INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The Role of Spectral Resolution in Foreign-Accented Speech Perception

Michelle R. Kapolowicz, Vahid Montazeri, and Peter F. Assmann

School of Behavioral and Brain Sciences
The University of Texas at Dallas
{michelle.kapolowicz, vahid.montazeri, assmann}@utdallas.edu

Abstract

Several studies have shown that diminished spectral resolution leads to poorer speech recognition in adverse listening conditions such as competing background noise or in cochlear implants. Although intelligibility is also reduced when the talker has a foreign accent, it is unknown how limited spectral resolution interacts with foreign-accent perception. It is hypothesized that limited spectral resolution will further impair perception of foreign-accented speech. To test this, we assessed the contribution of spectral resolution to the intelligibility of foreign-accented speech by varying the number of spectral channels in a tone vocoder. We also examined listeners' abilities to discriminate between native and foreign-accented speech in each condition to determine the effect of reduced spectral resolution on accent detection. Results showed that increasing the spectral resolution improves intelligibility for foreign-accented speech while also improving listeners' ability to detect a foreign accent but not to the level of accuracy for broadband speech. Results also reveal a correlation between intelligibility and accent detection. Overall, results suggest that greater spectral resolution is needed for perception of foreign-accented speech compared to native speech.

Index Terms: foreign-accented speech perception, foreign-accent detection, Chinese-accented English, spectral resolution

1. Introduction

Foreign-accented speech (FAS) is considered to carry an auditory perceptual distortion which requires more time and cognitive effort to understand [1]. Despite the initial difficulty with understanding FAS, listeners are generally able to adapt [2], [3], [4], [5], [6], [7], and [8]. In this paper, we examine the relationship between perceived accentedness and intelligibility during listeners' initial exposure to non-native talkers as a function of spectral resolution (SR). It is predicted that decreased SR will result in decreased intelligibility and accent detection for FAS.

Compared to cochlear implant (CI) users, normal hearing listeners have less difficulty perceiving native speech (NS) in quiet [9]. CI users also have more difficulty with talker variability, such as talkers with varying linguistic backgrounds [10], [11], and [12]. An attributable difference between CI users and normal hearing listeners is decreased SR in CI users. CI users have access to reduced SR mainly due to a limited number of physical channels. The importance of SR in speech perception is further supported by studies showing that intelligibility is reduced in adverse listening conditions, such as in the presence of competing background noise [13] and [14], due limited spectral resolution.

Evidence from CI users showed lower intelligibility for FAS compared to normal hearing listeners [15]. CI users have more difficulty detecting foreign accents than normal hearing listeners, which may limit their ability to make rapid perceptual adjustments required to adapt to the deviation from the expected target speech signal [16]. This evidence supports the hypothesis that there should be a systematic relationship between intelligibility and SR, as well as a relationship between accent detection and SR. A correlation between intelligibility and accent detection as a function of spectral resolution is also predicted. As aforementioned, it is expected that FAS will undergo a further reduction in intelligibility and accent detection when SR is reduced.

To parse the effect of reduced SR from other potential confounding factors found in CI users, we tested normal hearing listeners using speech processed through a vocoder, where the number of channels can be varied to limit SR cues available to the listener [17]. Given that there is a general advantage for higher SR in difficult listening situations, it is expected that perceiving FAS will also require greater SR compared to NS. This hypothesis was tested by examining the effect of SR on speech intelligibility.

By increasing SR, it is expected that the "foreign-accentedness" of the speech would be more obvious to listeners than when SR is reduced. On the other hand, greater SR generally improves speech perception. However, there is a potential conflict if increasing the SR also increases the distortion stemming from the foreign accent. This conflict is addressed in this study by investigating the relationship between perceived accentedness and intelligibility. It may also be argued that decreasing SR is itself a source of perceptual distortion. Although this may be case, for normal hearing listeners, only minimal SR cues are needed to reach near perfect intelligibility in quiet [17] and [18]. To control for this, outcomes from the perception of vocoded native speech are also investigated in this study.

2. Methods

2.1. Speech Materials

Audio recordings of low-context Harvard sentences [19] were obtained from 15 native (5 males, 10 females; age range: 18-38 years) and 18 non-native (9 males, 9 females; age range: 18-47 years) talkers of American English. All talkers were students at The University of Texas at Dallas. Non-native talkers, with a range of 2 weeks to 22 years of residency in Texas, were born and raised in Taiwan and reported using Mandarin as their native language. Non-native talkers were

http://dx.doi.org/10.21437/Interspeech.2016-1585

paid a nominal fee for producing the recordings. Native talkers have only ever resided in Texas and were monolingual. Native talkers were awarded research credits for participation. Both groups were given a brief hearing screening and reported no hearing impairments.

Talkers were instructed to repeat each sentence after listening to the sentence spoken by a male native American English talker and viewing a transcript of the sentence on a computer monitor. Recordings were made in a sound-attenuated booth using a Shure SM-94 microphone, Symetrix SX202 dual-microphone pre-amplifier and Tucker-Davis Technologies data acquisition hardware (MA1, RP2.1). Digital waveforms were stored on a computer disk at a rate of 48 kHz and 16-bit resolution. Sentences were RMS-equalized across all talkers.

2.2. Talker Group Assignment

Twenty listeners recruited for this task were monolingual native English-speaking university students ranging in age from 18 to 26 years who had only ever resided in Texas. Listeners were awarded research credits for participation and were screened for normal hearing. Sentences from the previously recorded talkers were presented in quiet through Sennheiser HD 598 headphones at a comfortable level in a sound booth. Participants were asked to listen to each sentence and type the words that they heard. Listeners heard two sentences from each talker, and no sentence was repeated. Presentation of talkers and sentences were randomized. Intelligibility scores were based on the percentage of words correctly heard based on what listeners typed. The average intelligibility score for each talker was used to provide intelligibility rankings. The bottom 6 talkers (4 females, 2 males) were used for the FAS group (66% group mean intelligibility score in quiet) and the top 6 talkers (4 females, 2 males) were used for the native talkers (96% group mean intelligibility score in quiet). Listeners were also asked to rate the degree of foreign-accentedness using a 9 point Likert scale, with 1 being *no foreign accent* and 9 being *heavily foreign-accented.*

2.3. Speech Processing

A sine-wave processor was implemented replicating the specifications of Dorman *et al*. [18]. Speech stimuli were first passed through a pre-emphasis filter (low-pass below 1200 Hz, -6 dB per octave). The filtered signals were then band-passed using sixth-order Butterworth filters into N logarithmically-spaced frequency bands (where N was either 3, 4, 5, or 9, based on the expectation reported in [18], showing that performance would reach a plateau within this range). The envelopes of the band-passed signals were then extracted with full-wave rectification followed by low-pass filtering using a second-order Butterworth filter with a cutoff frequency set to 160 Hz. N sinusoids were then generated with amplitudes equal to the RMS energy of the envelopes (computed every 4 ms) and frequencies equal to the center frequencies of the bandpass filters. The sinusoids generated for each band were then multiplied by the envelopes, filtered using the same bandpass filters, and finally summed across channels. For additional information, see [18].

2.4. Experimental Procedure

To test the effect of spectral resolution on FAS perception, 11 students at The University of Texas at Dallas (who did *not* participate in our talker group assignment procedure) with an age range of 18 to 25 were recruited for the listening experiment. Listeners (monolingual, native English talkers from Texas) were screened for normal hearing, and were awarded research credits for participation.

In each trial, the target signal was randomly selected (without replacement) from the previously recorded sentences. Participants were asked to type the words they heard. They were also required to specify whether or not the talkers had a foreign accent. Intelligibility scores were calculated as the ratio of the number of correctly identified key words to the total number of presented key words.

The experimental design was a 4 x 2 repeated measure design: 4 SR configurations (3, 4, 5, and 9 channels) x 2 accent conditions (native and foreign accented). The experiment was conducted in a double-walled sound booth. In each trial, participants were presented with stimuli through a Tucker-Davis sound system and Sennheiser HD 598 headphones. The stimuli were presented to the listeners at a comfortable level.

3. Results

3.1. Intelligibility

Figure 1 shows the results of intelligibility scores as a function of number of vocoder channels and talkers' accentedness. A repeated measures analysis of variance indicated a significant main effect of number of channels ($F(3,30) = 96.89$, $p < 0.01$), as well as a significant main effect of foreign accentedness ($F(1,10) = 186.2$, $p < 0.01$), and also a significant interaction of accentedness by number of channels ($F(3,30) = 3.49$, $p < 0.05$) on speech intelligibility scores.

Post-hoc comparisons using Bonferroni corrections revealed significant differences between native versus foreign-accented intelligibility scores for 3, 4, 5, and 9 channels (all $p < 0.01$). Additional post-hoc analyses were performed to compare intelligibility scores between channels for NS. Analyses showed that there was a significant difference between intelligibility scores for 3 and 4 channels, for 3 and 5 channels, for 3 and 9 channels, for 4 and 9 channels, and for 5 and 9 (all $p < 0.01$). Analyses also indicated that the difference between intelligibility scores for 4 and 5 channels was not significant, ($p = 1.00$). Post-hoc analyses were also performed to compare intelligibility scores between channels for FAS talkers. Analyses showed that there was a significant difference between intelligibility scores for 3 and 5 channels, for 3 and 9 channels, for 4 and 9 channels, and for 5 and 9 channels (all $p < 0.01$). Differences between intelligibility scores were not significant for 3 and 4 channels, ($p = 0.18$) nor for 4 and 5 channels, ($p = 1.00$).

3.2. Foreign Accent Detection

Figure 2 summarizes the perceived accent judgments (where 0 = unaccented and 100 = accented, averaged across talkers and listeners) as a function of number of vocoder channels and talkers' accentedness (NS, FAS). A mixed-effects logistic regression model on judgments of perceived accentedness indicated a significant effect of number of

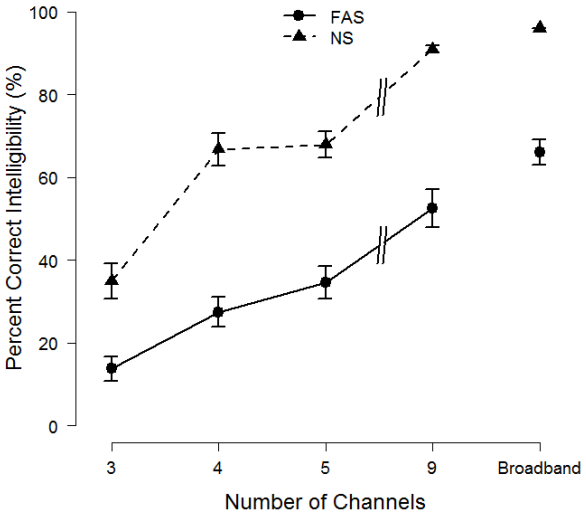

Figure 1: *Intelligibility scores across channels for NS and FAS. Error bars represent the standard error of the means. Broadband scores were collected from a previous unpublished study and are provided here for comparative purposes; the talkers are the same, but the listeners differ from the listeners in this study.*

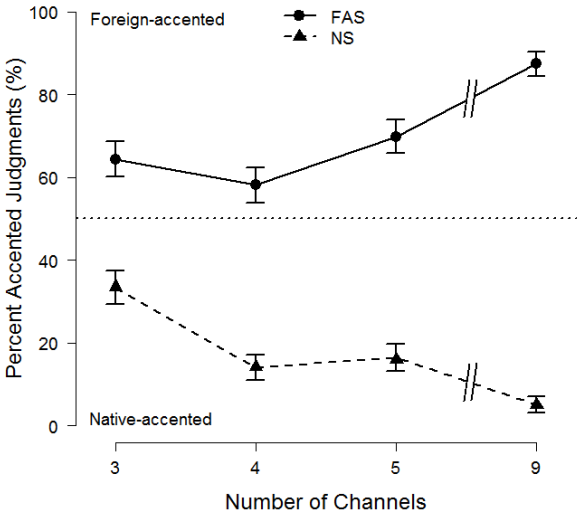

Figure 2: *Perceived accentedness across channels for NS and FAS. Error bars represent the standard error of the means. Scores approaching 0 correspond to participants' judgments as NS; scores close to 100 correspond to participants' judgments as FAS. Note the increase in accuracy of accent judgments for both FAS and NS conditions with increasing number of channels. Dotted horizontal line indicates chance.*

channels ($\chi 2 = 35.32$, $p < 0.01$) and a significant effect of talker accentedness ($\chi 2 = 24.27$, $p < 0.01$). For native talkers, a significant improvement was observed for 4 or more channels compared to 3 channels ($p < 0.01$ for each comparison). For foreign-accented talkers, only the 9-channel condition produced significantly better accent detection compared to 3 channels ($p < 0.01$).

3.3. Relationship between Intelligibility and Foreign Accent Detection in Vocoded Speech

Figure 3 presents a scatterplot of mean intelligibility and foreign accent detection scores for individual listeners in each channel condition. The plot shows a systematic relationship between accent detection, intelligibility and number of channels: Intelligibility and accent detection are both higher when the number of channels increases. However, the benefit associated with increasing the number of channels is larger for NS than for FAS. A significant linear relationship between accent detection and intelligibility was found for each talker group ($r = 0.48$, $p < 0.01$ for FAS and $r = -0.67$, $p < 0.01$ for NS). (Note that the signs of the correlations are reversed for the two talker groups because the abscissa shows accentedness judgments rather than proportion correct).

3.4. Relationship between Intelligibility and Foreign Accent Detection in Broadband Speech

Figure 4 shows the relationship between intelligibility scores and participants' ratings of foreign-accentedness for broadband FAS (in contrast to Figure 3 which shows accent detection for vocoded speech). A strong negative relationship was found ($r = -0.82$, $p < 0.01$). The comparison shows that listeners had more difficulty understanding talkers who were rated as having a heavier foreign accent.

4. Discussion

FAS introduces a type of auditory perceptual distortion that is intrinsic to the signal itself whereas other distortions, such as competing background noise, are external to the source of the signal [20]. The aim of the present study was to examine the effects of FAS and reducing the number of channels in a vocoder (two forms of intrinsic distortion) on speech intelligibility and accent detection. Our results show that greater SR is required when listening to FAS compared to NS.

Dorman *et al*. [18] showed that for NS in quiet, sentence intelligibility approaches ceiling with 4 channels. They found that adding more SR did not further benefit the listener. In the present study, there were no differences in intelligibility between 4 and 5 channels for NS in quiet, but there was a dramatic improvement approaching ceiling for 9 channels. The present study did not test SR using 6, 7, or 8 channels, (since [18] reported no perceptual benefit gained by increasing the number of channels from 5 to 9 for sentences spoken in quiet) so the precise point at which performance reaches a plateau is uncertain. Given the similarity in signal processing, the discrepancy between the two studies may be attributable to differences in speech materials. The present study used the more complex Harvard sentences, whereas the previous study used the more predictable HINT sentences (presented without competing noise).

Overall, intelligibility for FAS remained lower than for NS, as expected. Interestingly, unlike for NS with limited SR, listeners did not reach the same level of intelligibility for 9 channels as they did when they had access to broadband FAS. This leads to the conclusion that decreasing SR produces an additional deficit for understanding FAS. Also the benefit

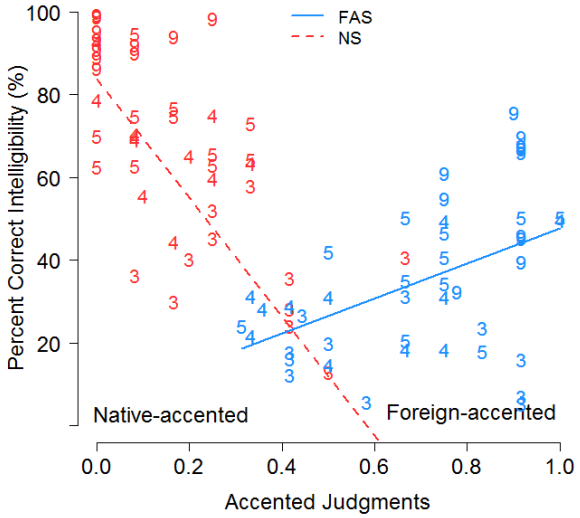

Figure 3 (colored online): *The relationship between accent detection and intelligibility for NS (red dashed line) and FAS (blue solid line). Numbers represent spectral resolution condition (number of channels).*

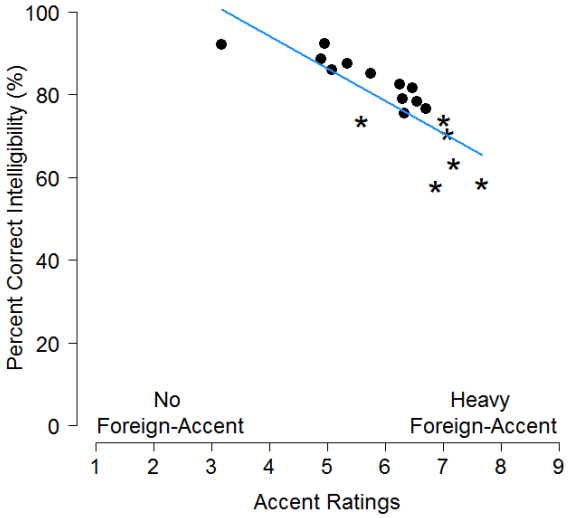

Figure 4 (colored online): *The relationship between accent rating and intelligibility for foreign-accented broadband speech. Stars represent subset of FAS used in the vocoder experiment.*

gained from increasing SR is limited for perceiving FAS when compared to NS. The question remains as to whether increasing SR further would benefit FAS perception as suggested by the results in [14]. Increasing the low-pass cutoff frequency and the number of channels can directly test this possibility and is the aim of ongoing studies.

In this study, we also examined listeners' abilities to determine whether talkers were native or foreign-accented. We found that accent detection increases with SR. The database used in this study consists of recorded speech samples from native and foreign-accented talkers; all foreign-accented talkers were rated as having a foreign accent. This ensures that, under broadband conditions, listeners were able to detect a foreign accent with 100% accuracy, unlike with reduced SR, where this task was more difficult.

The results of the present study also show a correlation between accent detection and intelligibility: Listeners are better able to detect differences between NS and FAS when the talkers are more intelligible. Somewhat unexpectedly, this implies that listeners can perceive FAS more accurately despite an increase in distortion from the accent. This result further strengthens the claim that degraded SR presents an added difficulty for the perception of FAS.

Although we found a strong correlation between accent ratings and intelligibility in broadband data, previous researchers have reported a weak correlation. Munro and Derwing [21] presented their FAS talkers with a cartoon story, and asked them to describe the story in their own words. In comparison, the present study elicited the low-context, low redundancy Harvard sentences. Although both studies used accented talkers with similar demographics, the difference in speech elicitation methods might explain this discrepancy.

Studies have shown the benefits of training on listeners' abilities to perceive vocoded speech [22] and [23] as well as lexically challenging words [24]. Studies have also shown that listeners have the ability to adapt to unprocessed FAS with increased exposure [2], [3], [4], [5], [6], [7], and [8]. As such, our future studies aim to focus on the effects of exposure and short-term training for the perception of FAS with limited SR to see if adaptation can occur without further increasing SR. This would allow us to determine the importance of SR for adaptation to FAS over time. This question is especially important for CI users whose devices provide reduced SR.

5. Conclusions

Our results, which tested normal hearing listeners on their ability to perceive FAS with decreased SR, corroborates evidence reported in [15] and [16] (both of which tested FAS perception in CI users). Taken together, these studies show the importance of spectral information in FAS perception. The data presented here reveals that listeners struggled more with accurately identifying whether or not a talker was foreign-accented when the SR was reduced. Listeners also showed a decrease in intelligibility with lower SR. There was a direct relationship between accent detection (both foreign and native) and intelligibility: More accurate detection of the presence or absence of a foreign accent was associated with higher intelligibility scores. Also, across all channel conditions, listeners were less accurate when detecting FAS compared to NS, and intelligibility scores were lower for FAS than for NS. This evidence strongly suggests that more SR is needed to perceive FAS than NS.

6. Acknowledgements

The authors would like to thank the students who participated in this research as well as James Kuang for assisting with data collection.

7. References

[1] K.J. Van Engen and J.E. Peelle, "Listening effort and accented speech," *Front Hum Neurosci*, 8, DOI: 10.3389/fhum.2014.00577, 2014.

[2] C.M. Clarke and M.F. Garrett, "Rapid adaptation to foreign-accented English," *J Acoust Soc Am*, 116, pp. 3647-3658, 2004.

[3] A.R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech," *Cognition*, 106, pp. 707-729, 2007.

[4] C. Floccia, J. Bunler, J. Goslin, and L. Ellis, "Regional and foreign accent processing in English: Can listeners adapt?" *J Psycholinguist Res*, 38, pp. 379-412, 2009.

[5] A.M. Trude, A.Tremblay, and S. Brown-Schmidt, "Limitations on adaptation to foreign accents," *J Mem Lang*, 69, pp. 349-367, 2013.

[6] M.M. Baese-Berk, A.R. Bradlow, and B.A. Wright, "Accent-independent adaptation to foreign accented speech," *J Acoust Soc Am*, 133, EL174-180, 2013.

[7] M.J. Witteman, A. Weber, and J.M. McQueen, "Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation," *Atten Percept Psychophys*, 75, pp. 537-556, 2013.

[8] M. J. Witteman, A. Weber, and J.M. McQueen, "Tolerance for inconsistency in foreign-accented speech," *Psychon Bull Rev*, 21, pp. 512-519, 2014.

[9] K.F. Faulkner and D.B. Pisoni, "Some observations about cochlear implants: challenges and future directions," *Neuroscience Discovery*, DOI: 10.7243/2052-6946-1-9, 2013.

[10] M. Cleary and D.B. Pisoni, "Talker discrimination by prelingually deaf children with cochlear implants: preliminary results," *Ann Otol Rhinol Laryngol*, 111, pp. 113-118, 2002.

[11] C.G. Clopper and D.B. Pisoni, "Perceptual dialect categorization by an adult cochlear implant user: a case study," *Int Congr*, 1273, pp. 235-238, 2004.

[12] M. Cleary, D.B. Pisoni, and K.I. Kirk, "Influence of voice similarity on talker discrimination in children with normal hearing and children with cochlear implants," *J Speech Lang Hear Res*, 48, pp. 204-223.

[13] P.C. Loizou, A. Mani, and M.F. Dorman, "Dichotic speech recognition in noise using reduced spectral cues," *J Acoust Soc Am*, 114, pp. 475-483, 2003.

[14] R.V. Shannon, Q.J. Fu, and J. Galvin III, "The number of spectral channels required for speech recognition depends on the difficulty of the listening situation," *Acta Otolaryngol*, 552, pp. 50-54, 2004.

[15] C. Ji, J.J. Galvin, Y. Chang, A. Xu, and Q.J. Fu, "Perception of speech produced by native and nonnative talkers by listeners with normal hearing and listeners with cochlear implants," *J Speech Lang Hear Res* 57, pp. 532-542, 2014.

[16] T.N. Tamati and D.B. Pisoni, "The perception of foreign-accented speech by cochlear implant users," in 18th International Congress of Phonetic Sciences, Glasgow, UK, August 2015.

[17] R.V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, 270, pp. 303-304, 1995.

[18] M.F. Dorman, P.C. Loizou, and D. Rainey, "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," *J Acoust Soc Am*, 102, pp. 2403-2411, 1997.

[19] IEEE Subcommittee, "IEEE Recommended Practice for Speech Quality Measurements", *IEEE Transactions on Audio and Electroacoustics*, 17, pp. 225-246, 1969.

[20] H. Lane, "Foreign accent and speech distortion," *J Acoust Soc Am*, 35, pp. 451-453, 1963.

[21] M. Munro and T.M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Lang Learn*, 45, pp. 73-97, 1995.

[22] A. Hervais-Adelman, M.H. Davis, I.S. Johnsrude, and R.P. Carlyon, "Perceptual learning of noise vocoded words: effects of feedback and lexicality," *J Exp Psychol Hum Percept Perform*, 34, pp. 460-474, 2008.

[23] T. Bent, J.L. Loebach, L. Phillips, and D.B. Pisoni, "Perceptual adaptation to sinewave-vocoded speech across languages," *J Exp Psychol Hum Percept Perform*, 37, pp. 1607-1616, 2011.

[24] M.H. Burk and L.E. Humes, "Effects of training on speech recognition performance in noise using lexically challenging words," *J Speech Lang Hear Res*, 50, pp. 25-40, 2007.