: Blackwell.
- Tartter, V. C. (1991). Identifiability of vowels and speakers from whispered syllables. *Perception and Psychophysics*, 49(4), 365–372. doi:10.3758/BF03205994.
- Van Engen, K. J. (2010). Similarity and familiarity: Second language sentence recognition in first-and second-language multi-talker babble. *Speech Communication*, 52(11), 943–953. doi:10.1016/j.specom.2010.05.002.
- Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native-and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America*, 121(1), 519–526. doi:10.1121/1.2400666.
- Varga, A., & Steeneken, H. (1993). Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3), 247–251. doi:10.1016/0167-6393(93)90095-3.
- Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9), 1162–1181. doi:10.1016/j.specom.2006.04.003.
- Vlasenko, B., Prylipko, D., & Wendemuth, A. (2012). *Towards robust spontaneous speech recognition with emotional speech adapted acoustic models*. Paper presented in the 35th German Conference on Artificial Intelligence, Saarbrücken, Germany.
- Wang, Y. T. (2005). *An exploration of the effects of reduced forms instruction on EFL college students' listening comprehension*. Unpublished master dissertation, National Tsing Hua University, Taiwan.
- Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, 52(4B), 1238–1250. doi:10.1121/1.1913238.
- Wong, S. W. L., Mok, P. P. K., Chung, K. K.-H., Leung, V. W. H., Bishop, D. V. M., & Chow, B.-W.-Y. (2017). Perception of native English reduced forms in Chinese learners: Its role in listening comprehension and its phonological correlates. *TESOL Quarterly*, 51(1), 7–31. doi:10.1002/tesq.273.
- Wrench, A. A., & Hardcastle, W. J. (2000). *A multichannel articulatory speech database and its application for automatic speech recognition*. Paper presented in the 5th seminar on speech production: Models and data, München, Germany.