Published in final edited form as:
Neuroimage. 2021 July 15; 235: 117887. doi:10.1016/j.neuroimage.2021.117887.

From acoustic to linguistic analysis of temporal speech structure: Acousto-linguistic transformation during speech perception using speech quilts

Tobias Overath[a,b,c,*], **Joon H. Paik**[b]

[a]Department of Psychology & Neuroscience, Duke University, Durham, North Carolina, 27708, U.S.A.

[b]Duke Institute for Brain Sciences, Duke University, Durham, North Carolina, 27708, U.S.A.

[c]Center for Cognitive Neuroscience, Duke University, Durham, North Carolina, 27708, U.S.A.

Abstract

Speech perception entails the mapping of the acoustic waveform to linguistic representations. For this transformation to succeed, the speech signal needs to be tracked over various temporal windows at high temporal precision in order to decode linguistic units ranging from phonemes (tens of milliseconds) to sentences (seconds). Here, we tested the hypothesis that cortical processing of speech-specific temporal structure is modulated by higher-level linguistic analysis. Using fMRI, we measured BOLD signal changes to 4 s long speech quilts with variable temporal structure (30, 120, 480, 960 ms segment lengths), as well as natural speech, created from a familiar (English) or foreign (Korean) language. We found evidence for the acoustic analysis of temporal speech properties in superior temporal sulcus (STS): the BOLD signal increased as a function of temporal speech structure in both familiar and foreign languages. However, activity in left inferior gyrus (IFG) revealed evidence for linguistic processing of temporal speech properties: the BOLD signal increased as a function of temporal speech structure only in familiar, but not in foreign speech. Network connectivity analyses suggested that left IFG modulates the processing of temporal speech structure in primary and non-primary auditory cortex, which in turn sensitizes the analysis of temporal speech structure in STS. The results thus suggest that acousto-linguistic transformation of temporal speech structure is achieved by a cortical network comprising primary and non-primary auditory cortex, STS, and left IFG.

[*]Corresponding author. t.overath@duke.edu (T. Overath).

Credit authorship contribution statement

Tobias Overath: Conceptualization, Investigation, Formal analysis, Writing – original draft, Writing – review & editing. **Joon H. Paik:** Investigation, Formal analysis, Writing – review & editing.

Overath Paik data code availability

Data and code used for the results reported in the manuscript will be made available upon reasonable request.

Declaration of Competing Interest

The authors declare no financial or other conflicts of interest.

1. Introduction

Speech perception entails the mapping of the acoustic waveform to linguistic representations (Poeppel et al., 2008, Kleinschmidt and Jaeger, 2015). Despite the development of detailed speech/language models (Hickok and Poeppel, 2007, Rauschecker and Scott, 2009, Friederici and Gierhan, 2013), the mechanisms of this acousto-linguistic transformation – how different acoustic properties of the speech signal are processed throughout the auditory system, and how they interface with linguistic representations such as syntax and semantics to enable communication – are still not fully characterized.

A candidate link from the analysis of acoustic speech structure to the linguistic processes underlying speech comprehension is the rich temporal structure of speech, which carries critical information for speech intelligibility (Shannon et al., 1995, Smith et al., 2002). For example, linguistic information is conveyed over multiple temporal scales, or temporal windows (Poeppel et al., 2008, Poeppel, 2003, Rosen, 1992, Stevens, 2000): phonemes have an average duration of 30–60 ms, syllables have average durations of 150–300 ms, words are generally longer still, and so forth. Thus, while some linguistic processes like phonetic categorization require the analysis of brief temporal properties in the acoustic signal, speech comprehension ultimately requires the integration over longer temporal windows to extract syntactic, semantic, and lexical information. In this context, the term 'linguistic processes' is construed broadly, encompassing processes underlying the analysis of syntax, semantics, or lexical access.

Recently, we introduced a novel algorithm that controls the temporal extent of natural speech structure via randomizing and then 'quilting' back together speech segments of a set duration (Overath et al., 2015). We showed that, whereas earlier processing centers in human auditory cortex – such as Heschl's gyrus, (HG), part of primary auditory cortex, or planum temporale (PT), a computational hub receiving information from primary auditory cortex (Griffiths and Warren, 2002, Kumar et al., 2007, Overath et al., 2007) – are not sensitive to the temporal speech structure, subsequent processing in the superior temporal sulcus (STS) increased as a function of segment length or natural temporal speech structure. This was specific to speech sounds and did not generalize to non-speech control sounds that shared many of the low-level acoustic features of speech (such as slow amplitude modulations) or other environmental sounds. Importantly, however, the acoustic manipulation (temporal speech structure via speech quilting) was performed in a language that was foreign to participants so as to focus on the analysis of acoustic temporal speech structure, independent of linguistic processes such as lexical-semantic and syntactic analyses. The approach could therefore not distinguish between acoustic and linguistic processes as they relate to the analysis of temporal speech structure.

This requires the dissociation of acoustic versus linguistic processes, which can be achieved by comparing the same acoustic manipulation in familiar and foreign languages. We use the term 'acoustic processes' here to refer to the neural encoding of acoustic speech properties, such as harmonicity, sound energy onset, or high-frequency bursts (which are important characteristics of vowels, plosives, or fricatives, respectively; Stevens, 2000). In contrast, 'linguistic processes' denote the lexical, syntactic, or phonological analyses of the acoustic signal that enable speech comprehension (Poeppel et al., 2008, Giraud and Poeppel, 2012). Controlling both the temporal scale of analysis and the linguistic content in one paradigm ensures that any signal manipulations will affect acoustic properties of the speech signal similarly in both languages; in contrast, such signal manipulations will affect linguistic processes only in the familiar language.

To date, only a few studies have taken this approach. For example, listeners are able to track hierarchical linguistic structure based on syntax and semantics only in a familiar, but not in a foreign language (Ding et al., 2015); this is also reflected in oscillatory entrainment to natural familiar speech, which is strongest in the delta band (Pérez et al., 2015). However, since most of these studies only investigated effects of language familiarity in either continuous speech (Ding et al., 2015, Pérez et al., 2015, Peña and Melloni, 2012) or intact single words (Strelnikov et al., 2011), they were unable to reveal which temporal scales in the speech signal are critical for, and amenable to, linguistic analysis.

Here we utilize speech quilting of familiar and foreign languages to map how processing temporal speech structure proceeds from acoustic analysis to linguistic analysis. We achieve this by simultaneously controlling (1) the temporal scale at which analysis occurs (via speech quilting) and (2) the linguistic content (via two different languages; one native, the other foreign). This ensures that neural responses that vary as a function of segment length, but are shared or similar for the two languages, represent an analysis at the signal-acoustics level; however, neural responses that differ based on language familiarity indicate the presence of linguistic processing.

2. Methods

2.1. Participants

The 21 participants (mean age = 23.57, range = 19–28, 9 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. Two participants were excluded from further analysis: one participant performed at chance for the speaker identification task in the scanner, while the other only completed 3 runs, leaving a total of 19 participants (mean age = 23.68, range = 19–28, 9 females). Participants provided written consent prior to participating in the study in accordance with the Duke University Health System Institutional Review Board.

2.2. Stimuli

Sounds were derived from recordings (44100 Hz sampling rate, 16 bit resolution) of four perfectly bilingual female English/Korean speakers reading from a book in either language (native English and Korean speakers judged the recordings as coming from native speakers). This ensured that voice cues are unavailable to differentiate between languages. The recordings were then used as source material for the quilting algorithm, following the same procedures as outlined in Overath et al. (2015). Briefly, a source signal is divided into equal-length segments, which are then pseudo-randomly rearranged, or stitched together, to create a new speech quilt signal. By using an L^2 norm when choosing adjacent segments to approximate the original segment-to-segment change in the original speech signal, and by using pitch-synchronous overlap-add (PSOLA; Moulines and Charpentier, 1990) to avoid sudden frequency jumps at segment 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 (see also Overath et al., 2015). For both languages (English and Korean), the stimuli of the 5 experimental conditions were 4 s long speech quilts made up of 30 ms, 120 ms, 480 ms, or 960 ms speech segments, as well as 4 s long original, unaltered excerpts from the recordings. The choice of segment lengths sub-samples those used in Overath et al. (2015), while also allowing a confirmation of the response plateau at ~500 ms for foreign speech, and testing its validity for speech-specific processing in a familiar language. Fig. 1 displays cochleograms of the 30 ms segment speech quilt and original speech conditions for English and Korean.

2.3. Experimental design

Prior to the main experiment in the scanner, participants were familiarized with the four speakers in a behavioral experiment. Trials consisted of original, unaltered 4 s long recordings of the four speakers and were presented via Sennheiser HD 380 Pro headphones using Psychophysics Toolbox (Brainard, 1997) in Matlab. During the first couple runs, participants saw the speaker identity while they listened to each trial (e.g. ‘Speaker 1’) on the monitor. During subsequent runs, participants identified the speaker identity for each trial via pressing keys 1–4 on the keyboard without any visual cues; feedback was provided for each trial (correct, wrong). Each individual run took approximately 4 min, and participants needed to reach at least 90% correct performance to proceed to the fMRI experiment.

The 10 experimental conditions from a 2 Language (English, Korean) × 5 Segment length (30 ms, 120 ms, 480 ms, 960 ms, Original) factorial design were presented in eight “runs”, each lasting ~6.5 min. Stimuli were presented in a pseudo-randomized fashion that boosted contrast selectivity, e.g. by ensuring that presentations of the 30 ms segment length speech quilt and original speech conditions of each language were close together in time (not more than two trials apart; contrasts between trials that are far apart in time can potentially be affected by the high-pass filter, see below). Each experimental condition was presented 32 times per scanning session (in addition to 32 silent trials of 4 s duration) with a mean inter-stimulus interval of 4 s (randomly sampled from a uniform distribution between 3–5 s). All stimuli were unique and were presented only once. Stimuli inside the scanner were presented using Psychophysics Toolbox (Brainard, 1997) running in Matlab at a comfortable listening level (~75 dB SPL) at 44100 Hz sampling rate and 16 bit resolution via Sensimetrics (www.sens.com) MRI-compatible insert earphones (Model S14); participants wore protective earmuffs to further reduce the background noise of the scanner environment.

In the scanner, participants performed the speaker identity task by pressing one of four buttons on an MRI-compatible button box to indicate which speaker they had heard for each trial; while participants had been trained on the original, unaltered 4 s long recordings, stimuli in the scanner also included the speech quilt stimuli. Participants were instructed to only register their response after the sound had ended (to avoid confounding the BOLD signal response with a motor execution response), and were given feedback on each trial (correct, wrong, missed) and for each run (overall percentage correct).

2.4. Image acquisition

Data were recorded on a GE MR750 3.0 Tesla scanner using an 8-channel head coil and a high-resolution echo-planar imaging (EPI) sequence yielding contiguous isotropic 2 × 2 × 2 mm voxels (110 × 110 matrix, FOV = 22, TE = 28 ms, flip angle = 90°, TR = 2.2 s). 36 slices were acquired for each volume in an interleaved ascending sequence to avoid signal bleeding between adjacent slices. The volume was centered on STG and spanned from the inferior colliculus (IC) to inferior frontal gyrus (IFG). A high-resolution 1 × 1 × 1 mm voxel-size T1-weighted MRI (FSPGR) scan (TR/TE: 2,089/3.18 ms, FOV: 256) was acquired for each participant to inform structure-function mapping.

2.5. Data analysis

Imaging data were analyzed using Statistical Parametric Mapping software (SPM12, http://www.fil.ion.ucl.ac.uk/spm). The first four of the 174 volumes in each run were discarded to control for T1 saturation effects. The remaining 1360 scans were realigned to the first volume in the first run, un-warped to correct for motion artifacts and re-sliced using sinc interpolation (SPM12, “Realign and Unwarp”), and slice time corrected to account for differences in slice acquisition time (SPM12, “Slice timing”); the structural scan of each participant was coregistered to the mean functional scan (SPM12, “Coregister”) and segmented into grey and white matter and cerebro-spinal fluid and spatially normalized to standardized stereotaxic MNI space (SPM12, “Segment”), before applying the resulting linear transformations to the EPIs and structural scan (SPM12, “Normalize: Write”). Finally, the EPIs were spatially smoothed to improve the signal-to-noise ratio using an isotropic 6 mm full-width at half-maximum (FWHM) Gaussian kernel (SPM12, “Smooth”).

The design matrix for each participant consisted of 10 regressors (corresponding to the 10 experimental conditions), which were derived by convolving the stimulus (modeled as a four-second box-car function) with SPM’s canonical hemodynamic response function. The silent periods were not modeled explicitly. Data were high-pass filtered at 1/256 Hz to remove slow drifts in the signal.

Standard whole-brain second-level group analyses in SPM were based on a random-effects (RFX) model within the context of the general linear model (Friston et al., 1995). For second-level group analyses, the smoothing of first-level functional contrast images was increased to an effective 8 mm FWHM Gaussian kernel to better allow for inter-individual anatomical variation.

The results were further investigated in anatomically and functionally defined regions-of-interest (ROI and fROI, respectively). Two cortical anatomical ROIs in HG (encompassing primary auditory cortex) and PT (part of non-primary auditory cortex) were based on published probability maps in Rademacher et al. (2001) and Westbury et al. (1999), respectively. Both ROIs were thresholded such that they only included voxels with at least 30% probability of belonging to either structure (see also Overath et al., 2015). Two subcortical anatomical ROIs in auditory structures in IC and the medial geniculate body

(MGB) were spherical ROIs (with a radius of 5 mm) centered on published coordinates of these structures ([−6 −34 −12] and [6 −34 −12] for IC (Griffiths et al., 2001); [−16 −28 −8] and [16 −28 −8] for MGB (Devlin et al., 2006)). For ROI analyses in these subcortical structures, the data were not smoothed.

The BOLD signal in ROIs was calculated using MarsBaR (Brett et al., 2002). Since the absolute level of BOLD signal varied between participants and ROIs, we normalized the BOLD signal with respect to the original speech condition for better comparison (see also Overath et al., 2015): for a given participant, the BOLD signal in each run to the five conditions of either language (English, Korean) was normalized to (i.e. divided by) the mean BOLD signal of their original, unaltered condition in the other 7 runs. For example, for the English speech conditions, the BOLD signal for the Eng30ms, Eng120ms, Eng480ms, Eng960ms, and EngOrig conditions in run 1 was normalized to the mean BOLD signal for EngOrig in runs 2 through 8. This was then averaged across runs for each participant. Such a leave-one-out cross-validation procedure ensures that computations are performed on independent data and results are not over-inflated, e.g. because of ‘double-dipping’ (Kriegeskorte et al., 2009).

Functional ROIs were determined separately for English and Korean via [English original > English 30 ms quilts] and [Korean original > Korean 30 ms quilts] functional contrasts ($p < 0.001$, uncorrected for multiple comparisons), both at the individual subject level (Indiv fROI) and at the group level (RFX fROI). To investigate the response in these fROIs, we employed a similar leave-one-out cross-validation procedure as described for the anatomical ROIs above, with two differences: 1) because differences between the languages with respect to the absolute level of the BOLD signal are not of interest for the effects of temporal structure, we defined separate fROIs for English and Korean conditions via [EngOrig > Eng30ms] and [KorOrig > Kor30ms] functional contrasts, respectively; 2) we computed separate fROIs from the data in 7 runs (e.g. 1–7, 2–8, 1 3–8, etc.), leaving out the data from the eighth remaining run (i.e. 8, 1, 2, respectively); this was done 8 times (once for each left-out run) to yield 8 separate fROIs. Only voxels that a) survived a significance threshold of p < 0.001 (uncorrected for multiple comparisons across the volume) and b) lay within the superior temporal lobe were evaluated. The response in each left-out run and for each language was then normalized with respect to the mean response to the EngOrig or KorOrig condition in the 7 runs that formed the corresponding fROI, respectively. For example, for the English speech conditions, the BOLD signal for the Eng30ms, Eng120ms, Eng480ms, Eng960ms, and EngOrig conditions in run 1 was normalized with respect to the mean BOLD signal for EngOrig in runs 2–8 in the fROI determined by data from runs 2–8. This procedure was repeated for the other 7 leave-one-out combinations. The results were then averaged across runs for each participant.

The procedure was similar for the fROI in left IFG: for the [EngOrig > Eng30ms] functional contrast for 7 runs, only voxels in the left-out run that a) survived a significance threshold of $p < 0.005$ (uncorrected) and b) lay within a mask defined by BA44 and BA45 (from the Anatomy Toolbox, version 1.5 (Eickhoff et al., 2005)) were included in the analysis. For some participants who had no supra-threshold voxels in a given 7-run fROI (each participant had 8 possible fROIs), we randomly chose one voxel within the mask; this was the case for 3 participants in 1, 3, and 1 runs, respectively; this procedure ensures that statistics can be run, while simultaneously penalizing data from those participants. The response in each left-out run was then normalized with respect to the mean response to the English original condition in the 7 other runs that formed the corresponding fROI, respectively. The data were then averaged across runs for each participant.

For the psychophysiological interaction (PPI) analysis (Friston et al., 1997), we chose as the seed region the fROI defined in the left IFG for each participant and fold, based on the procedure described above. We then searched for areas throughout the brain that were modulated by activity in left IFG as a function of segment length. The resulting first-level contrast images of the PPI analysis were averaged across folds, and the average PPI contrast image of each participant then fed into a second-level t-test.

All preprocessing and analysis procedures were run in voxel space. However, for better simultaneous visualization of activation in gyral and sulcal structures, the second-level random-effects contrast images were rendered on SPM's cortex_20484.surf.gii surface and then inflated (cf. Figs. 3, 5, and 8). Similarly, for better comparison between studies, Figures 3 and 5 display z-scores (converted from the original second-level t-statistic values), thresholded at z > 3.09 (which corresponds to a p-threshold of $p < 0.001$ for 18 degrees of freedom).

2.6. Statistics

Statistical analysis of the fMRI data was based on a random-effects general linear regression model, as implemented in SPM (Friston et al., 1995). Group statistical parametric maps for functional contrasts were thresholded at $p < 0.001$ (uncorrected for multiple comparison), while the PPI analysis was based on $p < 0.05$ (FWE corrected). Normalized BOLD signal data in ROIs were analyzed via two-way repeated-measures (RM) ANOVAs. As appropriate, RM ANOVAs for BOLD data included factors ROI (HG, PT, RFX fROI, Indiv fROI), Hemisphere (left, right), Segment length (30 ms, 120 ms, 480 ms, 960 ms, Original), and Language (English, Korean). RM ANOVAs for behavioral data (percentage correct) included factors Segment length and Language. The Greenhouse-Geisser corrected degrees of freedom are reported in cases where Mauchly's test indicated a violation of the assumption of sphericity.

3. Results

3.1. Behavioral results

Average behavioral performance (percent correct) in the speaker identification task was well above chance (25%) for all conditions (Fig. 2). Performance, assessed via a RM ANOVA with factors Segment length and Language, generally increased with segment length (main effect of Segment length $F_{(4,72)} = 10.23$, $p < 0.001$, $\eta^2_p = 0.36$), and was better for English than Korean (main effect of Language: $F_{(1,18)} = 35.79$, $p < 0.001$, $\eta^2_p = 0.67$). Performance for Korean speech quilts was slightly more variable, leading to a weak Language × Segment length interaction ($F_{(4,72)} = 2.62$, $p = 0.04$, $\eta^2_p = 0.13$). Post-hoc pairwise comparisons (Bonferroni corrected) between Languages revealed significant differences for all corresponding Segment levels (e.g. English 30 ms vs. Korean 30 ms; all $p < 0.02$). In addition, post-hoc pairwise comparisons (Bonferroni corrected) between Segment length levels within a language revealed that, for English, no pairwise comparisons were significantly different ($p > 0.05$), while for Korean only the 30 ms segment length condition differed significantly from all but the Korean original condition (all $p < 0.005$).

3.2. Acoustic analysis (effects of temporal speech structure)

We first searched for areas that showed an increase in BOLD signal as a function of increasing temporal speech structure. Fig. 3 shows this for the [EngOrig > Eng30ms] and [KorOrig > Kor30ms] group functional contrasts and reveals areas in STS for both English and Korean, as well as left IFG for English speech. To enable a direct comparison with Overath et al. (2015), who did not include original, natural speech in their study design, we also investigated the [Eng960ms > Eng30ms] and [Kor960ms > Kor30ms] group functional contrasts; the pattern of results was very similar to that revealed in Fig. 3 (not shown).

Next, we investigated the response in these fROIs located along STS, as well as in anatomically defined cortical ROIs of the auditory system, i.e. HG and PT. The responses in cortical ROIs and fROIs differed significantly (main effect of ROI: $F_{(2.08,37.43)} = 173.45$, $p < 0.001$, $\eta^2_p = 0.91$), and we therefore investigated their responses separately. The BOLD signal in ROIs of early cortical auditory areas (HG and PT) decreased slightly as a function of segment length (Fig. 4): RM ANOVAs for HG and PT with factors Hemisphere, Segment length, and Language revealed weak main effects of Segment length ($F_{(4,72)} = 4.01$, $p = 0.005$, $\eta^2_p = 0.18$; $F_{(2.59,46.67)} = 4.68$, $p = 0.008$, $\eta^2_p = 0.21$; for HG and PT, respectively). However, post-hoc pairwise comparisons revealed a significant difference ($p < 0.05$, Bonferroni corrected) only for Korean in PT, and only between the 30 ms vs. 480 ms and 30 ms vs. original speech quilt conditions.

In the group fROI (RFX) for English, the BOLD signal increased as a function of temporal speech structure (main effect of Segment length: $F_{(4,72)} = 62.57$, $p < 0.001$, $\eta^2_p = 0.78$) and differed between hemispheres (main effect of Hemisphere: $F_{(1,18)} = 7.37$, $p = 0.01$, $\eta^2_p = 0.29$); the effect of Segment length was more pronounced in the left hemisphere (interaction: $F_{(2.22,40.03)} = 9.52$, $p < 0.001$, $\eta^2_p = 0.35$). In the group fROI (RFX) for Korean, the BOLD signal increased as a function of Segment length ($F_{(2.45,44.09)} = 20.49$, $p < 0.001$, $\eta^2_p = 0.53$), while revealing an interaction with Hemisphere ($F_{(4,72)} = 6.16$, $p < 0.001$, $\eta^2_p = 0.26$).

For the individual fROIs (Indiv) for English and Korean the pattern was largely identical, but the size of the effects was generally larger. For the individual English fROI (Indiv), the BOLD signal increased as a function of segment length ($F_{(2.87,51.7)} = 163.49$, $p < 0.001$, $\eta^2_p = 0.9$), differed between hemispheres ($F_{(1,18)} = 12.37$, $p = 0.002$, $\eta^2_p = 0.41$), and revealed an interaction ($F_{(4,72)} = 25.71$, $p < 0.001$, $\eta^2_p = 0.59$). Post-hoc pairwise comparisons between adjacent segment length conditions showed significant differences ($p < 0.05$, Bonferroni corrected) between all but the 960 ms and original speech conditions.

For the individual Korean fROI (Indiv), the BOLD signal increased as a function of segment length ($F_{(4,72)} = 48.04$, $p < 0.001$, $\eta^2_p = 0.73$). Post-hoc pairwise comparisons between adjacent segment length conditions showed significant differences ($p < 0.05$, Bonferroni corrected) on the left between 30 ms, 120 ms, and 480 ms conditions, and on the right between 30 ms and 120 ms conditions.

Since performance in the speaker identification task (Fig. 2) varied somewhat as a function of Language and Segment length, we tested whether it could account for some portion of the BOLD signal responses reported here (Fig. 4). To this end, we set up a second-level factorial design in SPM12 with factors Subjects, Language, and Segment length, with each participant contributing 10 contrast images (one for each condition), and each participant's average behavioral performance for a given condition specified as a covariate. The measurements of factor Subject were treated as independent, and the variance as unequal; for both within-subject factors Language and Segment length, measurements were treated as dependent, and Variance as equal. An F-test on the behavioral covariate (thresholded at $p < 0.001$, uncorrected) revealed no effects in auditory cortex or left IFG; in addition, the effects of Segment length reported in Fig. 3 are maintained when including behavioral performance as a covariate. Taken together, these results suggest that the task performance does not significantly account for the BOLD signal changes in the areas of interest.

The speaker identification task also likely recruited areas in the temporal cortex that are involved in voice processing (Belin et al., 2000, Belin, 2006) and voice recognition (Zäske et al., 2017, Andics et al., 2010). We therefore tested whether these areas might be recruited differentially by the four speakers as a function of temporal speech structure. For each participant, we created first-level [Original > 30 ms quilt] contrast images separately for each speaker and language; these were then included in two separate second-level within-subjects factorial models for English and Korean stimuli. This allowed us to test for an interaction between speaker and temporal speech structure. No areas in auditory cortex or left IFG showed such an interaction (all $p > 0.001$, uncorrected).

We also investigated areas that showed a stronger response to short temporal speech structure than original speech via the functional contrast [30ms > Orig] (the [Eng30ms > EngOrig] and [Kor30ms > KorOrig] functional contrasts did not differ significantly from each other). This revealed a bilateral network of areas in the middle and anterior insula, inferior parietal cortex, and a small area within PT (Fig. 5).

Finally, we also investigated the sensitivity to temporal speech structure in earlier, subcortical structures of the auditory system. Fig. 6 shows that IC and MGB were unaffected by the manipulation of speech structure in the speech quilts. A RM ANOVA with factors ROI (IC, MGB), Hemisphere, Language, and Segment length revealed no statistically significant main effects or interactions (all $p > 0.05$).

3.3. Linguistic analysis (effects of language familiarity)

The results presented so far concern effects of temporal speech structure that are similar in nature for English and Korean; that is, they address an analysis at the level of signal acoustics, irrespective of language familiarity. Next, we turn to neural responses that differ between languages.

The response in the fROIs revealed two notable differences between languages: First, a RM ANOVA with factors Segment length, Language, and Hemisphere revealed that the size of the effect of segment length was larger for English than Korean (Segment length × Language interaction: $F_{(4,72)} = 31.00$, $p < 0.001$, $\eta^2_p = 0.63$). Second, the volume of the individual fROI in STS was significantly larger in the left hemisphere ($M = 1107.68$, SEM = 180.78) than the right hemisphere ($M = 443.84$, SEM = 122.29) only for English, but not for Korean ($M = 196.37$, SEM = 44.86 vs. $M = 138.16$, SEM = 35.63, for left and right hemispheres, respectively): a RM ANOVA with factors Language and Hemisphere revealed a significant interaction ($F_{(1,18)} = 43.64$, $p < 0.001$, $\eta^2_p = 0.71$).

Beyond the response in STS, the functional contrast [EngOrig > Eng30ms] also revealed an effect of segment length in the left IFG (see Fig. 3). The response in left IFG increased as a function of segment length only for English, but was flat for Korean (Fig. 7). (The [KorOrig > Kor30ms] functional contrast did not reveal any effects beyond the temporal lobes; therefore, the responses to Korean stimuli are normalized with respect to the English original stimuli in this instance.) A RM ANOVA with factors Language and Segment length revealed main effects of Language ($F_{(1,18)} = 171.31$, $p < 0.001$, $\eta^2_p = 0.91$) and Segment length ($F_{(4,72)} = 38.2$, $p < 0.001$, $\eta^2_p = 0.68$), which was due to a significant interaction ($F_{(4,72)} = 28.76$, $p < 0.001$, $\eta^2_p = 0.62$). Post-hoc pairwise comparisons between adjacent segment length levels revealed that, for English, only 30 ms vs. 120 ms, and 480 ms vs. 960 ms were not significantly different ($p > 0.05$, Bonferroni corrected). Between languages, all comparisons except for speech quilts with 30 ms segment length were significant ($p < 0.05$, Bonferroni corrected).

In order to further investigate the differential responses with respect to temporal speech structure between the left IFG on the one hand (Fig. 7), and auditory cortex in the temporal lobe on the other (Fig. 4), we performed a PPI analysis with left IFG as the seed region and an unconstrained search area (Friston et al., 1997). This analysis asks the question whether the response in left IFG has a modulatory effect else-where in the brain (though see Friston et al. (1997) with respect to ambiguity of directionality). This analysis revealed a modulatory effect in bilateral primary and non-primary auditory cortices (bilateral HG, PT, and STG) (Fig. 8, yellow). Note that these areas in auditory cortex are largely distinct from those revealed by the English group fROI (red; see also Fig. 3, top) and show only a modest amount of overlap (white).

4. Discussion

We provide evidence for the neural processes underlying the transformation from acoustic analysis to linguistic analysis of temporal speech structure. The results suggest that STS processes the acoustic properties of temporal speech structure, while left IFG transforms this acoustic information to linguistic representations. In addition, connectivity analyses suggest that this transformation modulates the processing of acoustic speech properties in earlier auditory areas in cortex.

Based on decades of research, current speech/language models are now able to delineate the major cortical structures and their putative roles in speech perception and production

(Hickok and Poeppel, 2007, Rauschecker and Scott, 2009, Friederici and Gierhan, 2013, Skeide and Friederici, 2016). However, while there is some evidence for how they interact and modulate each other, e.g. as a function of intelligibility (Leff et al., 2008, Park et al., 2015, Tuennerhoff and Noppeney, 2016), it is currently still unclear how they proceed from the analysis of speech-specific acoustic structure to linguistic analysis (acousto-linguistic transformation). This is mainly because the majority of studies on which current speech/ language models are based, either manipulated mainly linguistic content but not acoustic content (e.g. via syntactic violations; Friederici et al., 2000, Friederici et al., 2003, Wartenburger et al., 2004), mainly acoustic content (e.g. noise vocoding; Scott et al., 2000, Narain et al., 2003, Evans et al., 2014, Obleser et al., 2008), or both linguistic and acoustic content simultaneously (e.g. spectral rotation, time-reversed speech; Scott et al., 2000, Narain et al., 2003, Hasson et al., 2008, Lerner et al., 2011). In contrast, the current study was able to dissociate acoustic from linguistic analyses of temporal speech structure by comparing the cortical response to the same acoustic manipulation (temporal speech structure via speech quilting) in familiar and foreign languages. This ensured that differences between familiar and foreign languages as a function of temporal speech structure are mostly due to linguistic analysis of the familiar language: as the segment length increases, longer linguistic units such as syllables and words, or even brief sentences, become apparent and will automatically engage cortical areas that are involved in the analysis of syntax and semantics, but this will only be the case in a familiar language.

The results confirm that acoustic analysis of temporal speech structure takes place in STS: the BOLD signal increased as a function of temporal speech structure in both foreign and familiar languages. In general, the effects in STS for the current foreign language (Korean) were somewhat weaker than for the foreign language (German) in Overath et al. (2015). For example, while previously the normalized BOLD signal for speech quilts with 30 ms segment lengths was about 60% of that for 960 ms segment lengths, it was about 80% in the current study. Similarly, the current study revealed a significant difference between segment lengths only for the two shortest segment lengths used. These differences may be related to a number of factors. First, the effect estimation in Overath et al. (2015) was potentially more robust due to a total of 52 scanning sessions (and up to 4 scanning sessions for a given participant), as opposed to 19 individual scanning sessions in the current study. Second, the etymological difference between Korean and English is greater than that between German and English, and it is possible that the corresponding differences in temporal speech acoustics between Korean and English can explain some of these differences. In fact, we chose Korean precisely because of its etymological and linguistic distance to English (Chiswick and Miller, 2005), in an effort to obtain as ‘clean’ a measure of acoustic analysis of temporal speech structure as possible. Future studies will therefore need to explore further the degree to which the etymological dis/similarities between languages affect the acoustic analysis of speech structure in STS. Similarly, it will be important to confirm that the results are not specific to any potential idiosyncratic acoustic differences between English and Korean, but that they generalize to native vs. foreign language comparisons.

The response in the left IFG revealed a dissociation between the acoustic and linguistic analyses of temporal speech structure: it increased as a function of segment length only in the familiar language (English), but was unaffected by the same manipulation of temporal speech structure in the foreign language (Korean). This suggests a crucial role for the left IFG in the acousto-linguistic transformation of temporal speech structure, whereby left IFG extracts linguistic information from the temporal speech structure only if it matches familiar linguistic templates. However, precisely which aspect of temporal speech structure (e.g. phonological, syntactic, or semantic information) is driving this transformation in left IFG remains to be determined, since the present quilting approach does not differentiate between these characteristics of temporal speech structure. Such a differentiation would also be able to speak to the specific contributions towards the analysis of temporal speech characteristics of sub-regions within left IFG (BA44, BA45, BA47) that have been shown to sub-serve different functional roles with respect to syntax, phonology, and semantics (Friederici and Gierhan, 2013, Matchin, 2017).

Overath et al. (2015) demonstrated that the BOLD signal plateaued for temporal speech structure longer than ~500 ms. Importantly, since the inflection point was the same for time-compressed speech quilts – which, in a given segment, contain twice as much temporal structure as normal uncompressed speech – the plateau was attributed to intrinsic acoustic analysis properties of auditory cortex, rather than reflecting the analysis of intrinsic stimulus properties. In the current study, the BOLD signal did not show a clear plateau and generally continued to increase beyond 480 ms segment lengths. However, it should be noted that, while this trend was visible for Korean, it did not reach statistical significance. In contrast, the response to 960 ms speech quilts in English was significantly larger than that to 480 ms speech quilts, in both the left and right hemispheres. It is possible that the inclusion of the speaker identification task, and the associated increase in task difficulty and attention, may have led to a generally enhanced sensitivity to longer temporal windows of analysis (Overath et al. (2015)) simply asked participants to press a button after each sound). Future studies will need to determine whether analysis windows beyond ~500 ms are indeed malleable to task demands or attention, for example via a direct comparison of attended vs. ignored speech quilts as a function of segment length.

A novel finding in the current study concerns the possible role of a task on the response in auditory cortex and beyond. Overath et al. (2015) did not find any areas that showed a stronger response as a function of *decreasing* segment length. One possibility for this discrepancy is that, whereas the earlier study simply asked participants to press a button at the end of each sound, the behavioral speaker identification task in the current study required participants to engage more explicitly with the stimulus. In fact, speaker identification became somewhat more challenging and performance decreased as the segment length decreased. The areas that showed a stronger response to speech quilts with short segment lengths are associated with processing demands in linguistic tasks (Falkenberg et al., 2011, Yue et al., 2013). Thus, it is possible that this effect is due less to the analysis of temporal acoustics in speech signals, but more related to general attentional or task demands.

The performance in the speaker identification task varied as a function of segment length and language familiarity; however, there are several reasons we believe that these differences do not affect our conclusions. First, explicitly modeling participants' behavioral scores as a covariate revealed no areas in auditory cortex (or left IFG) that varied significantly ($p >$ 0.001, uncorrected) as a function of behavior. Second, with respect to task difficulty, all participants were well above chance performance (25%), suggesting that they were not struggling with the speaker identification task, just that they found it slightly more difficult for shorter segment lengths and for Korean stimuli. In addition, task difficulty is typically either unaffected by (Dräger et al., 2004, Demb et al., 1995), or correlated positively with (Keller et al., 2001, Desai et al., 2006) BOLD signal strength in language areas, while in the current study the predominant BOLD signal effect of segment length is stronger in English, for which the task was apparently somewhat easier. Third, the pattern of the behavioral performance does not explain or reflect the observed BOLD signal pattern, neither the BOLD signal increase for both languages as a function of segment length in STS, nor the differential BOLD signal responses in left IFG. Finally, Overath et al. (2015) used a task that had no stimulus-related difficulty (simply pressing a button at the end of each sound), but found essentially the same BOLD response shape in STS for a foreign language (German).

The decrease in speaker identification with decreasing temporal speech structure (Fig. 2) suggests that the quilting algorithm might disrupt acoustic cues that are important for paralinguistic processes involved in speaker identification. Such processes are likely to recruit regions in the temporal lobes that are involved in voice processing (Belin et al., 2000, Belin, 2006) and voice recognition (Zäske et al., 2017, Andics et al., 2010). It is possible that the speaker identification task interacted with the areas identified here in STS or left IFG. However, we found no evidence that any areas in the auditory cortex (or left IFG) responded differently to the four speakers as a function of temporal speech structure. This suggests that, at least in the current paradigm, speaker identification and temporal speech processing are independent.

While the response increase in STS as a function of segment length was bilateral for foreign speech (see also Overath et al., 2015), its extent was significantly left-lateralized for familiar speech. This suggests that the (pre-linguistic) acoustic analysis of temporal speech structure takes place in both hemispheres; however, if linguistic processes are able to become engaged, then left-hemispheric structures in STS are more strongly recruited. This provides further evidence for the view that speech perception is largely a bilateral process, for which lateralization emerges only once linguistic processes become engaged (Peelle, 2012).

Based on the current results we propose an extension of classical language models (Hickok and Poeppel, 2007, Rauschecker and Scott, 2009, Friederici and Gierhan, 2013, Skeide and Friederici, 2016), particularly with respect to the acousto-linguistic transformation of speech-specific temporal structure in the human brain. In this model, acoustic information is passed from primary and non-primary auditory cortices to STS for processing of temporal speech structure; this stage of processing is primarily concerned with the analysis of acoustic properties of temporal speech structure. If, however, this information contains familiar linguistic information (e.g. lexical, semantic, syntactic, phonemic cues), it is passed on to left IFG for linguistic analysis. The PPI analysis suggests that left IFG may then subsequently modulate the processing in earlier auditory areas (though see (Friston et al., 1997) for an alternative possibility of directionality), which in turn would induce greater sensitivity in STS for speech-specific temporal structure; the latter could explain the steeper slope for increasing speech-specific temporal structure for familiar speech (English) compared to foreign speech (Korean).

There is recent evidence in support of this view, both for a temporal progression of linguistic analysis from temporal to (left) frontal cortices, and in terms of top-down modulation of auditory cortex. With respect to the former, the analysis of acoustic speech information in temporal cortex precedes phonological analysis in left inferior frontal cortex (Toscano et al., 2018). Similarly, the lexical and semantic processes that lead up to and surround the uniqueness point of a word (the point at which the word can be uniquely identified and differentiated from other similarly sounding candidates) include a successive involvement of temporal to inferior frontal cortices (Kocagoncu et al., 2017); this succession is mirrored in the increase in temporal window of integration size for speech processing (Lerner et al., 2011). With respect to evidence for the role of top-down feedback in speech perception, while anatomical studies in non-human primates (Hackett et al., 1999) and humans (Saur et al., 2008) have demonstrated a link between left IFG and auditory cortex, its functional relevance is supported by numerous studies showing an involvement of left IFG when processing degraded speech (Davis and Johnsrude, 2003, Davis and Johnsrude, 2007, Giraud et al., 2004, Zekveld et al., 2006). Pertaining specifically to temporal processes in speech perception, which are the focus here, top-down signals from frontal cortex increase the ability of auditory cortex to track temporal speech envelope modulations in the delta and theta bands (Park et al., 2015).

In conclusion, by simultaneously controlling temporal speech structure and linguistic familiarity, the current study was able to disambiguate the neural contributions underlying acoustic and linguistic analyses of temporal speech structure. The results thereby inform our understanding of where and how linguistic information interfaces with, and modulates the temporal analysis of speech.

Acknowledgements

This work was supported in part by US National Institutes of Health grant R21DC016386 and Duke University startup funds to T.O. The authors thank Kimberly Paige Mihalsky for assistance with data collection and the three anonymous reviewers for their constructive feedback.

References

Andics A, McQueen JM, Petersson KM, Gál V, Rudas G, Vidnyánsky Z., 2010. Neural mechanisms of voice recognition. Neuroimage 52 (4), 1528–1540. [PubMed: 20553895]

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B., 2000. Voice-selective areas in human auditory cortex. Nature 403 (6767), 309. [PubMed: 10659849]

Belin P., 2006. Voice processing in human and non-human primates. Philos. Trans. R. Soc. Lond B Biol. Sci 361 (1476), 2091–2107. [PubMed: 17118926]

Brainard DH., 1997. The psychophysics toolbox. Spat Vis. 10 (4), 443–446. [PubMed: 9176954]

Brett M, Anton J-L, Valabregue R, Poline JB, 2002. Region of interest analysis using an SPM toolbox (abstract). Neuroimage 16 (Suppl).

Chiswick BR, Miller PW., 2005. Linguistic distance: a quantitative measure of the distance between English and other languages. J. Multilingual Multicultural Develop 26 (1), 1–11.

Davis MH, Johnsrude IS., 2003. Hierarchical processing in spoken language comprehension. J. Neurosci 23, 3423–3431. [PubMed: 12716950]

Davis MH, Johnsrude IS., 2007. Hearing speech sounds: top-down influences on the interface between audition and speech perception. Hear Res. 229, 132–147. [PubMed: 17317056]

Demb JB, Desmond JE, Wagner AD, Vaidya CJ, Glover GH, Gabrieli JD., 1995. Semantic encoding and retrieval in the left inferior prefrontal cortex: a functional MRI study of task difficulty and process specificity. J. Neurosci 15 (9), 5870–5878. [PubMed: 7666172]

Desai R, Conant LL, Waldron E, Binder JR., 2006. fMRI of past tense processing: the effects of phonological complexity and task difficulty. J. Cogn. Neurosci 18 (2), 278–297. [PubMed: 16494687]

Devlin JT, Sillery EL, Hall DA, Hobden P, Behrens TEJ, Nunes RG, et al., 2006. Reliable identification of the auditory thalamus using multi-modal structural analyses. Neuroimage 30 (4), 1112–1120. [PubMed: 16473021]

Ding N, Melloni L, Zhang H, Tian X, Poeppel D., 2015. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci 19, 158–164. [PubMed: 26642090]

Dräger B, Jansen A, Bruchmann S, Förster AF, Pleger B, Zwitserlood P, et al., 2004. How does the brain accommodate to increased task difficulty in word finding?: a functional MRI study. Neuroimage 23 (3), 1152–1160. [PubMed: 15528114]

Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, et al., 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 25 (4), 1325–1335. [PubMed: 15850749]

Evans S, Kyong JS, Rosen S, Golestani N, Warren JE, McGettigan C, et al., 2014. The pathways for intelligible speech: multivariate and univariate perspectives. Cereb. Cortex 24 (9), 2350–2361. [PubMed: 23585519]

Falkenberg LE, Specht K, Westerhausen R., 2011. Attention and cognitive control networks assessed in a dichotic listening fMRI study. Brain Cogn. 76 (2), 276–285. [PubMed: 21398015]

Friederici AD, Gierhan AME., 2013. The language network. Curr. Opin. Neurobiol 23 (2), 250–254. [PubMed: 23146876]

Friederici AD, Meyer M, von Cramon DY., 2000. Auditory language comprehension: an event-related fMRI study on the processing of syntactic and lexical information. Brain Lang. 7 (2), 85–96.

Friederici AD, Rüschemeyer SA, Hahne A, Fiebach CJ., 2003. The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. Cereb. Cortex 13 (2), 170–177. [PubMed: 12507948]

Friston KJ, Holmes AP, Worsley KJ, Poline JB, Frith CD, Frackowiak RS., 1995. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp 2, 189–210.

Friston KJ, Büchel C, Fink GR, Morris J, Rolls E, Dolan RJ., 1997. Psychophysical and modulatory interactions in neuroimaging. Neuroimage 6, 218–229. [PubMed: 9344826]

Giraud AL, Poeppel D., 2012. Speech perception from a neurophysiological perspective. In: Poeppel D, Overath T, Popper AN, Fay RR (Eds.), The Human Auditory Cortex. Springer Handbook of Auditory Research: Springer Science + Business Media, LLC, pp. 225–260.

Giraud AL, Kell C, Thierfelder C, Sterzer P, Russ MO, Preibisch C, et al., 2004. Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. Cereb. Cortex 14 (3), 247–255. [PubMed: 14754865]

Griffiths TD, Warren JD., 2002. The planum temporale as a computational hub. Trends Neurosci. 25 (7) 348–253. [PubMed: 12079762]

Griffiths TD, Uppenkamp S, Johnsrude I, Josephs O, Patterson RD., 2001. Encoding of the temporal regularity of sound in the human brainstem. Nat. Neurosci 4 (6), 633–637. [PubMed: 11369945]

Hackett TA, Stepniewska I, Kaas JH., 1999. Prefrontal connections of the parabelt auditory cortex in macaque monkeys. Brain Res. 817 (1–2), 45–58. [PubMed: 9889315]

Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N, 2008. A hierarchy of temporal receptive windows in human cortex. J. Neurosci 28 (10), 2539–2550. [PubMed: 18322098]

Hickok G, Poeppel D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci 8 (5), 393–402. [PubMed: 17431404]

Keller T, Carpenter P, Just MA., 2001. The neural bases of sentence comprehension: an fMRI examination of syntactic and semantic processing. Cereb. Cortex 11, 223–237. [PubMed: 11230094]

Kleinschmidt DF, Jaeger TF., 2015. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. Psychol. Rev 122 (2), 148–203. [PubMed: 25844873]

Kocagoncu E, Clarke A, Devereux BJ, Tyler LK., 2017. Decoding the cortical dynamisc of sound-meaning mapping. J. Neurosci 37 (5), 1312–1319. [PubMed: 28028201]

Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci 12 (5), 535–540. [PubMed: 19396166]

Kumar S, Stephan KE, Warren JD, Friston KJ, Griffiths TD., 2007. Hierarchical processing of auditory objects in humans. PLoS Comp Biol 3 (6), e100.

Leff AP, Schofield TM, Stephan KE, Crinion JT, Friston KJ, Price CJ., 2008. The cortical dynamics of intelligible speech. J. Neurosci 28 (49), 13209–13215. [PubMed: 19052212]

Lerner Y, Honey CJ, Silbert LJ, Hasson U., 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci 31 (8), 2906–2915. [PubMed: 21414912]

Matchin WG., 2017. A neural retuning hypothesis of sentence-specificity in Broca’s area. Psychon. Bull. Rev 25 (5), 1682–1694.

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

Narain C, Scott SK, Wise RJ, Rosen S, Leff A, Iversen SD, et al., 2003. Defining a left-lateralized response specific to intelligible speech using fMRI. Cereb. Cortex 13 (12), 1362–1368. [PubMed: 14615301]

Obleser J, Eisner F, Kotz SA., 2008. Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. J. Neurosci 28 (32), 8116–8123. [PubMed: 18685036]

Overath T, Cusack R, Kumar S, von Kriegstein K, Warren JD, Grube M, et al., 2007. An information theoretic characterisation of auditory encoding. PLoS Biol. 5 (11), e288. [PubMed: 17958472]

Overath T, McDermott JH, Zarate JM, Poeppel D., 2015. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat. Neurosci 18 (6), 903–911. [PubMed: 25984889]

Pérez A, Carreiras M, Dowens MG, Duñabeitia JA., 2015. Differential oscillatory encoding of foreign speech. Brain Lang. 147, 51–57. [PubMed: 26070104]

Park H, Ince RA, Schyns PG, Thut G, Gross J., 2015. Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. Curr. Biol 25 (12), 1649–1653. [PubMed: 26028433]

Peña M, Melloni L., 2012. Brain oscillations during spoken sentence processing. J. Cogn. Neurosci 24 (5), 1149–1164. [PubMed: 21981666]

Peelle JE., 2012. The hemispheric lateralization of speech processing depends on what “speech” is: a hierarchical perspective. Front. Hum. Neurosci 6, 309. [PubMed: 23162455]

Poeppel D, Idsardi WJ, van Wassenhove V., 2008. Speech perception at the interface of neurobiology and linguistics. Philos. Trans. R. Soc. Lond B Biol. Sci 363 (1493), 1071–1086. [PubMed: 17890189]

Poeppel D., 2003. The analysis of speech in different temporal integration windows: cerebral lateralization as ’asymmetric sampling in time. Speech Commun. 41 (1), 245–255.

Rademacher J, Morosan P, Schormann T, Schleicher A, Werner C, Freund HJ, et al., 2001. Probabilistic mapping and volume measurement of human primary auditory cortex. Neuroimage 13 (4), 669–683. [PubMed: 11305896]

Rauschecker JP, Scott SK., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci 12, 718–724. [PubMed: 19471271]

Rosen S., 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. Philos. Trans. R. Soc. Lond B Biol. Sci 336 (1278), 367–373. [PubMed: 1354376]

Saur D, Kreher BW, Schnell S, Kümmerer D, Kellmeyer P, Vry MS, et al., 2008. Ventral and dorsal pathways for language. Proc. Natl. Acad. Sci U S A 105 (46), 18035–18040. [PubMed: 19004769]

Scott SK, Blank CC, Rosen S, Wise RJ., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123 (12), 2400–2406. [PubMed: 11099443]

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304. [PubMed: 7569981]

Skeide MA, Friederici AD., 2016. The ontogeny of the cortical language network. Nat. Rev. Neurosci 17 (5), 323–332. [PubMed: 27040907]

Smith ZM, Delgutte B, Oxenham AJ., 2002. Chimaeric sounds reveal dichotomies in auditory perception. Nature 416 (6876), 87–90. [PubMed: 11882898]

Stevens KN., 2000. Acoustic Phonetics. MIT Press, Cambridge, MA.

Strelnikov K, Massida Z, Rouger J, Belin P, Barone P., 2011. Effects of vocoding and intelligibility on the cerebral response to speech. BMC Neurosci. 12, 122. [PubMed: 22129366]

Toscano JC, Anderson ND, Fabiani M, Gratton G, Garnsey SM., 2018. The time-course of cortical responses to speech revealed by fast optical imaging. Brain Lang. 184, 32–42. [PubMed: 29960165]

Tuennerhoff J, Noppeney U., 2016. When sentences live up to your expectations. Neuroimage 124, 641–653. [PubMed: 26363344]

Wartenburger I, Heekeren HR, Burchert F, Heinemann S, De Bleser R, Villringer A., 2004. Neural correlates of syntactic transformations. Hum. Brain Mapp 22 (1), 72–81. [PubMed: 15083528]

Westbury CF, Zatorre RJ, Evans AC., 1999. Quantifying variability in the planum temporale: a probability map. Cereb Cortex 9 (4), 392–405. [PubMed: 10426418]

Yue Q, Zhang LI, Xu G, Shu H, Li P., 2013. Task-modulated activation and functional connectivity of the temporal and frontal areas during speech comprehension. Neuroscience 237, 87–95. [PubMed: 23357111]

Zäske R, Awwad Shiekh Hasan B, Belin P, 2017. It doesn’t matter what you say: fMRI correlates of voice learning and recognition independent of speech content. Cortex 94, 100–112. [PubMed: 28738288]

Zekveld AA, Heslenfeld DJ, Festen JM, Schoonhoven R., 2006. Top-down and bottom-up processes in speech comprehension. Neuroimage 32, 1826–1836. [PubMed: 16781167]

Significance statement

Where and how the acoustic information contained in complex speech signals is mapped to linguistic information is still not fully explained by current speech/language models. We dissociate acoustic from linguistic analyses of speech by comparing the same acoustic manipulation (varying the extent of temporal speech structure) in two languages (native, foreign). We show that acoustic temporal speech structure is analyzed in superior temporal sulcus (STS), while linguistic information is extracted in left inferior frontal gyrus (IFG). Furthermore, modulation from left IFG enhances sensitivity to temporal speech structure in STS. We propose a model for acousto-linguistic transformation of temporal speech structure in the human brain that synthesizes these results.

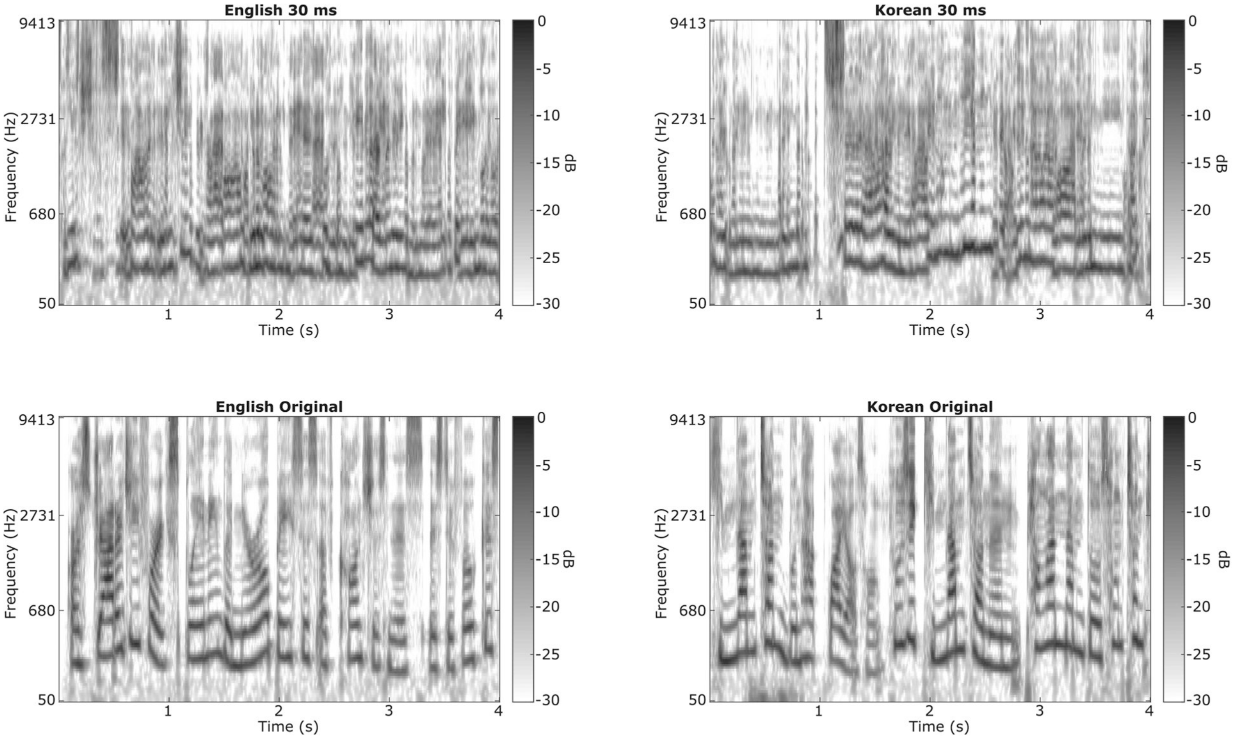

Fig. 1.

Example cochleograms of quilts made with 30 ms segments of the source signal (top), or of the unaltered original source signal (bottom), displayed for English (left column) and Korean (right column). The four cochleograms are based on four different source signals.

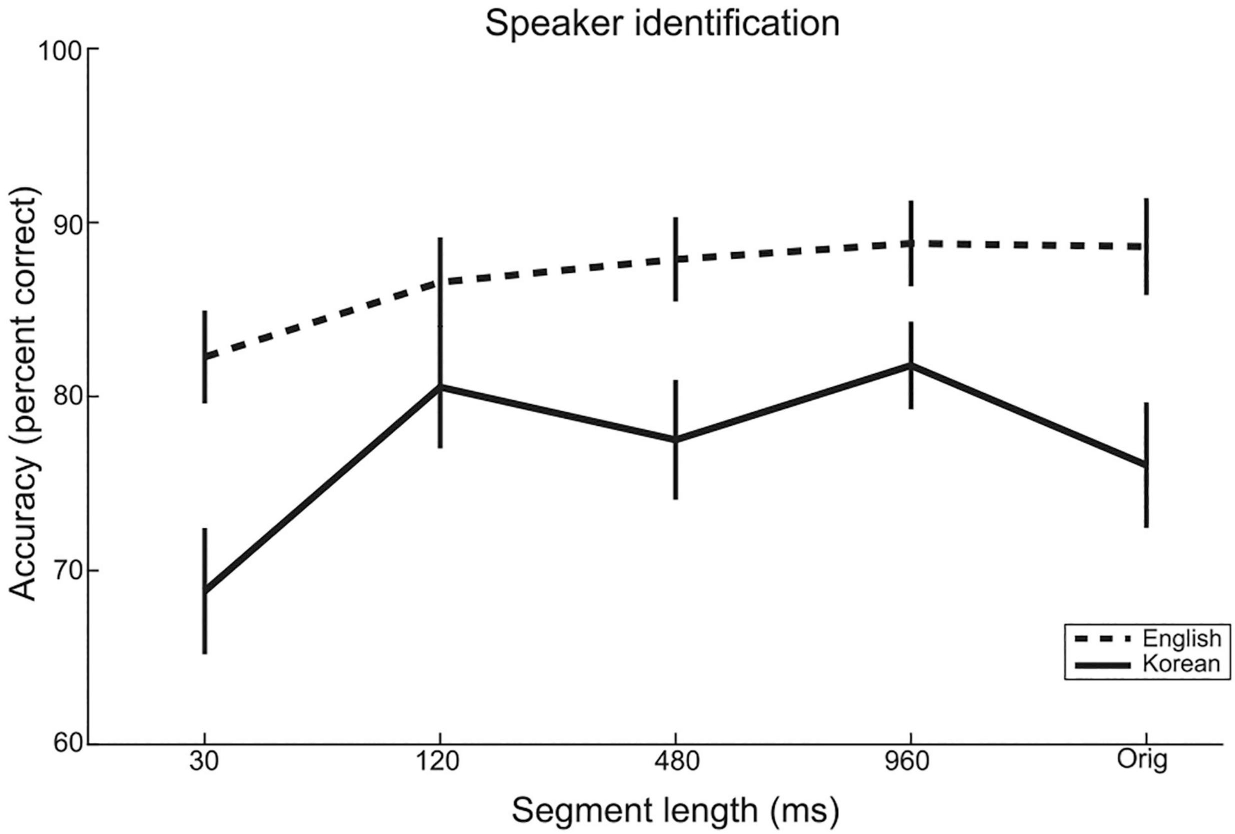

Fig. 2.

Percent correct performance for the speaker identification task for English (dashed) and Korean (solid) speech quilts as a function of segment length. Error bars denote ± 1 SEM.

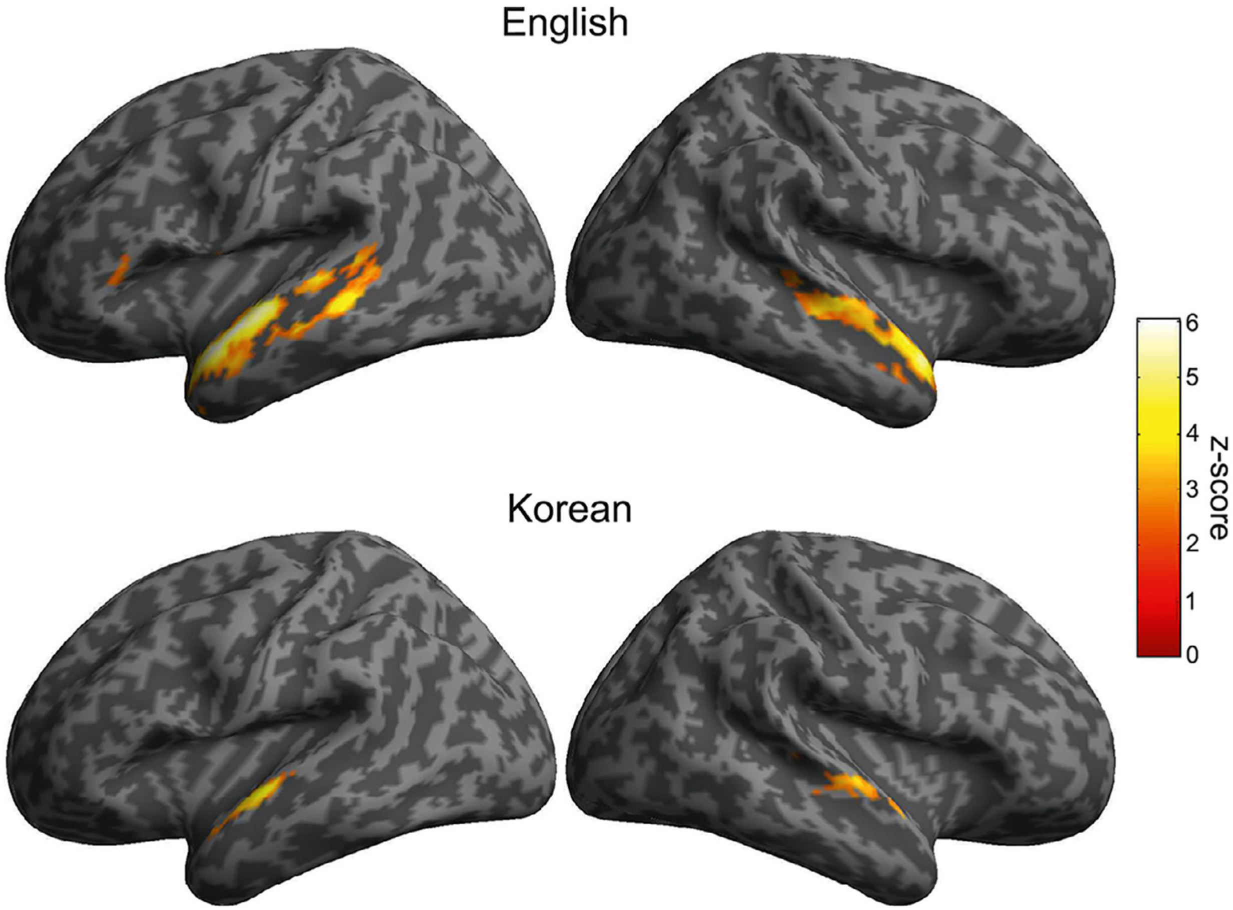

Fig. 3.

Areas showing significantly stronger BOLD signal to original speech compared to speech quilted with 30 ms segment lengths in English (top) and Korean (bottom).

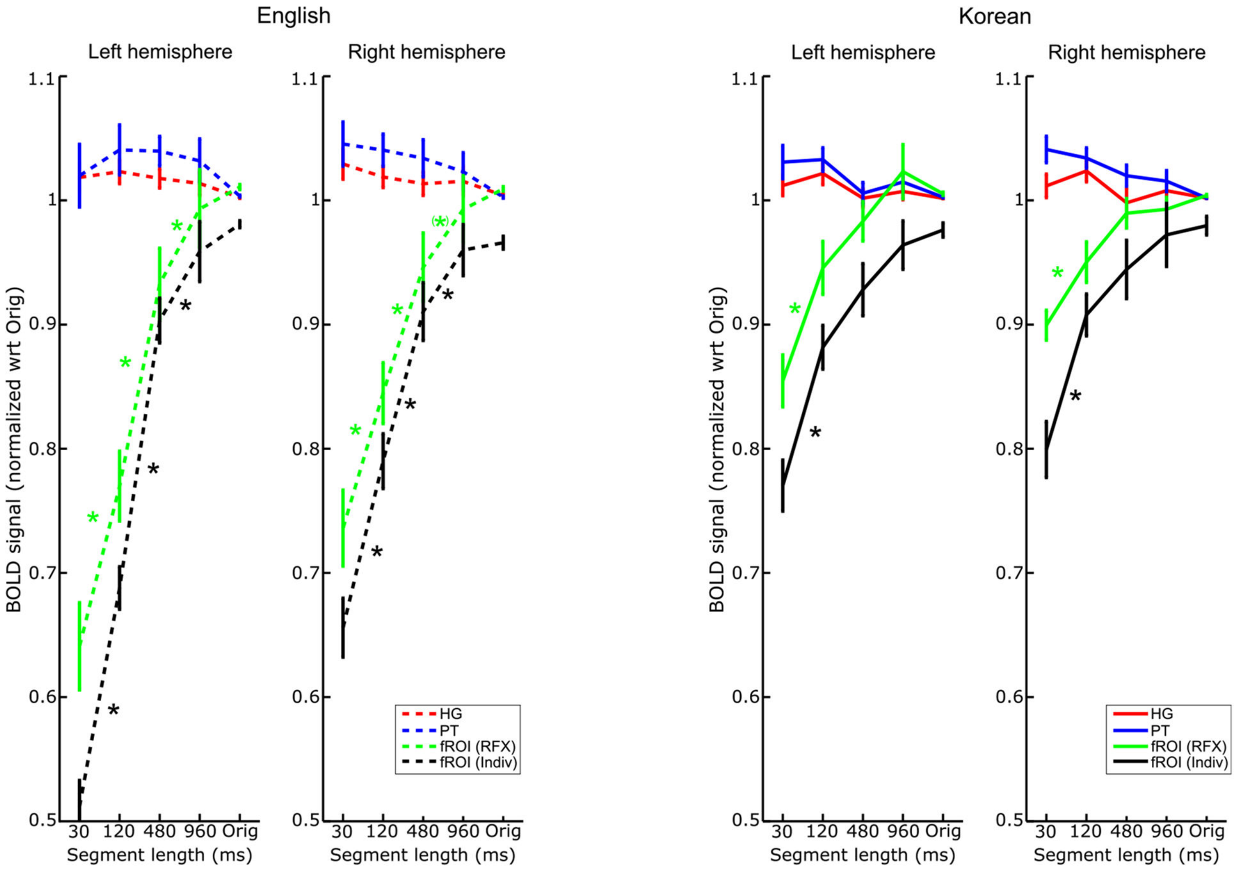

Fig. 4.

Response in anatomical ROIs HG (red) and PT (blue), and functional ROIs of the group fROI shown in Fig. 3 (green) and individual fROIs (black), shown separately for the two languages and two hemispheres. Error bars denote ± 1 SEM. Asterisks denote significant pairwise comparisons ($p < 0.05$, Bonferroni corrected) between adjacent segment length conditions. For each language, responses are normalized within each ROI to the response to the original speech condition in the left-out run (see Methods).

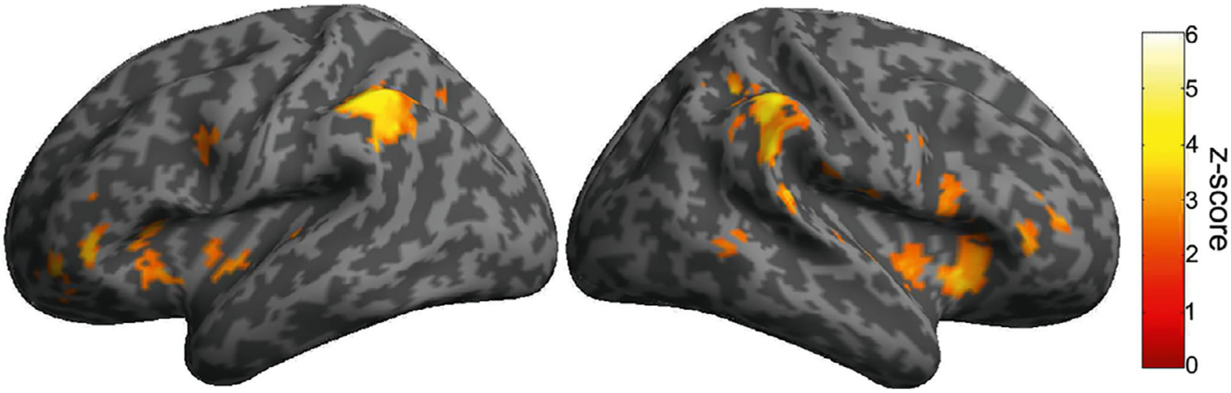

Fig. 5.

Areas showing significantly stronger BOLD signal to speech quilted with 30 ms segment lengths compared to original speech (across languages).

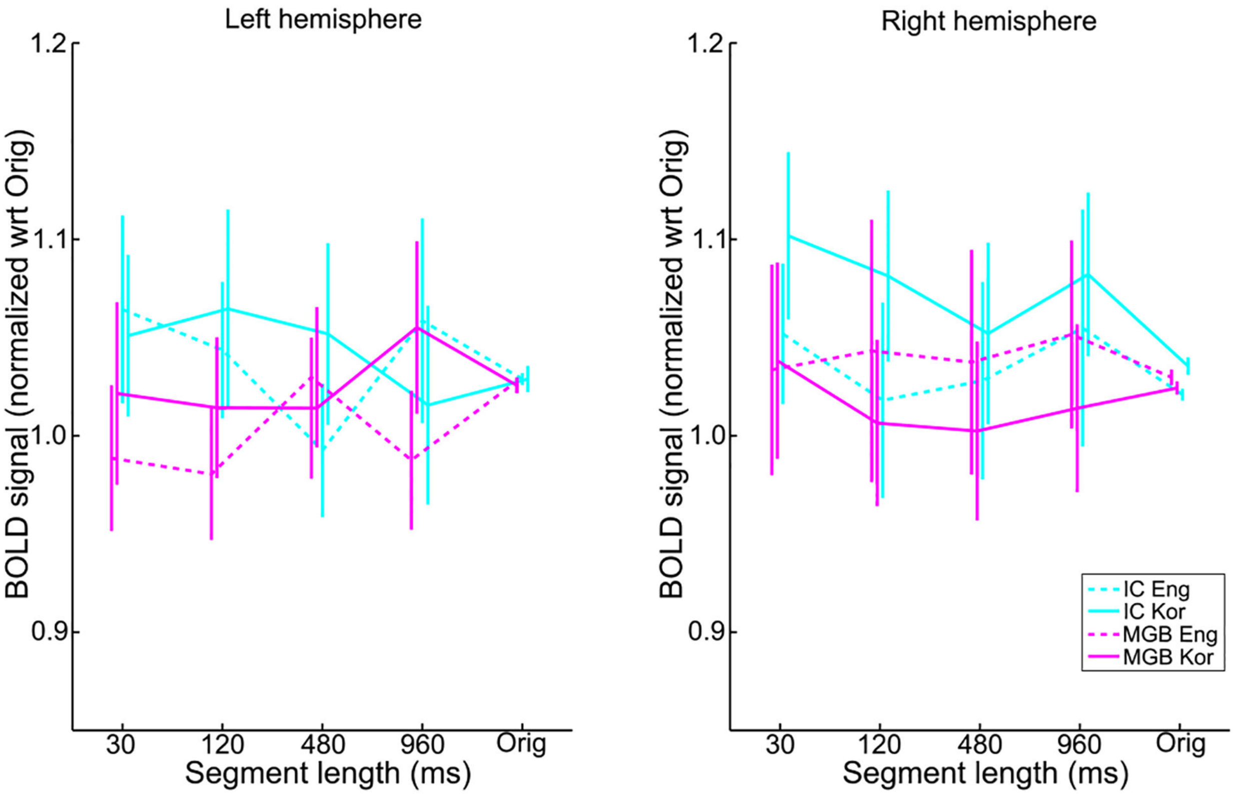

Fig. 6.

Response in IC and MGB. Error bars denote ± 1 SEM. For each language, responses are normalized within each ROI to the response to original speech in the left-out run (see Methods).

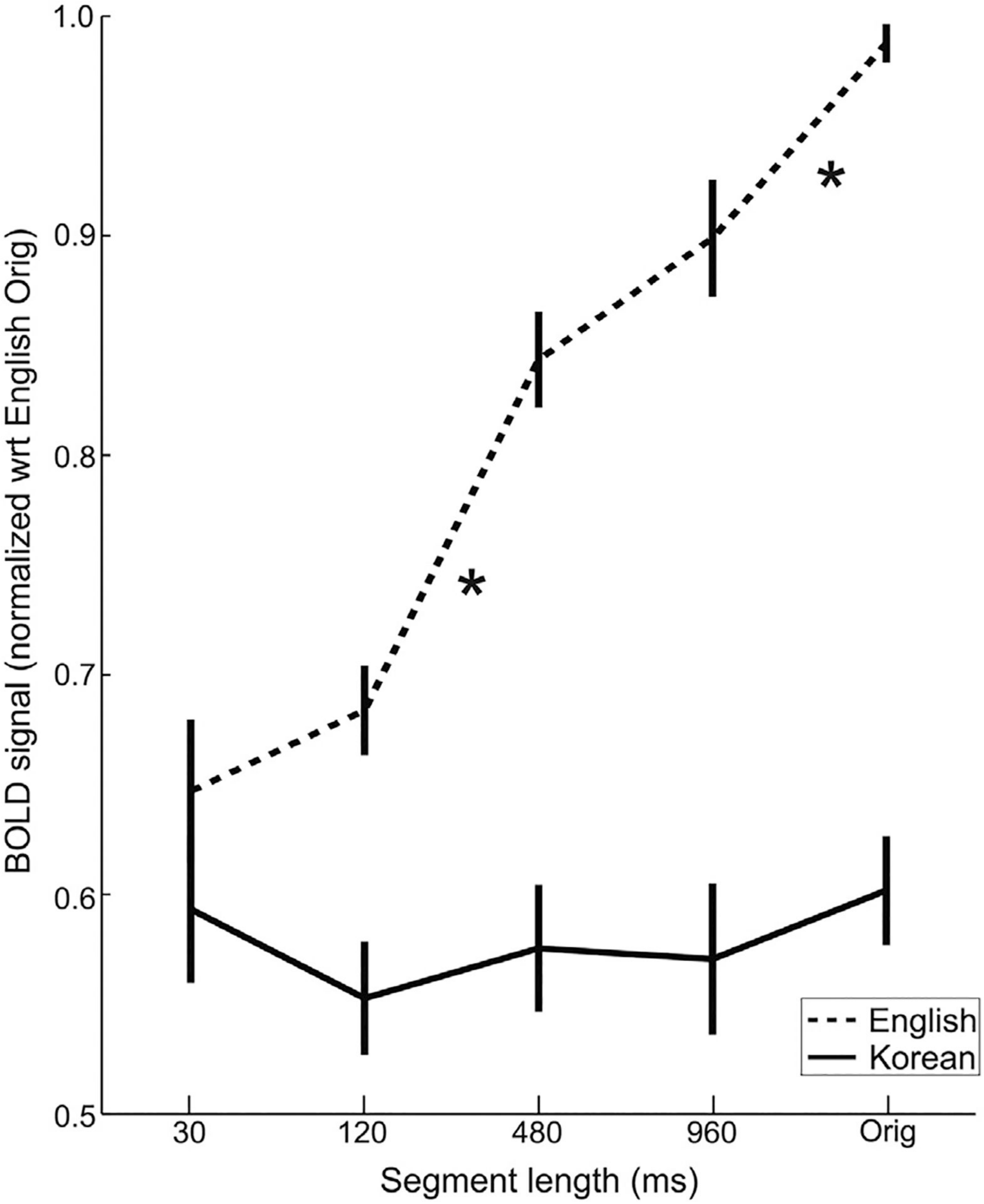

Fig. 7.

Response in individually defined (leave-one-out) fROIs in left IFG to speech quilts and original speech in English (dotted) and Korean (solid). Error bars denote ± 1 SEM. Asterisks denote significant pairwise comparisons ($p < 0.05$, Bonferroni corrected) between adjacent segment length levels within a language.

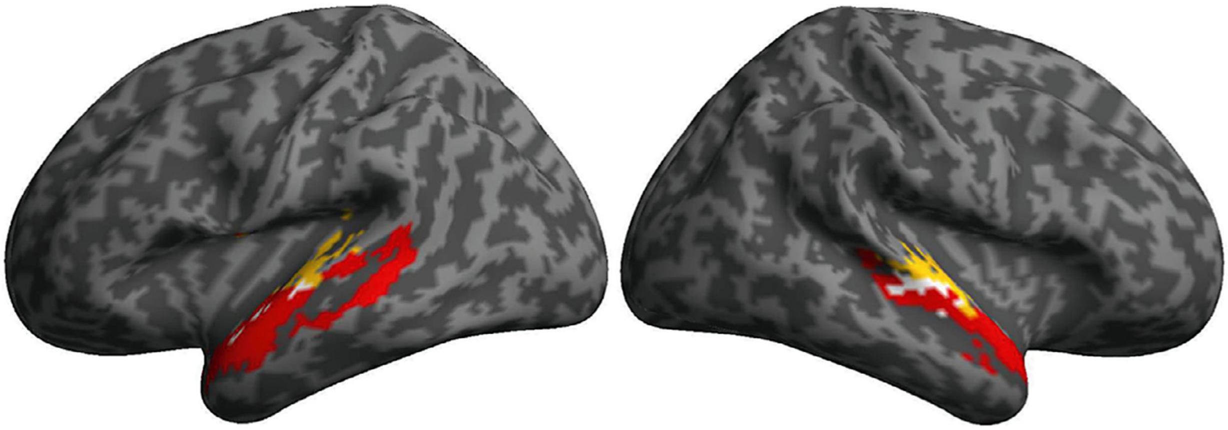

Fig. 8.

Regions (in yellow) showing significant modulation by segment length (in English), as revealed by the PPI analysis with a seed in left IFG. For reference, the areas highlighted in red show a stronger response to English original speech than English speech quilted with 30 ms segments (this is the same as the English fROI (RFX) in Fig. 3, top). Areas of overlap between the PPI results and the English fROI (RFX) are shown in white.