NATIVE AND NON-NATIVE SPEECH RECOGNITION IN NOISE: NEURAL MEASURES OF AUDITORY AND LEXICAL PROCESSING

Jieun Song, Luke Martin & Paul Iverson

Speech, Hearing & Phonetic Sciences, University College London
jieun.song@ucl.ac.uk, luke.martin.17@ucl.ac.uk, p.iverson@ucl.ac.uk

ABSTRACT

Non-native listeners have more difficulty understanding speech in noise than do native listeners, but it remains unclear to what extent these difficulties arise from early auditory or later linguistic processes. The present study investigated this using EEG when native English and Korean subjects listened to English sentences in listening conditions that varied the demands on peripheral and central processes (single-talker and unintelligible babble). Speech comprehension by non-native listeners was poorer overall and was more adversely affected by noise than that of native listeners. However, neural entrainment to the speech envelope was greater for non-native than native listeners, indicating greater attention to acoustics, but was also more affected by noise. Context-related differences in lexical processing (N400 effect) were greater for native than non-native listeners, and both groups had greater N400 effects for single-talker maskers. The results demonstrate that listeners vary in terms of how they modulate their processing under difficult conditions.

Keywords: informational masking, second-language speech processing, N400, cortical tracking of the amplitude envelope of speech, EEG

1. INTRODUCTION

Speech perception in noisy environments is more difficult for non-native listeners, and this likely arises from multiple levels of processing. For example, non-native speech perception difficulties may arise from a pre-linguistic, auditory level, because one’s first language experience alters their perceptual space [10, 11]. However, non-native listeners also have phonological representations that are less precise, as well as being less able to draw on other linguistic information, such as lexical or semantic cues (e.g., [2, 15]). Furthermore, the cognitive load of adverse listening conditions can add to the general speech recognition difficulties that listeners have with non-native speech (see [12] for a review).

Recent advances in EEG (electroencephalography) have started to allow researchers to unpick these levels of processing in relatively naturalistic speech tasks. For example, auditory processing can be assessed in terms of the entrainment of the neural signal to the speech amplitude envelope (e.g., [1, 14]). Entrainment to the speech envelope can decrease when the amount of spectral detail in the speech signal is reduced (e.g., [8, 18]) and has sometimes been shown to have a positive relationship with speech comprehension (e.g., [18]). More traditional EEG measures, such as N400, can be used to simultaneously assess lexical processing. N400 is associated with the ease of lexical access (e.g., [19]) or the ease of semantic integration of the word with its preceding context (e.g., [3]). Previous N400 studies have shown that lexical processing can be hindered by acoustic degradation or background noise [5, 16].

A previous study used these measures to compare native and non-native speech processing, under a condition in which listeners attended to a target talker played simultaneously with a single-talker distractor [21]. One could expect that non-native listeners would be worse at tracking the speech envelope of a target talker given that their perceptual and linguistic representations are less well tuned to that language (e.g., rhythm or syllable structure), but the results indicated that they actually tracked the target talker better. It is likely that this task required more focused attention for them to perform, relative to native speakers, and that this increased attention modulated their auditory processing. In contrast, native speakers were able to modulate their lexical processing for L2-accented speech and had greater N400 differences depending on the predictability of the sentences.

The aim of the present study was to investigate these issues further under various listening conditions that place differential demands on the peripheral and central processing. That is, this study examined speech recognition in the presence of single-talker and unintelligible babble maskers, using the same neural and behavioural measures as the previous study [21]. Babble would be expected to primarily involve energetic masking (EM; reduced audibility of the target speech sounds at the periphery [4]). However, single-talker distractors involve more informational masking (IM; e.g., segregation of the target speech from competing streams and the linguistic interference from the masker; [9, 20]). IM likely places greater demands on attentional and cognitive resources. In addition, spatial cues were manipulated in the present study (i.e., the spatial separation between target and masker), to examine the ability to focus attention to particular locations.

The present study recorded EEG when native English and Korean subjects listened to English sentences in the presence of a single-talker and an unintelligible babble noise. The masker was co-located with the target straight ahead or located 45° away from the target. Listeners were asked to press a button whenever they heard a semantically anomalous sentence, and the accuracy of this response was used as a measure of their speech comprehension performance.

2. METHODS

2.1. Subjects

Twenty-four monolingual native speakers of British English and Korean (12 each) participated in the EEG experiment. They were adults under the age of 35 (mean age: English - 23.8, Korean - 29.1) with no self-reported hearing or language impairments, and were all right-handed. The Korean subjects were second-language speakers of English who had started learning English at 12 years old on average and had lived in English-speaking countries for an average of 20 months as adults.

2.2. Materials

English sentences were recorded by a female native speaker of Standard Southern British English. The sentences had different levels of final-word cloze probability to allow for measurement of the N400 response [22]; high cloze probability sentences consisted of highly constraining sentence contexts followed by congruent final words (e.g., *There are three pictures hanging on the wall.*); low cloze probability sentences were neutral sentences (e.g., *There are many dirty marks on the wall*); semantically anomalous sentences consisted of highly constraining sentence contexts followed by incongruent final words (e.g., *There are three pictures hanging on the pain*).

English stories read by the same female speaker were used as the single-talker masker. The babble masker was created by combining twelve recordings from the same talker together. In the no-masker condition, the sentences were presented without any noise. To manipulate spatial cues, the targets and maskers were processed with head-related transfer functions, simulating the auditory effects of locating talkers straight ahead (0°) or 45° towards the left ear. The target signal was always presented at 0° and the noise was either at the same location or placed 45° to the left. In order to equalise intelligibility between the two conditions, a signal-to-noise-ratio (SNR) of 3dB was used for the co-located condition, whereas -7dB was used for the 45° separation condition. The noise level was higher for the spatially separated masker because this condition would otherwise be much easier.

2.3. Procedure

Subjects were instructed to pay attention to individual target sentences and ignore continuous stories (i.e., single-talker noise) or babble noise in the background. They were also asked to press a button whenever they heard a semantically anomalous sentence. The experiment consisted of 10 blocks (2 blocks * 5 conditions) with each lasting approximately 4 minutes. The order of the blocks was randomised for each subject.

2.4. EEG recording and analysis

EEG was recorded with a Biosemi Active Two System with 64 electrodes (Ag/AgCl) mounted on an elastic cap and 7 external electrodes (nose, left and right mastoids, two vertical and horizontal EOG electrodes), with a sampling rate of 2048Hz. Electrode impedances were kept between ±25kΩ.

Preprocessing of the EEG recordings was performed offline in Matlab; they were re-referenced to the average of left and right mastoids and high-pass filtered at 0.1Hz. The EEG data for the N400 analysis was also low-pass filtered at 40 Hz. Noisy channels were interpolated. In order to correct for eye artefacts, an independent component analysis (ICA) was used. All preprocessing was conducted using the Fieldtrip toolbox [17], except for filtering which used the ERPlab toolbox [16] of EEGlab [7].

2.4.1. Coherence Analysis

Cortical entrainment to the amplitude envelope of speech was measured using a stimulus reconstruction method; the Multivariate Temporal Response Functions [6] were generated in backward models that related the EEG data from each subject back to the Hilbert envelopes of the sentences that they listened to. The degree of phase coherence was computed as a function of frequency (0.5 Hz resolution) between the predicted amplitude envelopes from the EEG data via this model and the original envelopes of the stimuli, which assessed how much the EEG signals were phase-locked to the amplitude envelopes of the target speech.

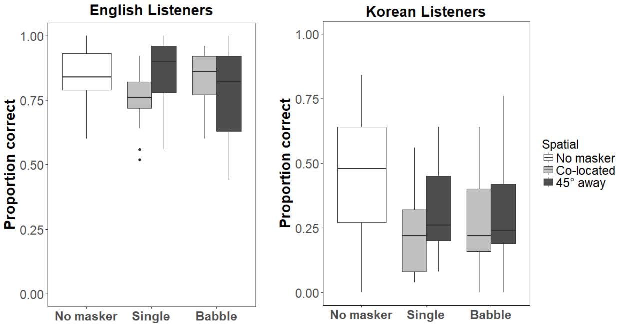
2.4.2. N400 analysis

The EEG data were segmented into epochs that were timed-locked to the onset of each final word. Trials were rejected if the amplitude was not within the range of ±150μV. N400 amplitudes were calculated by subtracting the baseline average over the 200 ms pre-stimulus interval from the post-stimulus interval and averaging the amplitude in the 300-500 ms window. N400 amplitudes were averaged across five midline electrodes (Fz, FCz, Cz, CPz & Pz).

3. RESULTS

As displayed in Fig. 1, Korean subjects were poorer at detecting anomalous sentences than were English subjects overall, and their performance was more adversely affected by noise. A logistic mixed-model analysis was conducted with listener group and masker type (i.e., no masker, single-talker masker, and babble masker) as independent variables and with button response (correct vs. incorrect) as the dependent variable. Each subject and sentence stimulus were added to the model as random intercepts. Significance of fixed effects was evaluated by comparing two nested models with and without each factor in all mixed-model analyses of this paper. The results confirmed that the interaction of listener group and masker type was significant, $\chi^2(2) = 21.25$, $p < 0.001$. Specifically, the difference in performance between the no-masker and other conditions was significantly larger for Korean listeners than for English listeners, b = - 0.80, z = -3.4, $p < 0.001$, and so was the difference between the single-talker and babble noise condition, b = - 0.71, z = -3.31, $p < 0.001$. There were also significant main effects of listener group, $\chi^2(1) = 25.57$, $p < 0.001$, and masker type, $\chi^2(2) = 9.36$, $p = 0.009$.

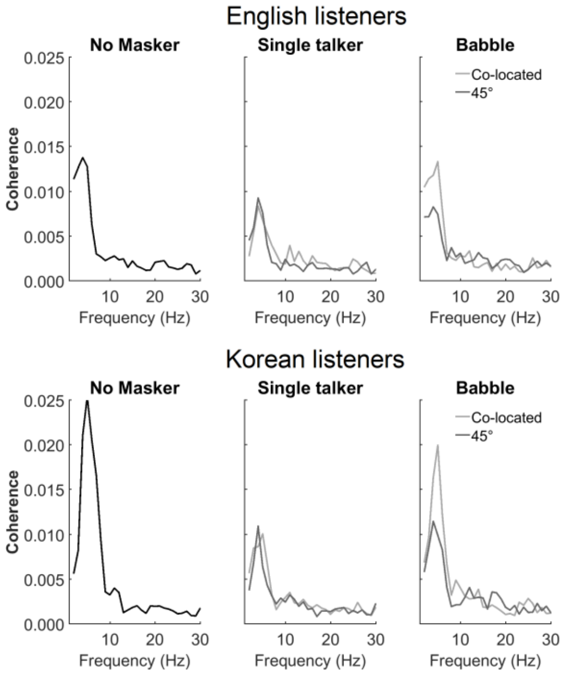
Figure 1: Boxplots showing the proportions of correct identification of anomalous sentences by listener group and noise condition



As shown in Fig. 2, Korean listeners had greater entrainment to the amplitude envelopes of target talkers than did English listeners in the no-masker condition. However, entrainment by Korean listeners decreased substantially when there was background noise. A mixed-model analysis was performed on the coherence results; coherence values averaged in the delta-theta range (2-8Hz) were used as the dependent variable, listener group and masker type as independent variables, and each subject as a random intercept. The results demonstrated that the interaction of listener group and masker type was significant, $\chi^2(2) = 16.07$, $p < 0.001$. Specifically, the difference in coherence between the no-masker and noise conditions was greater for Korean than for English listeners, b = - 0.0035, t(92) = - 3.89, $p < 0.001$. The main effect of masker type was significant, $\chi^2(2) = 28.82$, $p < 0.001$, but the main effect of listener group was not, $p = 0.106$.

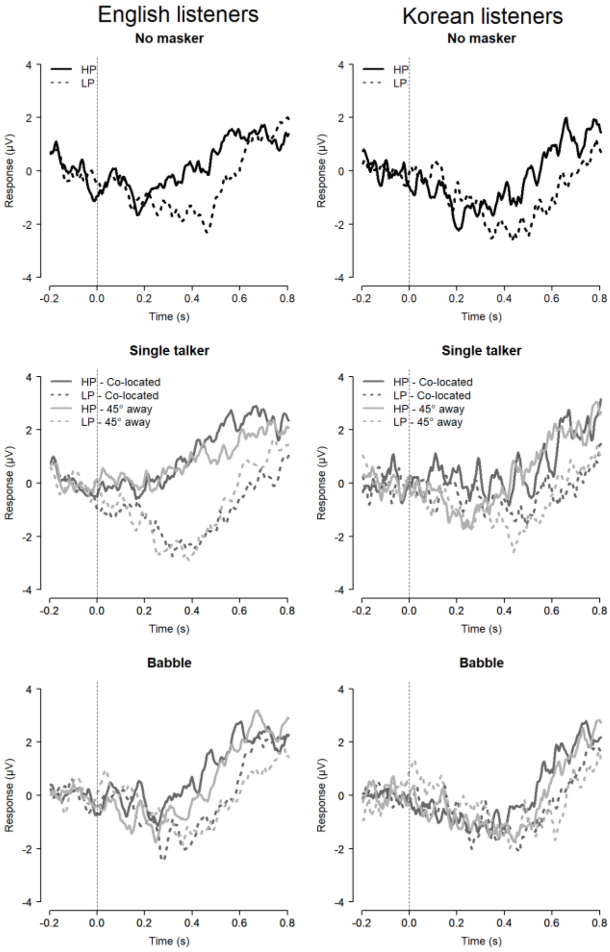
To examine the effect of spatial separation on target speech tracking, an additional mixed-model analysis was performed only for the noise conditions. The main effect of spatial cues was significant, $\chi^2(1) = 8.79$, $p = 0.003$. Target-speech entrainment was higher when the target and masker were co-located, likely because a higher SNR was used for this condition. The main effect of listener group or the interaction of the two variables was not significant, $p > 0.05$.

Figure 2: Plots showing coherence values as a function of frequency by listener group and noise condition



A mixed-model analysis was also conducted with N400 amplitudes as the dependent variable, sentence type (high vs. low cloze probability), listener group, and masker type as independent variables, and with by-subject random intercepts. The N400 amplitude was larger (i.e., more negative) for low than high cloze probability sentences as shown in Fig. 3, indicating greater effort for lexical processing when words were less predictable. However, this context-related N400 difference (i.e., N400 effect) was smaller for Korean than English listeners. There were a significant interaction of sentence type and listener group, $\chi^2(1) = 11.66$, $p < 0.001$, and a main effect of sentence type, $\chi^2(1) = 43.89$, $p < 0.001$. In addition, there was a significant interaction of masker type and sentence type, $\chi^2(2) = 9.37$, $p = 0.009$; the N400 difference between high and low cloze probability sentences was significantly greater when sentences were presented in the single-talker noise than in the unintelligible babble noise, b = -1.39, t(210) = - 3.049, $p = 0.0026$. The main effects of listener group and masker type and other interactions were not significant, $p > 0.05$.

Figure 3: Grand average ERP waveforms to sentence-final words by listener group, sentence type (HP: high cloze probability sentences, LP: low cloze probability sentences), and noise condition.



4. DISCUSSION

The present study demonstrated that listeners differ in the ways they modulate their processing in difficult conditions. Target-talker entrainment by non-native listeners was greater than that of native listeners in the no-masker condition, replicating the previous finding [21]. This likely occurred because they focused greater attention on the acoustic signal to compensate for their great difficulties with the L2 speech.

However, speech recognition by non-native listeners was less robust to adverse conditions than for native listeners at this auditory level. The reduced speech tracking likely resulted from non-native listeners having poor representations of the acoustic input at a peripheral level or by difficulties in selecting and attending to target speech signals. Even though non-native listeners can use additional attentional mechanisms to enhance entrainment in difficult listening conditions, it appears that this can break down more rapidly when their recognition is overwhelmed by the additional demands of the background noise.

The results also suggested that both groups of listeners relied more on lexical processing when the distractor involved greater informational masking. In general, a larger N400 indicates greater effort in the lexical access process, and a greater N400 difference between high- and low-predictability words indicates greater use of context. These N400 effects were larger when the distractor involved informational masking (i.e., single talker) than when subjects heard a constant babble. It thus appears that listeners modulated their lexical processing to overcome the effect of single-talker maskers, by searching lexical candidates more carefully or increasing reliance on semantic cues to aid lexical access. Moreover, this strategy was available to non-native listeners, despite the fact that their degree of lexical processing was lower overall.

This study thus demonstrates that speech recognition in noise is a complex process, in which listeners are able to modulate attention to the acoustic signal and the amount of lexical processing, dependent both on the demands of the listening conditions and the language experience of the listeners.

5. REFERENCES

[1] Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., Merzenich, M. M. 2001. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc Natl Acad Sci USA* 98(23), 13367-13372.

[2] Bradlow, A.R., Alexander, J. 2007. Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *J Acoust Soc Am*. 121(4), 2339-2349.

[3] Brown, C., Hagoort, P. 1993. The Processing Nature of the N400: Evidence from Masked Priming. *J Cogn Neurosci.*5(1), 34-44.

[4] Brungart, D. S. 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am* 109, 1101–1109.

[5] Carey. D., Mercure, E., Pizzioli, F., Aydelott, J. 2014. Auditory semantic processing in dichotic listening: Effects of competing speech, ear of presentation, and sentential bias on N400s to spoken words in context. *Neuropsychologia* 65,102-112.

[6] Crosse, M. J., Di Liberto, G. M., Bednar, A., Lalor, E. C. 2016. The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Front Hum Neurosci* 10(604), 1-14.

[7] Delorme, A., Makeig, S. 2004. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* 134(1), 9-21.

[8] Ding, N., Chatterjee, M., Simon, J. Z. 2014. Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage* 88, 41-46.

[9] Freyman, R. L., Balakrishnan, U., Helfer, K.S. 2004. Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *J Acoust Soc Am.*115(5), 2246-2256.

[10] Hattori, K., Iverson, P. 2009. English /r/-/l/ category assimilation by Japanese adults: individual differences and the link to identification accuracy. *J Acoust Soc Am.*125(1), 469-479.

[11] Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., Siebert, C. 2003. A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87(1), B47–B57.

[12] Lecumberri, M. L.G., Cooke, M., Cutler, A. 2010. Non-native speech perception in adverse conditions: A review. *Speech Commun.* 52(11-12), 864-886.

[13] Lopez-Calderon, J., Luck, S. J. 2014. ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Front Hum Neurosci.* 8, 1-14.

[14] Luo, H., Poeppel, D. 2007. Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron* 54(6), 1001-1010.

[15] 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(3), 686.

[16] Obleser, J., Kotz, S. A. 2011. Multiple brain signatures of integration in the comprehension of degraded speech. *Neuroimage* 55(2), 713-723.

[17] Oostenveld, R., Fries, P., Maris, E., Schoffelen, J. M. 2011. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci*, 1-9.

[18] Peelle, J. E., Gross, J., Davis, M. H. 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb Cortex.* 23(6), 1378-1387.

[19] Van Petten, C., Kutas, M. 1990. Interactions between sentence context and word frequency in event-related brain potentials. *Mem Cognit.* 18(4), 380-393.

[20] Rosen, S., Souza, P., Ekelund, C., Majeed, A. A. 2013. Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *J Acoust Soc Am.*133(4), 2431-2443.

[21] Song, J., Iverson, P. 2018. Listening effort during speech perception enhances auditory and lexical processing for non-native listeners and accents. *Cognition* 179, 163-170.

[22] Stringer, L. M. 2015. Accent intelligibility across native and non-native accent pairings: investigating links with electrophysiological measures of word recognition. (Unpublished MPhil dissertation). University College London, London.