Measuring speech-in-noise intelligibility for spontaneous speech: The effect of native and non-native accents

Jieun Song and Paul Iverson

Speech Hearing and Phonetic Sciences, University College London
jieun.song@ucl.ac.uk, p.iverson@ucl.ac.uk

ABSTRACT

Previous research has found that listeners understand talkers who speak with the same accent as themselves better than others. The aim of the current study was to investigate how speech intelligibility is modulated by this talker-listener accent interaction when native and non-native listeners hear spontaneous speech. To this end, native Southern British English listeners and native Korean listeners were tested on the recognition of read and spontaneous speech spoken with a native English accent (Standard Southern British English) and non-native English accents (Finnish and Korean-accented English). The results demonstrated that native listeners have an intelligibility benefit for their own accent over non-native accents when they listen to spontaneous speech as well as read speech. However, native Korean listeners had a trend for them to have higher intelligibility for Korean-accented speech only in the spontaneous speech condition.

Keywords: spontaneous speech, speech-in-noise recognition, non-native speech perception, accented speech

1. INTRODUCTION

The speech that we hear in everyday listening situations is far more variable than the carefully read speech that is elicited in a laboratory setting. Word forms often differ from their citation forms as they undergo casual speech processes such as assimilation (e.g., *lea[m] bacon;* [4]) in connected speech. Furthermore, conversational speech involves more extreme cases in which multiple phonemes or syllables are deleted [7]. Nonetheless, listeners are very skilled at processing such deviant word forms in the speech stream as they automatically incorporate a wide range of linguistic knowledge to decode the structure and meaning of the speech [5, 8, 12, 14, 15].

Despite the widespread agreement that investigating speech recognition in realistic environments is important, most previous research on speech perception has involved laboratory speech [10] or investigated the perception of casual speech processes occurring within words or phrases (e.g., [3, 13]).

Speech recognition in realistic communication settings is also affected by the accent of the talkers and listeners especially in noisy conditions. Specifically, listeners understand talkers who speak with the same accent as themselves more easily than others (e.g., [1, 6, 9, 16, 17, 20, 21]) However, most of the previous findings were based on intelligibility of read speech materials.

One could expect that this accent effect found in read speech extends to spontaneous speech. However, it is possible that the effect of accent is stronger when we listen to spontaneous speech, given that applying casual speech processes or making acoustic-phonetic modifications in different speaking styles can be language-specific [18]. Non-native listeners may thus have difficulty compensating for the casual speech processes of the target language, unless they have the same processes in their native language [19].

Moreover, the additional variability in spontaneous speech might in itself cause difficulties in speech recognition for non-native listeners, given that their second-language processes and representations are less developed. For example, non-native listeners may not be able to exploit semantic and contextual information as freely as native listeners to compensate for degraded or variable phonetic information (e.g., [11]).

The aim of this study is to investigate the effects of talker-listener accent differences for native and non-native listeners when they hear spontaneous speech. Native English listeners and native Korean listeners listened to read sentences and spontaneous utterances in noise, in a native English accent (Standard Southern British English) and two non-native English accents (Finnish and Korean accented English). Subjects completed a picture evaluation task that was developed to measure speech recognition difficulties for spontaneous speech while having some degree of control over the lexical and semantic content.

2. METHOD

2.1. Subjects

Twelve monolingual native speakers of Standard Southern British English (mean age: 23.4 years, age range: 18-28 years) and twelve monolingual native speakers of Korean (mean age: 28.9 years, age range: 21-35 years) participated in the experiment. All of the participants had no self-reported hearing or language disorders and were living in London at the time of testing. The Korean subjects reported that they had started learning English at school from the age of 12 years old on average (range: 8-13 years) and they had lived in England for an average of 10 months (range: 3-36 months) as adults.

2.2. Stimuli

In order to obtain spontaneous speech, we conducted the Diapix task [22]. The Diapix task was designed to elicit spontaneous speech from two speakers while they are conversing with each other to find differences between sets of pictures. This task is suitable for eliciting spontaneous speech while having some degree of control over the lexical content of the conversation. For the read speech condition, we used the Basic English Lexicon (BEL) sentences [2].

Two female speakers of each of these three accents - Standard Southern British English, Finnish-accented English and Korean-accented English - took part in the recording (age range: 18-30 years, mean age: 24.7 years). The Finnish speakers reported that they had never lived in English-speaking countries, but they had learned English since they were nine years old. The Korean speakers reported that they had lived in London for approximately eight to twelve months and learned English since they were twelve. Two speakers of each accent took park in the Diapix task together in a no-barrier, normal listening condition. The spontaneous speech obtained in the Diapix task was edited such that one stimulus comprised a section of speech produced by one talker describing a specific part of a Diapix scene. Each stimulus in the spontaneous speech condition was four to five seconds long on average.

Speech-shaped noise was generated for each talker in each speaking style using the smoothed long-term average spectrum of their recordings. The read speech materials were mixed with the noise at the signal-to-noise ratio of -8 dB and the spontaneous speech materials at the signal-to-noise ratio of -4 dB. These noise levels were chosen based on pilot work that indicated that they achieved similar intelligibility levels between the two types of speech materials.

2.3. Procedure

In this study, a new method was developed to measure speech recognition difficulties for both read sentences and utterances from spontaneous speech. Instead of repeating back what they heard as in the previous studies, listeners were presented with a picture on the screen as they listened to a stimulus and had to decide whether what they heard matched the picture. This paradigm is particularly appropriate for spontaneous speech because spontaneous speech is often relatively unstructured and it doesn’t lend itself very well to word-by-word repetition.

The stimuli were counterbalanced between listeners. Each listener listened to 195 stimuli of read speech and 126 stimuli of spontaneous speech. In each speaking style condition (i.e., read speech vs. spontaneous speech), the same number of stimuli were used for each accent. Half of the trials showed pictures that matched the speech stimuli and the other half of the trials showed random pictures that did not match the speech stimuli. The order of the stimuli was randomised.

3. RESULTS

A linear mixed-effects analysis was conducted with the accent of the talkers (English, Finnish, Korean), the accent of the listeners (English, Korean) and the speaking style (read speech, spontaneous speech) included as fixed effects, and subjects and stimuli as random effects. The dependent variable was the accuracy of responses in the speech-in-noise recognition task.

Figures 1 and 2 display the mean proportion correct for each listener and speaker group. The two-way interaction between talker accent and listener accent was significant, $\chi^2(2) = 55.0$, $p < 0.01$, as were the main effects of talker accent and listener accent, $\chi^2(2)= 28.9$, $p < 0.01$ and $\chi^2(1) = 105.2$, $p < 0.01$, respectively. English listeners showed highest recognition performance on the native accent, followed by Finnish accented English, and the lowest performance on Korean accented English. In contrast, the recognition accuracy of Korean listeners was not significantly different for the different accents. Korean listeners also had lower recognition accuracy than English listeners overall.

The main effect of speaking style was significant, $\chi^2(1)=15.1246$, $p <0.01$. The recognition accuracy for spontaneous speech was higher overall than the recognition accuracy for read speech, indicating that the two different SNR levels did not completely equalize performance levels in the two conditions.

However, the three-way interaction between speaking style, talker accent and listener accent was not significant, $p > 0.05$ and the two-way interaction between the speaking style and the listener accent was only marginally significant, $\chi^2(1)= 2.9362$, $p = 0.087$. As shown in Figure 1, English listeners were affected by accents very similarly in read speech and spontaneous speech conditions. However, Korean listeners showed some indication of a trend for them to understand Korean accented speech better than other accents in the spontaneous speech condition, although it did not reach significance with the present number of subjects.

Figure 1: Speech-in-noise recognition accuracy of English listeners. English listeners showed the highest recognition accuracy for Standard Southern British English, followed by Finnish accented English and Korean accented English.

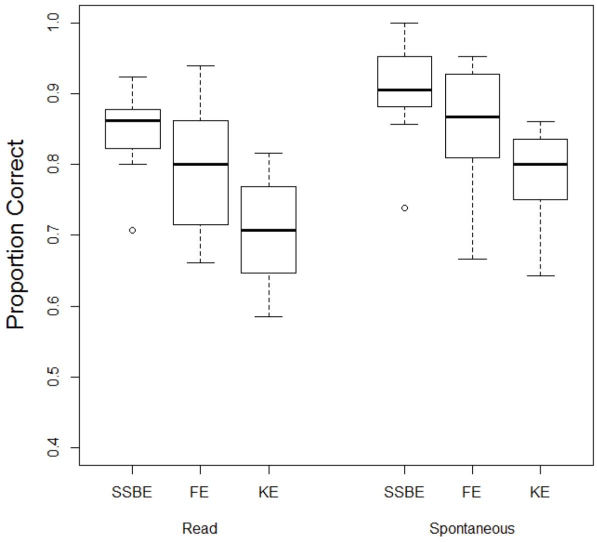
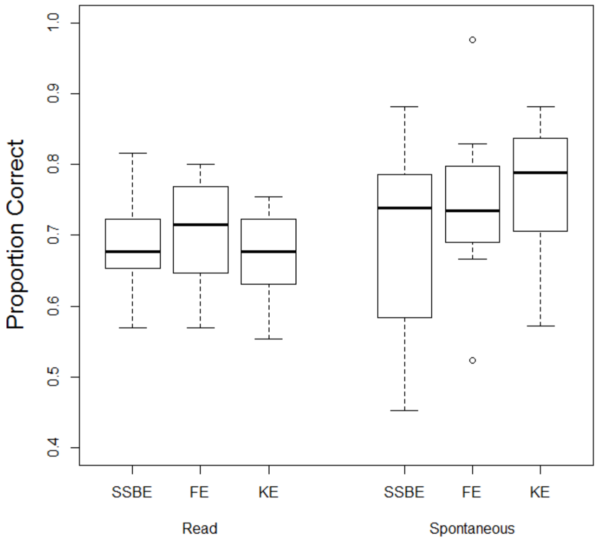


Figure 2: Speech-in-noise recognition accuracy of Korean listeners. The recognition accuracy was not significantly different among the different accent conditions.



4. DISCUSSION

In this study, we set out to investigate how speech-in-noise intelligibility for spontaneous speech is affected by the accents of the talkers and the listeners. The results supported previous work demonstrating an interaction of listener and speaker accent [1, 17, 21]. However, there was less of an effect of conversational speech than anticipated. There was no apparent interaction of accent and speech style for native listeners. For Korean listeners, there was some indication of an interaction, with stronger accent effects for conversational speech, but this didn't reach significance.

To some extent this is an encouraging result, in that it suggests that previous work using read speech likely extends to more naturalistic listening conditions. It is still likely that the reduction processes etc. in conversational speech are a complicating factor in non-native speech perception, at least to some extent. However, the potential magnitude of the accent and speaking style interactions are small enough to require more statistical power than was available here.

5. REFERENCES

[1] Bent, T., Bradlow, A. R. 2003. The interlanguage speech intelligibility benefit, *J. Acoust. Soc. Am.* 114, 1600–1610.

[2] Calandruccio, L., Smiljanic, R. 2012. New sentence recognition materials developed using a basic non-native english lexicon. *Journal of Speech, Language, and Hearing Research*. 55, 1342–1355.

[3] Ernestus, M., Baayen, H., Schreuder, R. 2002. The recognition of reduced word forms. *Brain and Language*. 81, 162–173.

[4] Gaskell, M. G., Marslen-Wilson, W. D. 1996. Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*. 22(1), 144-158.

[5] Gaskell, M. G., Marslen-Wilson, W. D. 1997. Integrating form and meaning: a distributed model of speech perception. *Lang. Cognit. Process*. 12, 613–656.

[6] Imai, S., Flege, J. E., and Walley, A. 2005. Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners. *J. Acoust. Soc. Am.* 117(2), 896–907.

[7] Johnson, K. 2004. Massive reduction in conversational American English, In: Yoneyama, K., Maekawa, K. (eds), *Spontaneous Speech: Data and Analysis*. *Proc. 1st Session of the 10th International Symposium* Tokyo, 29-54.

[8] Luce, P. A., Pisoni, D. B. 1998. Recognizing spoken words: the neighbourhood activation model. *Ear Hearing*. 19, 1–36.

[9] Major, R. C., Fitzmaurice, S. M., Bunta, F., Balasubramanian, C. 2002. The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*. 36, 173–190.

[10] Mattys, S. L., Davis, M. H., Bradlow, A. R., Scott, S. K. 2012. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*. 27(7/8), 953-978.

[11] Mayo, L. H., Florentine, M., Buus, S. 1997. Age of second-language acquisition and perception of speech in noise. *J. Speech Lang. Hear. Res.* 40, 686–693.

[12] McClelland, J. L., Elman, J. L. 1986. The TRACE model of speech perception. *Cognit. Psychol.* 18, 1–86.

[13] Mitterer, H., Ernestus, M. 2006. Listeners recover /t/s that speakers reduce: Evidence from /t/-lenition in Dutch. *Journal of Phonetics*. 34, 73–103.

[14] Norris, D. 1994. Shortlist: a connectionist model of continuous speech recognition. *Cognition*. 52, 189–234.

[15] Norris, D., and McQueen, J.M. 2008. Shortlist B: a Bayesian model of continuous speech recognition. *Psychol. Rev*. 115, 357–395.

[16] Pinet, M., Iverson, P. 2010. Talker-listener accent interactions in speech-in-noise recognition: Effects of prosodic manipulation as a function of language experience. *J. Acoust. Soc. Am*. 128, 1357–1365.

[17] Pinet, M., Iverson, P., Huckvale, M. 2011. Second-language experience and speech-in-noise recognition: Effects of talker–listener accent similarity. *J. Acoust. Soc. Am*. 130 (3), 1653-1662.

[18] Smiljanic, R., Bradlow, A. R. 2008. Stability of temporal contrasts across speaking styles in English and Croatian, *Journal of Phonetics*. 36, 91–113.

[19] Tuinman, A., Cutler, A. 2011. L1 knowledge and the perception of casual speech processes in L2. In: Wrembel, M., Kul, M., Dziubalska-Kolaczyk, K. (eds), *Achievements and perspectives in SLA of speech: New Sounds 2010. Volume I.* Frankfurt am Main: Peter Lang, 289-301.

[20] van Wijngaarden, S. J. 2001. Intelligibility of native and non-native Dutch speech, *Speech Commun.* 35, 103–113.

[21] van Wijngaarden, S. J., Steeneken, H. J. M., Houtgast, T. 2002. Quantifying the intelligibility of speech in noise for non-native listeners, *J. Acoust. Soc. Am.* 111, 1906–1916.

[22] Van Engen, K., Baese-Berk, M., Baker, R., Choi, A., Kim, M., Bradlow, A. 2010. The Wildcat corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*. 53(4), 510-540.