The neural processing of phonemes is shaped by linguistic analysis

TOBIAS OVERATH[1,2,3,*] AND JACKSON C. LEE[1]

[1] *Duke Institute for Brain Sciences, Duke University, Durham, NC, USA*

[2] *Center for Cognitive Neuroscience, Duke University, Durham, NC, USA*

[3] *Department of Psychology and Neuroscience, Duke University, Durham, NC, USA*

Speech perception entails the mapping of the acoustic waveform to its linguistic representation. For this transformation to succeed, the speech signal needs to be tracked across multiple temporal scales in order to decode linguistic units ranging from phonemes to sentences. Here, we investigate how linguistic knowledge, and the temporal scale of linguistic analysis, influence the neural processing of a fundamental linguistic unit, the phoneme. To obtain control over the linguistic scale of analysis, we use a novel speech-quilting algorithm (Overath *et al.*, 2015) to control the acoustic structure available at different linguistic units (phoneme, syllable, word). To obtain control over the linguistic content, independent of the temporal acoustic structure, we construct speech quilts from both familiar (English) and foreign (Korean) languages. We recorded electroencephalography in healthy participants and show that the neural response to phonemes, the phoneme-related potential, is shaped by linguistic context only in a familiar language, but not in a foreign language. The results suggest that the processing of the acoustic properties of a fundamental linguistic unit, the phoneme, is already shaped by linguistic analysis.

INTRODUCTION

Speech is an intrinsically temporal signal with a rich temporal structure: Its linguistic constituents, such as phonemes, syllables, words, or sentences, all have characteristic durations, ranging from tens of milliseconds (in the case of phonemes) to hundreds (words) or thousands of milliseconds (sentences) (Rosen, 1992; Stevens, 2000; Poeppel, 2003). Our understanding of the neural architecture supporting speech perception has increased substantially over the last two decades (Hickok and Poeppel, 2007; Friederici and Gierhan, 2013). However, where and how the acoustic analysis of temporal speech structure interfaces with linguistic representations (such as syntax, lexicon, or semantics) is still poorly understood. While there is evidence that speech is analyzed at different temporal analysis scales, which are instantiated via a hierarchical organization across auditory and frontal cortices (Hasson *et al.*, 2008; Lerner *et al.*, 2011), these apply to relatively long temporal analysis windows

*Corresponding author: t.overath@duke.edu

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 6: Adaptive Processes in Hearing, August 2017, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C.-Dalsgaard, L. Tranebjærg, T. Andersen, and T. Poulsen. The Danavox Jubilee Foundation, 2017.
 ISBN: 978-87-990013-6-1.

commensurate with words, sentences, and paragraphs. In contrast, less is known about the neural representation of smaller linguistic units, in particular phonemes and syllables, which form the 'building blocks' upon which longer linguistic structures are built.

Previous studies have typically used isolated phonemes, consonant-vowel transitions, or words to investigate the underlying neural processes (Phillips *et al.*, 2000; Sanders and Neville, 2003; Tremblay *et al.*, 2003; Martin *et al.*, 2008). However, by presenting linguistic units in isolation, the role of predictive, top-down linguistic processes such as learned phonological, morphological or syntactical rules (Park *et al.*, 2015; Kocagoncu *et al.*, 2017), which are ubiquitous and automatic in natural speech perception, remained unclear. More recently, Khalighinejad *et al.* (2017) used continuous, natural speech to demonstrated that different categories of phonemes (e.g., vowels, nasals, fricatives, or plosives) have distinct neural correlates, or phoneme-related potentials (PRP). However, this approach is unable to differentiate between acoustic and linguistic processes, since listening to natural speech in a familiar language automatically recruits both.

What is needed, therefore, is an experimental approach that dissociates acoustic from linguistic processes during the analysis of temporal speech structure. Such an approach requires two essential features: (1) control over the linguistic structure, or units at which analysis occurs; (2) control over the linguistic content. We propose the following paradigm that allows the dissociation of acoustic and linguistic speech processes: To obtain control over the linguistic units of analysis, we modify a novel sound-quilting algorithm (Overath *et al.*, 2015) to control acoustic structure at the level of different linguistic units (phonemes, syllables, words) by shuffling and then stitching them together. This approach yields new 'speech quilts' that preserve the natural temporal speech structure only up to the linguistic unit, but not beyond. To obtain control over the linguistic content, independent of the temporal acoustic structure of linguistic units, we construct speech quilts from both familiar (English) and foreign (Korean) languages. This approach ensures that any changes at the signal-acoustics level affect both languages identically, while manipulating the linguistic percept differently. Thus, neural responses that vary as a function of the size of the linguistic unit (phoneme, syllable, word) will imply the presence of linguistic processing, while neural responses that are unaffected by linguistic unit will imply aspects of acoustic processing.

In this study, we investigated how acquired linguistic knowledge influences the neural processing of a fundamental linguistic unit, the phoneme, in different contexts. We recorded electroencephalography (EEG) from participants while they listened to speech quilts carrying information at the level of phonemes, syllables, or words, as well as natural speech, in either a familiar (English) or foreign language (Korean). We hypothesized that the PRP would be modified as a function of linguistic context only in a familiar language, due to linguistic processes, but not in a foreign language.

METHODS

Participants

The 18 right-handed participants (mean age = 23, range = 18-31, 10 females) were native speakers of American English, with no knowledge of Korean. All reported to have normal hearing and no history of neurological or psychiatric diseases. Participants provided written consent prior to participating in the study, in accordance with the Duke University Institutional Review Board.

Stimuli

The stimuli were derived from mono recordings (44100-Hz sampling rate, 16-bit resolution) of four female bilingual English/Korean speakers reading from a book in either language (native English and native Korean speakers judged the recordings as coming from native speakers, respectively). The recordings were then segmented into phonemes, syllables, and words using the Penn Phonetics Lab Forced Aligner Toolkit (Yuan and Lieberman, 2008) for English, and the Korean Phonetic Aligner Program Suite (Yoon and Kang, 2013) for Korean. The alignment was then manually checked for eventual segmentation errors by a native English and Korean speaker, respectively. Korean is a phonetic language that shares no etymological roots with English and has a different grammatical structure (Sohn, 1999).

We placed a number of constraints on the quilted stimuli. 1) Phonemes had to be between 20-80 ms in duration, syllables between 100-240 ms, and words between 300-600 ms to be included in the phoneme, syllable, or word quilts, respectively. 2) Syllables that were also words were excluded from consideration in the syllable quilts. 3) Two identical phonemes could not be next to each other, since this does not happen in normal English or Korean speech. The relative phoneme distribution (frequency of occurrence of a given phoneme across conditions) was not affected by these constraints: Phoneme frequency profiles within a language were significantly correlated ($0.85 < \rho < 0.99$, all $p < 0.001$).

The stimuli are based on a slight modification of the quilting algorithm introduced in Overath *et al.* (2015), such that instead of quilting equal-length segments, here we quilt linguistic units. Briefly, a source signal is divided into linguistic units (here either phonemes, syllables, or words), which are then pseudorandomly rearranged and stitched together to create a new speech quilt signal. By using an L^2 norm when choosing adjacent linguistic units to approximate the original unit-to-unit change in the original speech signal, and by using pitch-synchronous overlap-add (PSOLA) (Moulines and Charpentier, 1990) to avoid sudden frequency jumps at unit boundaries, the quilting algorithm ensures that low-level acoustic attributes (e.g., amplitude modulation rate, frequency spectrum) in the speech quilt are similar to those in the original speech signal. All stimuli are 6 s long and are speech quilts made up of phonemes, syllables, and words, as well as original, unaltered excerpts from the recordings.

Experimental procedure

Participants were familiarized with recordings of the four different speakers, and then performed a behavioral task in which they listened to brief recordings in English or Korean and were asked to identify the speaker for each trial (irrespective of language). Participants responded by pressing one of keys 1, 2, 3, or 4 for speakers 1-4.

In the EEG experiment, each condition of a 2 Language (English, Korean) × 4 Linguistic unit (phoneme-, syllable-, word-quilt, natural speech) design was presented a total of 48 times (12 exemplars per speaker) over the course of 4 runs. The inter-trial-interval was 2 s. Participants performed the same speaker identification task as in the prior behavioral experiment (irrespective of language and linguistic unit).

Stimuli were presented at a comfortable listening level (~60 dB) through high-fidelity Sennheiser HD-25 on-ear headphones via a low-latency Fireface UC USB sound card, using Psychophysics Toolbox Version 3 (Brainard, 1997) running in Matlab.

EEG recording and analysis

EEG data were recorded on a 63-channel active electrode system (Brain Vision ActiChamp, Brain Products) using a customized, extended coverage, elastic electrode cap (EASYCAP, Herrsching, Germany) (Woldorff *et al.*, 2002). This cap provides extended coverage of the head from just above the eyebrows to below the inion posteriorly and has electrodes that are equally spaced across the cap. Two fronto-lateral electrodes track horizontal eye movements, while an additional external electrode just underneath the left eye tracks vertical eye movements. Data are recorded at a 1000-Hz sampling rate (with a DC to 260 Hz bandpass) referenced to the right mastoid, and are then re-referenced off-line to the average of the left and right mastoids.

Data were analyzed using EEGLAB (Delorme and Makeig, 2004) and custom-written Matlab scripts. Standard artifact rejection algorithms and independent component analysis (ICA) implemented in EEGLAB were used to remove eye-blink and physiological noise artifacts. The PRP analysis largely followed that outlined in Khalighinejad et al. (2017): Data were z-scored, epoched between −100 and 600 ms relative to phoneme onset, baseline corrected (−100 to 0 ms), and bandpass filtered between 2-15 Hz. To determine time windows of interest for our subsequent analyses, we centered 50-ms windows around the P50, N100, and P200 peaks derived from the PRP across languages and all electrodes. Results are shown for a region-of-interest (ROI) containing 9 fronto-central electrodes around electrode FCz.

RESULTS

Behavioral performance in the speaker identification task improved with linguistic unit length, and interacted with language familiarity (Fig. 1): A repeated measures (RM) ANOVA revealed main effects of Linguistic unit ($F_{(3,51)} = 12.87$, $p < 0.001$, $\eta^2_p = 0.43$), Language ($F_{(1,17)} = 9.49$, $p = 0.007$, $\eta^2_p = 0.36$), and an interaction ($F_{(3,51)} = 4.51$, $p = 0.007$, $\eta^2_p = 0.21$). Post-hoc pairwise comparisons (Bonferroni

corrected) revealed that performance was significantly better in English than in Korean, except in the phoneme quilt condition ($p > 0.05$).

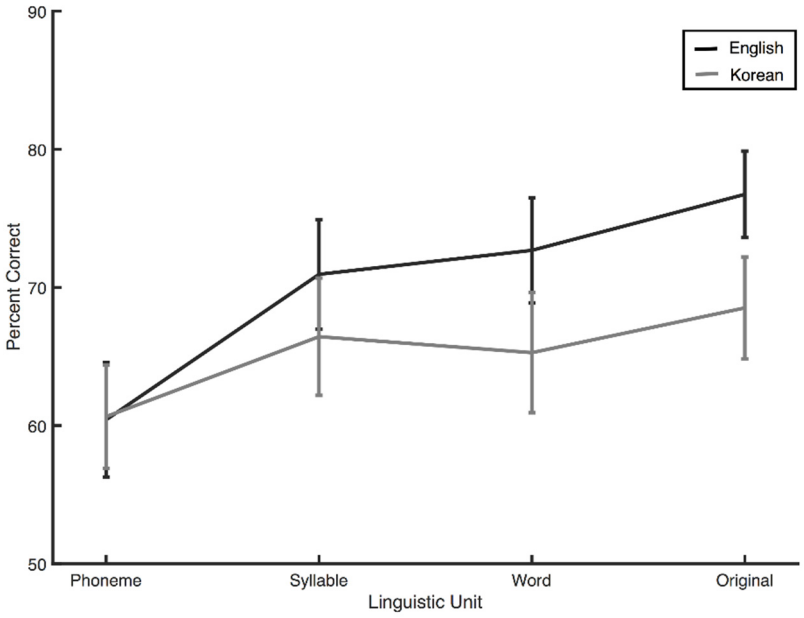

Fig. 1: Mean percent correct performance (±SEM) in the speaker identification task. Performance was well above chance (25%).

Next, we computed the grand-average PRP across all phonemes. As in Khalighinejad *et al.* (2017), the PRP showed a clear succession of P50 and N100 components, as well as a weak P200, in both English and Korean (Fig. 2). For each component, we ran RM ANOVAs with factors Language (English, Korean) and Linguistic Unit (phoneme-, syllable- , word-quilts, and natural speech; Table 1). The P50 component revealed main effects for both factors, as well as an interaction. The N100 component showed main effects for both factors, while the P200 component revealed a main effect of Linguistic Unit and an interaction.

	P50 (35-85 ms)			**N100 (90-140 ms)**			**P200 (180-230 ms)**		
	Unit	**Lang.**	**Inter.**	**Unit**	**Lang.**	**Inter.**	**Unit**	**Lang.**	**Inter.**
F-value	15.79	5.53	4.37	3.71	12.43	n.s.	3.23	n.s.	3.21
p-value	< 0.001	0.031	0.008	0.017	0.003	n.s.	0.03	n.s.	0.031
η^2_p	*0.48*	*0.25*	*0.2*	*0.18*	*0.42*	n.s.	*0.16*	n.s.	*0.16*

Table 1: RM ANOVA with factors Linguistic Unit and Language. F-value degrees of freedom are (3,51) for Linguistic Unit and (1,17) for Language factors.

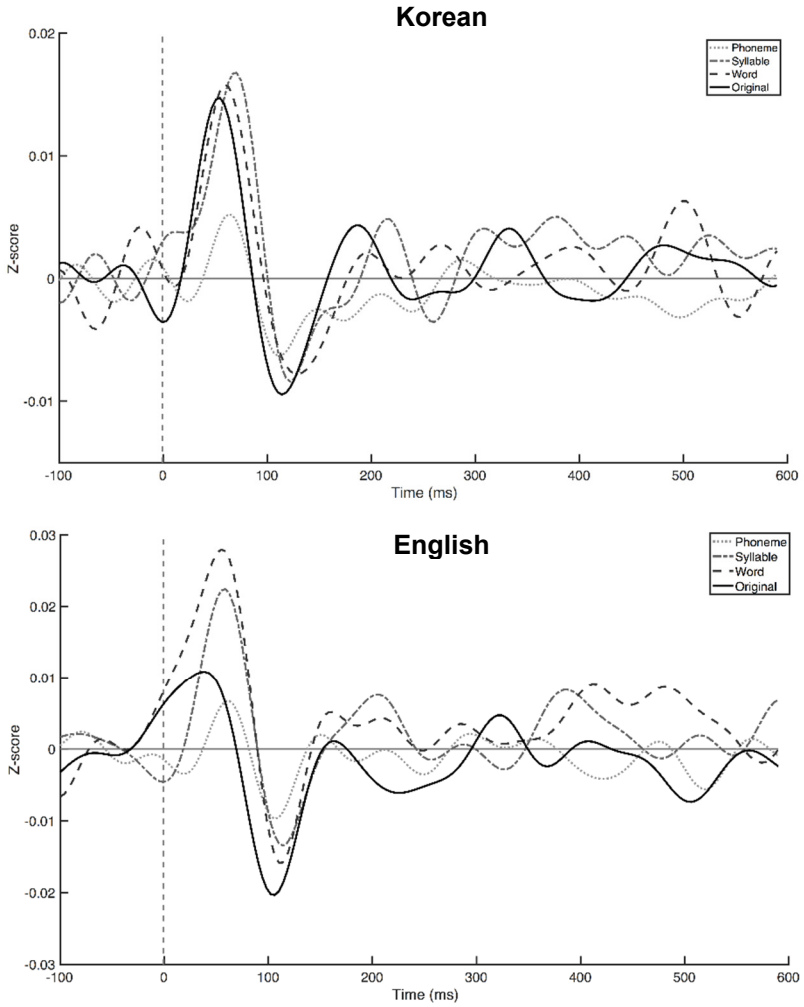

Fig. 2: The phoneme-related potential (PRP) for Korean (top) and English (bottom) conditions (phoneme-, syllable-, and word quilts, as well as natural speech). Note that in Korean the N100 component of the PRP is not significantly affected by the linguistic context; In English, the strength of the N100 component increases from phoneme-quilts to natural speech.

To investigate in more detail the effect of linguistic unit in English and Korean, we computed RM ANOVAs separately for each language. In English, P50, N100, and P200 differed in magnitude as a function of Linguistic Unit ($F_{(3,51)} = 12.44$, $p < 0.001$, $\eta^2_p = 0.42$; $F_{(3,51)} = 3.86$, $p = 0.015$, $\eta^2_p = 0.19$; and $F_{(3,51)} = 3.75$, $p = 0.016$, $\eta^2_p = 0.16$, respectively). The N100 revealed the clearest effect of linguistic unit, whereby its absolute magnitude increased monotonically from phoneme quilts to natural speech. In Korean, only the early P50 component was affected by linguistic unit ($F_{(3,51)} = 5.05$, p = 0.004, $\eta^2_p = 0.23$); However, post-hoc pairwise comparisons revealed that this was driven by a more categorical, rather than graded, effect, whereby the P50 in phoneme quilts was significantly different from any of the other conditions.

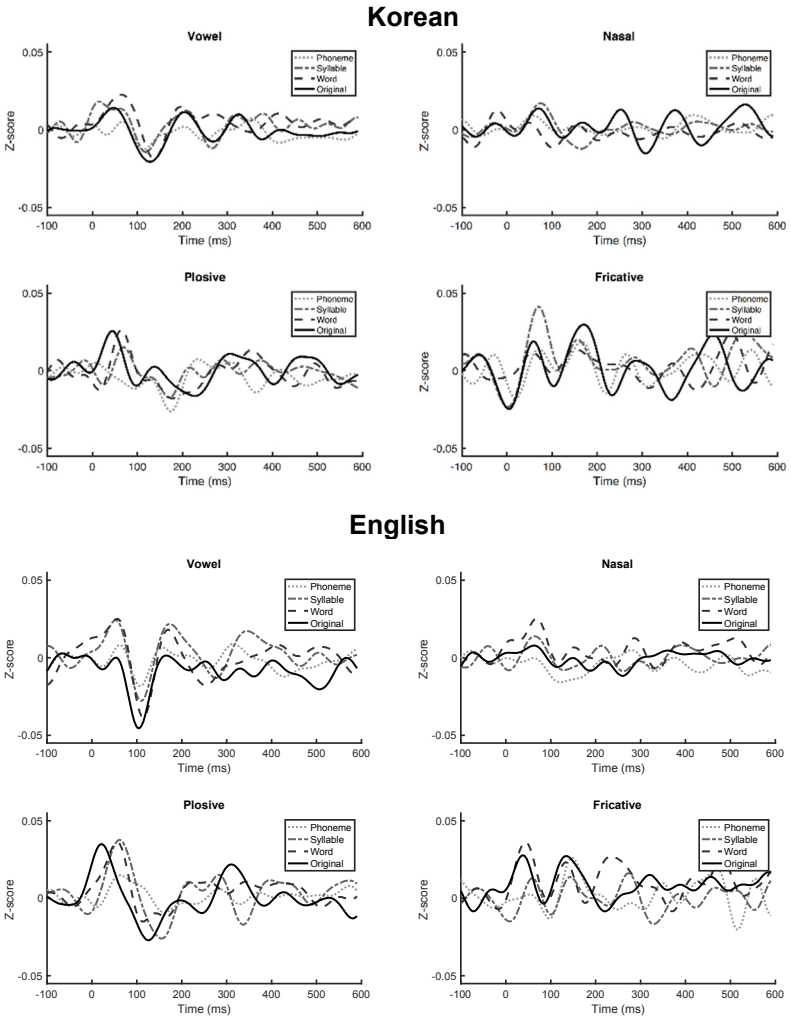

Fig. 3: Category-specific (vowel, nasal, plosive, fricative) PRPs as a function of linguistic unit (phoneme-, syllable-, word-quilt, natural speech) in Korean (top) and English (bottom). Note the overall similarity between languages for the original speech conditions. In English, the N100 component of the vowel PRP shows the clearest gradated PRP.

The analyses so far have treated all phonemes the same; however, phonemes can be classified by their manner of articulation (Ladefoged and Johnson, 2010), and we next investigated the main classes of plosives, fricatives, nasals, and vowels, to determine whether different phoneme categories are differentially affected by the linguistic context as a function of language. Figure 3 shows the PRP for each phoneme category in each language as a function of linguistic unit (phoneme-, syllable-, or word-quilt, as well as natural speech). We focus here on the N100 component, which showed a significant effect of linguistic unit in English. In English, the vowel PRP showed a

clear graded response ($F_{(3,51)}$ = 4.45, p = 0.007, η^2_p = 0.21); Post-hoc pairwise comparisons (Bonferroni corrected) revealed significant differences between the phoneme-quilt and natural speech conditions (p = 0.025), as well as a tendency between phoneme-quilt and word-quilt conditions (p = 0.055). The other three phoneme categories did not show a graded response as a function of linguistic unit. In Korean, no phoneme category was affected by linguistic unit size in any systematic way.

DISCUSSION

The preliminary results reported here demonstrate that the processing of a fundamental linguistic unit, the phoneme, is already shaped by linguistic analysis, but only if a linguistic repertoire is available. In the familiar language, the phoneme-related potential showed a graded N100 response as the size of the linguistic unit increased (from phoneme quilts to normal speech); This was most pronounced for vowels. In contrast, the PRP was generally unaffected by linguistic context in a foreign language.

The design of directly comparing the effect of linguistic unit size in two languages allowed the dissociation of acoustic and linguistic neural processes. Acoustic processing would be shared between languages, while linguistic processing would be indicated by a differentiation of the response (e.g., PRP) as a function of linguistic context. In the current study, the similarity of the PRP in natural English and Korean speech (e.g., Fig. 2) therefore reveals a shared mechanism for processing acoustic properties that are common to phonemes in both languages. For example, the vowel PRP in both English and Korean displayed a characteristic N100 similar to that in Khalighinejad *et al.* (2017) (Fig. 3). In contrast, the linguistic context within which phonemes appeared influenced the PRP in a systematic manner only in the familiar, but not the foreign language. This suggests that a linguistic repertoire (e.g., syntax, lexicon, or semantics), when available, shapes the processing of acoustic properties of temporal speech structure, even at a fundamental level such as the phoneme.

The results have implications for our understanding of how acoustic and linguistic representations interface already at an early level of speech processing. For example, difficulties in speech perception in children with developmental dyslexia (Molinaro *et al.*, 2016), or older adults with hidden hearing loss (Plack *et al.*, 2014), might arise from a compromised acousto-linguistic transformation at fast temporal scales such as those of phonemes. More generally, the results inform speech and language models that need to explain a fundamental question in speech perception: Where and how the analysis of the acoustic speech signal is transformed into linguistic representations that enable speech comprehension.

ACKNOWLEDGEMENTS

The authors would like to thank Frankie Pennington and Joon Hyun Paik for manually checking phoneme, syllable, and word onsets and offsets for English and Korean recordings, respectively.

REFERENCES

Brainard, D.H. (**1997**). "The psychophysics toolbox," Spat. Vis., **10**.

Delorme, A., and Makeig, S. (**2004**). "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics," J. Neurosci. Methods, **134**, 9-21.

Friederici, A.D., and Gierhan, A.M.E. (**2013**). "The language network," Curr. Opin. Neurobiol., **23**, 250-254.

Hasson, U., Yang, E., Vallines, I., Heeger, D.J., and Rubin, N. (**2008**). "A hierarchy of temporal receptive windows in human cortex," J. Neurosci., **28**, 2539-2550.

Hickok, G., and Poeppel, D. (**2007**). "The cortical organization of speech processing," Nat. Rev. Neurosci., **8**, 393-402.

Khalighinejad, B., da Silva, G.C., and Mesgarani, N. (**2017**). "Dynamic encoding of acoustic features in neural responses to continuous speech," J. Neurosci., **37**, 2176-2185.

Kocagoncu, E., Clarke, A., Devereux, B.J., and Tyler, L.K. (**2017**). "Decoding the cortical dynamisc of sound-meaning mapping," J. Neurosci., **37**, 1312-1319.

Ladefoged, P., and Johnson, K. (**2010**). *A Course in Phonetics*. Boston: Wadsworth.

Lerner, Y., Honey, C.J., Silbert, L.J., and Hasson, U. (**2011**). "Topographic mapping of a hierarchy of temporal receptive windows using a narrated story," J. Neurosci., **31**, 2906-2915.

Martin, B.A., Tremblay, K.L., and Korczak, P. (**2008**). "Speech evoked potentials: from the laboratory to the clinic," Ear Hearing, **29**, 285-313.

Molinaro, N., Lizarazu, M., Lallier, M., Bourguignon, M., and Carreiras, M. (**2016**). "Out-of-synchrony speech entrainment in developmental dyslexia," Hum. Brain Mapp., **37**, 2767-2783.

Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun., **9**, 453-467.

Overath, T., McDermott, J.H., Zarate, J.M., and Poeppel, D. (**2015**). "The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts," Nat. Neurosci., **18**, 903-911.

Park, H., Ince, R.A., Schyns, P.G., Thut, G., and Gross, J. (**2015**). "Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners," Curr. Biol., **25**, 1649-1653.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., McGinnis, M., and Roberts, T. (**2000**). "Auditory cortex accesses phonologcal categories: an MEG mismatch study," J. Cogn. Neurosci., **12**, 1038-1055.

Plack, C.J., Barker, D., and Prendergast, G. (**2014**). "Perceptual consequences of "hidden" hearing loss," Trends Hear., **18**, 1-11.

Poeppel, D. (**2003**). "The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'," Speech Commun., **41**, 245-255.

Rosen, S. (**1992**). "Temporal information in speech: acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. Lond. B. Biol. Sci., **336**, 367-373.

Sanders, L.D., and Neville, H.J. (**2003**). "An ERP study of continuous speech processing: I. Segmentation, semantics, and syntax in native speakers," Brain Res. Cogn. Brain Res., **15**, 228-240.

Sohn, H.-M. (**1999**). *The Korean Language*. Cambridge: Cambridge University Press.

Stevens, K.N. (**2000**). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Tremblay, K.L., Friesen, L., Martin, B.A., and Wright, R. (**2003**). "Test-retest reliability of cortical evoked potentials using naturally produced speech sounds," Ear Hearing, **24**, 225-232.

Woldorff, M.G., Liotti, M., Seabolt, M., Busse, L., Lancaster, J.L., and Fox, P.T. (**2002**). "The temporal dynamics of the effects in occipital cortex of visual-spatial selective attention," Brain Res. Cogn. Brain Res., **15**, 1-15.

Yoon, T.-J., and Kang, Y. (**2013**). "The Korean Phonetic Aligner Program Suite," http://korean.utsc.utoronto.ca/kpa/

Yuan, J., and Lieberman, M. (**2008**). "Speaker identification on the SCOTUS corpus," J. Acoust. Soc. Am., **123**, 3878.