INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Simultaneous articulatory and acoustic distortion in L1 and L2 Listening: Locally time-reversed "fast" speech

Mako Ishida[1, 2]

[1]Sophia University, Japan,
[2]Japan Society for the Promotion of Science, Japan

Abstract

The current study explores how native and non-native speakers cope with simultaneous articulatory and acoustic distortion in speech perception. The articulatory distortion was generated by asking a speaker to articulate target speech as fast as possible (fast speech). The acoustic distortion was created by dividing speech signals into small segments with equal time duration (e.g., 50 ms) from the onset of speech, and flipping every segment on a temporal axis, and putting them back together (locally time-reversed speech). This study explored how "locally time-reversed fast speech" was intelligible as compared to "locally time-reversed normal speech" measured in Ishida, Samuel, and Arai (2016). Participants were native English speakers and native Japanese speakers who spoke English as a second language. They listened to English words and pseudowords that contained a lot of stop consonants. These items were spoken fast and locally time-reversed at every 10, 20, 30, 40, 50, or 60 ms. In general, "locally time-reversed fast speech" became gradually unintelligible as the length of reversed segments increased. Native speakers generally understood locally time-reversed fast spoken words well but not pseudowords, while non-native speakers hardly understood both words and pseudowords. Language proficiency strongly supported the perceptual restoration of locally time-reversed fast speech.

Index Terms: locally time-reversed speech, fast speech, perceptual restoration, speech perception, L1 vs. L2.

1. Introduction

1.1. Locally time-reversed speech

What was striking about the first research of locally time-reversed speech by Saberi and Perott (1999) was that a spoken sentence was intelligible even when every certain length of speech signal (e.g., 50 ms) was flipped on a temporal axis [1]. The constituents of speech signal were shifted forward or backward from the original position by the local time reversal, but listeners were able to retrieve the "shifted" or "scattered" information from the reversed segments, and integrate them to perceptually restore the original speech [1, 2, 3, 4, 5, 6, 7]. In general, locally time-reversed speech was intelligible when the reversed segment length was relatively short [1, 2, 3, 4, 5, 6, 7]. However, speech became gradually unintelligible as the length of the reversed segments became longer [1, 2, 3, 4, 5, 6, 7]. The dispersed information, by the local time reversal, can be perceptually retrieved and integrated when the reversed segment length is relatively short.

1.2. Perceptual Restoration: Lexicality

The restorability of locally time-reversed speech also depends on lexicality. Kiss et al. (2008) examined the intelligibility of locally time-reversed speech at a sentence level, and suggested that listeners were able to understand speech when sentences were made of real words – sentences containing only pseudowords were hardly understood [5]. Further, Ishida, Samuel, and Arai (2016) examined the intelligibility of locally time-reversed speech solely at a lexical level, and suggested that locally time-reversed words were significantly more intelligible than locally time-reversed pseudowords [3]. In addition, Garataloup et al. (2009) examined the intelligibility of locally time-reversed words and pseudowords by flipping *syllables* in time [4]. The results suggested that words were significantly more intelligible than pseudowords when syllables were locally time-reversed. It seems that lexicality strongly supports the perceptual restoration of locally time-reversed speech.

1.3. Perceptual Restoration: Phonemic Constituents

At the same time, the intelligibility of locally time-reversed speech also depends on phonemic constituents of speech. Ishida, Samuel, and Arai (2016) suggested that fricative-dominant words and pseudowords (that contained many fricatives as compared to stops) were significantly more intelligible than stop-dominant words and pseudowords (that contained many stops as compared to fricatives) when locally time-reversed [3]. It seems that fricative consonants have relatively symmetric waveforms, while stop consonants have relatively asymmetric waveforms (with the initial burst at the onset of speech as well as VOT), and this seems to have impacted the intelligibility of speech. That is, fricative-dominant items (symmetric waveforms) generally retain their original contours of speech even when locally time-reversed, while stop-dominant items (asymmetric waveforms) drastically change their contours. The intelligibility of speech is, seemingly, severely impaired when the amplitude envelope of speech signals drastically changes from the original shape. The acoustic characteristics of phonemes in speech seem to affect the restorability of speech.

1.4. Perceptual Restoration: Fast Speech

On the other hand, the intelligibility of speech also depends on speech rate as well as subsequent pronunciation variations. In daily situations, people tend to speak fast, and adjacent phonemes and words are often connected and pronounced together (connected speech). Dalby (1986) reported that the number of syllables in fast spoken American English was frequently reduced [8]. For example, the word

http://dx.doi.org/10.21437/Interspeech.2017-83

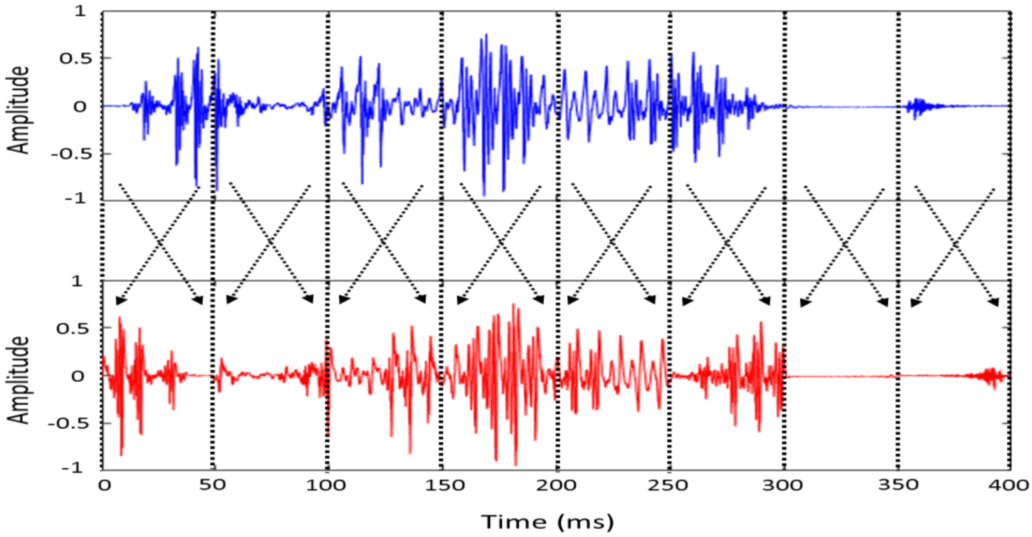

Figure 1: *The manipulation of local time reversal. The upper panel shows the original waveform of a fast spoken word "academic" in blue. The lower panel shows the waveform of locally time-reversed fast speech "academic" in red – every 50 ms of original speech was locally time-reversed.*

"probably" [prɑbəbli] was pronounced as [prɑbli] (two phonemes omitted), [prɑli] (three phonemes omitted), and [prɑˈ] (five phonemes omitted). Johnson (2004) [9] analyzed the American English Corpus "Variation in Conversation Corpus" (38560 content words, 49362 function words) [10, 11, 12], and reported that the number of phonemes was reduced in 10% of content words and in 20% of function words [9]. In addition, phonemes were altered in 25% of content words and in 40% of function words [9]. For example, the word "until" [ʌntɪl] was pronounced as [əntʌ_] with the omission of the last phoneme [l], and the deviation of three phonemes: [ʌ] was pronounced as [ə], [ɪ] as [ʌ], [l] as none. In addition, Brown and Kondo-Brown (2006) analyzed casual speech in English and reported that many adjacent words are often combined and pronounced as connected speech [13]. In general, connected speech sounds different from clearly articulated speech. For example, "Where are you?" was pronounced as "Wheraya?", and "Did you eat yet?" was pronounced as "J'eat jet?" [14, 15, 16]. These sound alterations are not necessarily predictable, since every person has a different speaking style. Listeners presumably need to retrieve relevant cues, and map these cues onto lexical items in their mental lexicon, to understand speech.

1.5. Perceptual Restoration: Language Proficiency

Moreover, the restorability of locally time-reversed sentences also relates to the listener's language proficiency. Kiss et al. (2008) suggested that native German speakers were generally more successful in perceptual restoration of locally time-reversed German sentences than non-native speakers [5]. In general, semantically coherent sentences (i.e., both sentential and lexical contexts were available) were generally more intelligible than semantically incoherent sentences (i.e., only lexical context was available) when the same temporal inversion was imposed. Both sentential and lexical contexts seem to be critical for perceptual restoration, but the availability of contextual cues seems to depend on the listener's language proficiency [cf. 17, 18]. Native speakers would find it relatively easy to make connections between the perceived sounds and lexical items to understand speech, but non-native speakers would find it challenging to understand the pronunciation variations.

2. Experiment

The current study explores how native and non-native listeners cope with simultaneous articulatory and acoustic distortion in speech perception (i.e., fast speech rate and local time distortion). Listeners listened to a locally time-reversed fast spoken word (or pseudoword) in a male voice, followed by a normally spoken unreversed word (or pseudoword) in a female voice, and judged if the first and second speakers said the same or different words. There were three research questions: (1) How much is locally time-reversed fast speech intelligible when the reversed segment length gradually increases? (fast speech rate + temporal distortion) (2) Are locally time-reversed fast spoken words better restored than locally time-reversed fast spoken pseudowords (lexicality)? (3) How much do native and non-native speakers' performance differ in perceptual restoration (language proficiency)? This study explores the impact of fast speech rate on the intelligibility of locally time-reversed speech, and the influence of lexicality and language proficiency in perceptual restoration.

2.1. Materials

The same set of 60 stop-dominant words and 60 stop-dominant pseudowords were adopted from Ishida, Samuel and Arai (2016) [3]. Here, each pair of a word and a pseudoword was different only by one phoneme (e.g., "aca**d**emic" vs. "aca**b**emic"). The target stimuli, which were to be locally time-reversed, were recorded in a sound proof room by using a digital audio recorder (SONY PCM-D50) and a microphone (SONY ECM-MS957). These items were spoken fast by a male native American English speaker. The average duration of the fast spoken words and pseudowords was 664 ms, while the average duration of normally spoken words and pseudowords in Ishida, Samuel, and Arai (2016) was 1,025 ms [3]. Thus, the fast spoken items were 1.54 times faster than the original items. The stimuli were first recorded at a sampling rate of 48 kHz with 16-bit resolution, and downsampled to 16 kHz (16 bit) and saved as WAV files. These items were locally time-reversed at every 10, 20, 30, 40, 50, or 60 ms (Figure 1). The reversed segment length was increased by 10 ms steps, in order to observe any subtle transitional differences of intelligibility. The joints of the adjacent reversed segments were cross-faded (i.e., 5-ms linear onset and offset ramps were imposed) to prevent any additional noise or clicks.

2.2. Participants

2.2.1. Native speakers

A total of 30 native English speakers from Stony Brook University (23 female, 7 male, ave. 20.17 years old) participated in this study. No participants reported any hearing or speech impairments. They received course credits for their participation.

2.2.2. Non-native speakers

A total of 30 native Japanese speakers who spoke English as a second language (18 female, 12 male, ave. 33.73 years old) participated in this study. Their English proficiency was intermediate, based on the placement test of DIALANG which measured the vocabulary size of test-takers [19, 20]. Their average score was 429.93 out of 1,000 full marks, the third level of proficiency from the top, out of 6 levels. No participants reported any hearing or speech impairments. Participants received monetary remuneration for their participation.

2.3. Procedure

This study adopted the same-different task to examine the intelligibility of locally time-reversed fast spoken words and pseudowords, following the procedure taken in Ishida, Samuel, and Arai (2016) [3]. As before, participants listened to a locally time-reversed word (or pseudoword) in a male voice followed by an intact word (or pseudoword) in a female voice, and judged if the first and second speakers said the same or different words. Here, the gender of the first and second speakers was intentionally changed, so that the participant's judgment of same-different would be based on the lexical activation, not the voice quality of the first and second speakers. There were four possible pairs in the trials: word-word (same), pseudoword-pseudoword (same), word-pseudoword (different), and pseudoword-word (different). Words and pseudowords were different only by one phoneme, and this subtle difference was intentionally created, in order to make the judgement of same-different difficult. The *d'* parameter of the signal detection theory was then computed [21]. The miss rates and false alarm rates were computed based on error responses: i.e., responding 'same' when the paired items were different, and 'different' when the paired items were the same. For each subject, *d'* was computed for each cell of the experimental design (stop-dominant words and pseudowords with 10, 20, 30, 40, 50, and 60 ms reversal windows). A miss rate or false alarm rate at ceiling or floor was replaced with the value of 1/2N (floor) or 1 – 1/2N (ceiling) in which N equals to the number of items.

In the experiment, participants listened to the stimuli facing a computer monitor, and responded 'same' or 'different' by pressing a button on a response pad. Each subject experienced 240 trials, comprised of 4 pairs x 60 items. The 60 items were divided into 6 subsets of 10 items, with each subset assigned to one of the 6 segment durations (10, 20, 30, 40, 50, and 60 ms). Six groups of participants, in a Latin square, were used to counterbalance the 6 subsets of items across the 6 segment durations. The 240 trials for each subject were individually randomized. The inter-stimulus-interval was 400 ms, and the total duration of experiment was approximately 15 minutes.

For native speakers, the experiment took place in a sound proof room of Stony Brook University in the United States. The stimuli were presented diotically over headphones (SONY MDR-V900HD) at a participant's comfortable listening level.

For non-native speakers, the experiment took place in a sound proof room of NTT Communication Science Laboratories in Japan. The stimuli were presented diotically over headphones (SONY MDR-CD900ST) at a participant's comfortable listening level.

3. Results

The current study explored how native and non-native speakers of English perceptually restore fast spoken stop-dominant words and pseudowords when locally time-reversed. This study examined the effects of a fast speech rate as well as lexical context and language proficiency on the intelligibility of locally time-reversed speech. The *d'* parameter of signal detection theory was computed for each participant, assuming that approx. *d'* = 1 is a good indication of intelligibility (i.e., performance was well above chance). The results (Figure 2) suggested that native speakers understood locally time-reversed fast spoken words really well (*d'* = 2.61, 2.50, 2.29, 2.01, 1.72, 1.15 respectively for 10, 20, 30, 40, 50, 60 ms), while struggling to understand locally time-reversed fast spoken pseudowords as the reversed segment length became longer (*d'* = 1.15, 0.83, -0.06, -0.10, -0.22, -0.11 respectively for 10, 20, 30, 40, 50, 60 ms). On the other hand, non-native speakers hardly understood locally time-reversed fast spoken words (*d'* = 0.97, 0.92, 0.81, 0.75, 0.68, 0.51 respectively for 10, 20, 30, 40, 50, 60 ms) as well as locally time-reversed fast spoken pseudowords (*d'* = 0.30, 0.18, 0.10, -0.46, -0.12, 0.25 respectively for 10, 20, 30, 40, 50, 60 ms). Overall, locally time-reversed fast spoken words were generally more intelligible than pseudowords, but the lexical advantage was much more evident for native speakers than for non-native speakers. Locally time-reversed fast speech was relatively intelligible for native speakers, but unintelligible for non-native speakers. The combination of natural articulatory distortion and artificial acoustic distortion could be overcome to a reasonable extent by native speakers, but not by non-native speakers.

An ANOVA was performed with language (native and non-native speakers of English) as a between-subject factor, and lexical status (word vs. pseudoword) and reversed segment length (10, 20, 30, 40, 50, or 60 ms) as within-subject factors. The results showed that perceptual restoration by native and non-native speakers was significantly different, F (1, 58) = 63.038, p < .001, partial η^2 = .52. While both native

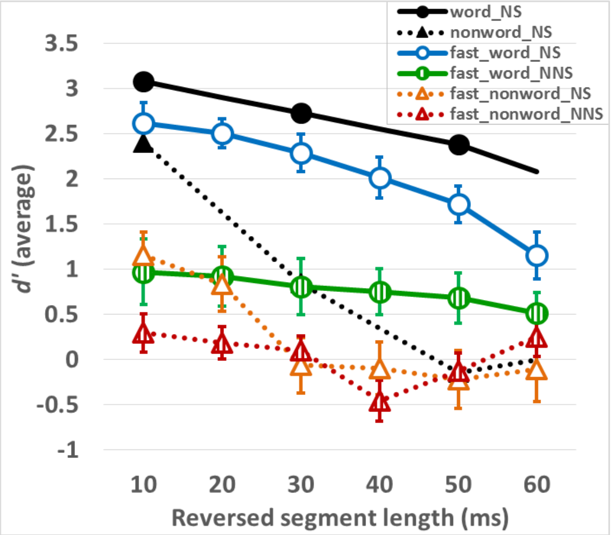

Figure 2: *The intelligibility of locally time-reversed fast speech in color (blank and striped circles and triangles), as compared to locally time-reversed normal speech in black from Ishida, Samuel, & Arai (2016). NS: native speakers of English. NNS: Non-native speakers of English (= native Japanese speakers who speak English as a second language).*

and non-native speakers perceptually restored locally time-reversed words better than locally time-reversed pseudowords, *F* (1, 58) = 132.093, *p* < .001, partial η^2 = .70, there was a significant interaction between the lexical status and language, *F* (1, 58) = 23.542, *p* < .001, partial η^2 = .29; the lexical advantage was more evident among native speakers than among non-native speakers. While the intelligibility of locally time-reversed speech dropped significantly for both native and non-native speakers when the reversed segment length increased, *F* (5, 290) = 50.674, *p* < .001, partial η^2 = .47, there was also a significant interaction between the reversed segment length and language, *F* (5, 290) = 18.884, *p* < .001, partial η^2 = .25; native speakers were much more tolerant of local time distortion than non-native speakers. There was also a significant interaction between the reversed segment length and lexical status, *F* (5, 290) = 13.166, *p* < .001, partial η^2 = .19; the intelligibility of words deteriorated significantly more than pseudowords when the reversed segment length increased, because of a floor effect for pseudowords. Finally, there was a significant three-way interaction among the reversed segment length, lexical status, and language, *F* (5, 290) = 2.918, *p* = .014, partial η^2 = .05; the locally time-reversed words were significantly more intelligible than pseudowords across all the reversed segment lengths, but this result was significantly more robust for native speakers than for non-native speakers.

4. Discussion

The combination of natural articulatory distortion (i.e., fast speech rate) and artificial acoustic distortion (i.e., local time reversal) severely impaired the perceptual restoration by native and non-native speakers, but native speakers were much more tolerant of heavy distortion than non-native speakers. In fact, fast spoken words were intelligible to native speakers even when the maximum distortion was imposed (*d'* = 1.15 when every 60 ms was flipped in time), while comparable intelligibility was barely observed among non-native speakers even when the temporal distortion was minimal (*d'* = 0.97 when every 10 ms was flipped in time). Moreover, the restorability of *pseudowords* by native speakers for the minimum distortion (*d'* = 1.15) was similar to the restorability of *words* by non-native speakers for the minimum distortion (*d'* = 0.97). The restorability of fast spoken *unfamiliar words in a first language* is, possibly, comparable to the restorability of relatively *familiar words in a second language*.

At the same time, lexical context supported perceptual restoration for both native and non-native listeners. However, the lexical advantage was more evident for native speakers than for non-native speakers: i.e., the gap of intelligibility between words and pseudowords was bigger among native speakers than among non-native speakers. It seems that the accessibility to lexical context determines the restorability of heavily distorted speech. The listener's mental lexicon seems to play a significant role for the perceptual restoration of locally time-reversed fast spoken words and pseudowords.

5. Conclusions

Simultaneous articulatory (i.e., fast speech) and acoustic distortion (i.e., local time reversal) surely impaired the intelligibility of speech. However, locally time-reversed "fast" speech was perceptually restorable to some extent, depending on context and the listener's language proficiency. The current study suggested that the intelligibility of locally time-reversed 'fast' speech gradually deteriorated as the length of reversed segments increased, as was also observed with locally time-reversed 'normal' speech in Ishida, Samuel, and Arai (2016) [3]. In addition, fast spoken words were much more intelligible than fast spoken pseudowords when locally time-reversed – lexical context strongly supported perceptual restoration. Overall, native speakers understood fast spoken words but struggled to understand pseudowords, while non-native speakers hardly understood both fast spoken words and pseudowords when locally time-reversed – language proficiency strongly supported perceptual restoration. It seems that the temporal sequence of speech can be jumbled up to some extent, as long as the constituents of speech stay within a certain time range from the original position on a temporal axis. It is possible that speech is processed chunk by chunk from the onset of speech, and the chunk of information is integrated for speech perception. Locally time-reversed speech would be intelligible when the temporal sequence of speech is jumbled within this chunk. The length of the chunk (i.e., a minimum processing unit) can differ depending on a language or the listener's language proficiency, but this remains as a question for the future study.

6. Acknowledgements

I would like to sincerely thank Dr. Takayuki Arai for his comments on the earlier draft. I also would like to thank Dr. Arty Samuel who kindly supported me to conduct research in Stony Brook University in the United States, and Dr. Makio Kashino who kindly supported me to conduct research in NTT Communication Science Laboratories in Japan. I also appreciate all the participants for this study. This study was supported by JSPS research grants (#15J00033 & #17J00285).

7. References

[1] K. Saberi, and D.R. Perrott, “Cognitive restoration of reversed speech”. *Nature*, vol. 398, pp. 760, 1999.

[2] S. Greenberg and T. Arai, “The relation between speech intelligibility and the complex modulation spectrum”, *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, pp. 473–476, 2001

[3] M. Ishida, A.G. Samuel, and T. Arai, “Some people are “more lexical” than others”, *Cognition*, vol. 151, pp. 68-75, 2016.

[4] C. Grataloup, M. Hoen, E. Veuillet, L. Collet, F. Pellegrino, and F. Meunier, “Speech processing: An interactive process”, *Journal of Speech, Language, and Hearing Research*, vol. 52, pp. 827-838, 2009.

[5] M. Kiss, T. Cristescu, M. Fink, and M. Wittmann, “Auditory language comprehension of temporally reversed speech signals in native and nonnative speakers”, *Acta Neurobiologiae Experimentalis*, vol. 68, no. 2, pp. 204–213, 2008.

[6] I. Magrin-Chagnolleau, M. Barkat, and F. Meunier, “Intelligibility of reverse speech in French: a perceptual study”, *Proceedings of the 7th. International Conference on Spoken Language Processing (Interspeech 2002)*, pp. 1669-1672, 2002.

[7] R.E. Remez, E.F. Thomas, K.R. Dubowski, S.M. Koinis, N.A.C. Porter, N.U. Paddu, M. Moskalenko, and Y.S. Grossman, “Modulation sensitivity in the perceptual organization of speech”, *Attention, Perception & Psychophysics*, vol. 75, pp. 1353-1358, 2013.

[8] J. Dalby, “Phonetic structure of fast speech in American English”, *Bloomington: Indiana University Linguistics Club*, 1986.

[9] K. Johnson, “Massive reduction in conversational American English”, In K. Yoneyama, and K. Maekawa, (ed.). *Spontaneous speech: Data and analysis* (pp. 29-54). Tokyo: The National Institute for Japanese Language, 2004.

[10] M.A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, “The ViC corpus of conversational speech”, Manuscript submitted to IEEE Transactions on Speech and Audio Processing: Special Issue on Spontaneous Speech Processing. 2003.

[11] M.A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, W. “The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability”, *Speech Communication*, vol. 45, pp. 90-95, 2005.

[12] M.A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, *Buckeye corpus of conversational speech* (2nd release) [www.buckeyecorpus.osu.edu]. Columbus, OH: Department of Psychology, Ohio State University (Distributor), 2007.

[13] J.D. Brown, and K. Kondo-Brown, (Eds.). (2006). *Perspectives on teaching connected speech to second language speakers*, Honolulu, HI: University of Hawai‘i, National Foreign Language Resource Center, 2006.

[14] J.D. Brown, and A.G. Hilferty, “The effectiveness of teaching reduced forms for listening comprehension”, *Paper presented at the TESOL Convention*, Honolulu, Hawai'i, 1982.

[15] J.D. Brown, and A.G. Hilferty, “The effectiveness of teaching reduced forms for listening comprehension”, *RELC Journal*, vol. 17, no. 2, pp. 59-70, 1985.

[16] J.D. Brown, and A.G. Hilferty, “Understanding reduced forms”, In D. Nunan (Ed.). *New ways in teaching listening* (pp. 124-127), Washington, DC: TESOL, 1995.

[17] M. Ishida, and T. Arai, “Perception of an existing and non-existing L2 English phoneme behind noise by Japanese native speakers”, In *INTERSPEECH-2015*, pp. 3408-3411, 2015.

[18] M. Ishida, and T. Arai, “Missing phonemes are perceptually restored but differently by native and non-native listeners”, *SpringerPlus*, vol. 5. No.1, pp. 1-10, 2016.

[19] J.C. Alderson, *Diagnosing foreign language proficiency: The interface between language learning and assessment*, London and New York: Continuum, 2006.

[20] Lancaster University, (2014, June 5), “DIALANG”, Retrieved from “https://dialangweb.lancaster.ac.uk/”.

[21] Abdi, H. “Signal detection theory (SDT)”. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 886-889). Thousand Oaks, CA: Sage, Inc., 2017.