Shared and distinct representational dynamics of phonemes and prosody in ventral and dorsal speech streams

Authors: Seung-Cheol Baek[1,2,3], Seung-Goo Kim[1], Burkhard Maess[4], Maren Grigutsch[5], Daniela Sammler[1,5,*]

[1]Research Group Neurocognition of Music and Language, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany.

[2]Institute of German Linguistics, Philipps-University Marburg, Marburg, Germany.

[3]Center for Mind, Brain and Behavior, Universities of Marburg and Gießen, Germany.

[4]Methods and Development Group Brain Networks, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany.

[5]Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany.

*Correspondence should be addressed to Daniela Sammler (daniela.sammler@ae.mpg.de).

Abstract

Phonemes and prosodic contours are fundamental building blocks of speech used to convey complementary meanings. Perceiving these elements is thought to involve a neural abstraction from acoustic to categorical representations, occurring serially in time and segregated in space. Using magnetoencephalography and behavioural psychophysics combined with time-resolved representational similarity and multivariate transfer entropy analyses, we show that the abstraction of phonemic and prosodic information does not fully align with this serial hierarchical processing. Instead, we found partially overlapping acoustic and categorical representations in space and time, for both elements. Notably, building on similarly organised acoustic representations along the ventral stream, categorical representations of prosody extended to right premotor cortex along the dorsal stream, whereas those of phonemes focally involved posterior temporal regions. These results highlight shared principles of joint and distributed processing, yet partially distinct mechanisms for the abstraction of phonemes and prosody, key to access the multilayered meaning of speech.

Introduction

Speech conveys meaning at multiple levels. The propositional meaning coded by words—composed of phonemes—is often complemented by speaker meaning, that is, a speaker's communicative intentions conveyed by the tone of voice—the prosody—of an utterance[1]. A crucial first step to fully comprehend both levels is to extract the fundamental elements that underlie these meanings from the myriad of acoustic cues in the continuous speech stream. Transient acoustic patterns code phonemes[2], which are combined to form words, while characteristic patterns of fundamental frequency (F_0) form prosodic contours that mark, for example, whether a speaker intends to make a statement or asks a question[3,4]. Although these acoustic cues are merged and mixed in the dynamic continuous speech signal, human experience often involves the effortless abstraction of the respective cues into discrete categories of phonemes and prosodic contours[5,6]. However, the neural substrate underlying this abstraction over time remains poorly understood, along with how it compares between these linguistic elements that emerge from distinct physical cues.

Previous studies have produced a coarse anatomical map of frontotemporal cortical areas involved in phonemic and prosodic processing, organised into ventral and dorsal streams[7–11]. However, a more fundamental question than functional localisation that remains unanswered is how various linguistic elements are encoded in these dual-stream networks, that is, how they evolve in space, time, and across different levels of abstraction. Although evidence from functional neuroimaging and electrophysiology has provided some insight into where and when phonemic[12–18] and prosodic[19–21] information is represented, the limited spatial or temporal resolution of non-invasive methods and the restricted coverage of intracranial recordings hamper understanding the precise spatiotemporal dynamics of phonemic and prosodic representations, as well as their information flow between regions via feedforward or feedback connections. Moreover, are acoustic cues and categories of both phonemes and prosodic contours (henceforth: prosody) encoded serially in time and segregated in space along the dual streams—from primary to non-primary auditory regions

along the posterior-to-anterior axis of the temporal lobe[13,22–24] or from posterior temporal areas to motor and inferior frontal areas[25–27]—as suggested by classical hierarchical processing models[8,9,28], or alternatively, are different abstraction levels jointly represented in distributed areas[12,15,29,30]? These questions have major implications for understanding how the brain forms abstract linguistic representations from a dynamic acoustic signal.

In addition, whether the processing of phonemes and prosody relies on the same or distinct computational principles remains an open question. While traditional views posit shared acoustic analysis[8,28], phonemes and prosody are often claimed to tap into different hemispheric preferences[10–12,19,23]. Furthermore, recent intracranial studies indicate the segregated encoding of phonological features and pitch information within the ventral stream[21,31]. Although these findings are thought to reflect partially different mechanisms, it is still unclear to what degree neural infrastructures are shared or dissociated across the dual-stream networks in encoding phonemic and prosodic information at different levels of abstraction. Resolving this question is critical for understanding the nature and extent of specialisation in the human brain for the fundamental building blocks of speech at different time scales, which may allow their concurrent processing to comprehend spoken language.

To address these questions, we acquired magnetoencephalography (MEG) data from 29 participants during a psychophysical experiment with single words varying in word-initial phoneme and prosodic pitch contour. Building on participants' behaviour and MEG recordings, we developed a novel approach that combined time-resolved representational similarity analysis[32,33] (RSA) with multivariate transfer entropy[34,35] (mTE) analysis to reliably trace phonemic and prosodic representations at different levels of abstraction in space and time, as well as their information flow. We show that both phonemic and prosodic information is represented along the ventral stream, transferring from posterior to anterior temporal areas. Yet, prosody additionally recruits the dorsal stream, including bilateral premotor and right inferior frontal regions. Moreover, phoneme and prosody acoustics are represented slightly earlier than their categories, confirming the abstraction of information. However, the two abstraction levels are not as serially ordered and spatially segregated as typically

claimed, calling for a revision of classical hierarchical models of speech processing[8,9,28]. Apart from these similarities of acoustic representations of phonemes and prosody across the ventral stream, their categorical representations differ in spatial extent. Phoneme categories are represented locally in bilateral posterior temporal areas, and can be enhanced by top-down focus on phonemes, particularly in the left pSTS. In contrast, prosody categories are distributed more broadly, including bilateral anterior temporal and right premotor regions, implying partially different mechanisms of abstraction for prosody. Our findings demonstrate that the brain similarly abstracts phonemic and prosodic information, yet flexibly adapts regional involvement and information flow within the dual-stream networks to the respective cues and their associated linguistic functions. This may allow listeners to derive and disentangle complementary building blocks of speech merged in the auditory signal, as a fundamental step towards comprehending the multilayered meanings conveyed in the continuous speech stream.

Results

To track and compare the representational dynamics of phonemes and prosody in the brain, we created a stimulus set of 7,442 mono-syllabic words that varied in 61 evenly spaced steps along two orthogonal dimensions (middle panel of Fig. 1a): (i) the voice onset time (VOT) of the word-initial phoneme shifting from /b/ to /p/, and (ii) the pitch contour of the word shifting from statement (falling) to question (rising). Stimuli were generated via audio morphing of natural recordings of the German words, “Bar” (engl., “pub”) and “Paar” (engl., “pair”), uttered as both statement and question by two German native speakers (one female). A subset of 25 stimuli per speaker, that is, 5 x 5 morph levels out of the 61 x 61 steps, were selected individually for each of our 29 participants (Fig. 1a). The selection was based on participants’ performance in pre-tests (see Supplementary Information 1), which served to centre stimuli on the individual phonemic and prosodic category boundaries, and to match the difficulty of categorising the phonemes and prosody, for both speakers.

During the MEG recordings, participants were asked in 2-alternative forced-choice manner whether the speaker said “Bar” or “Paar”, or uttered “statement” or “question”, in alternating phoneme and prosody blocks presenting the 25 distinct stimuli of one speaker (Fig. 1b). The comparison of the slopes of the psychometric functions fitted to the proportions of “Paar” or “question” responses per speaker and morph level along the respective dimension confirmed that stimuli were matched in difficulty as intended[36]: Slopes differed neither between tasks ($F_{1,28} = 0.203,\ p > 0.05,\ \eta_G^2 = 0.002$) nor speakers ($F_{1,28} = 1.105,\ p > 0.05,\ \eta_G^2 = 0.012$), and no significant interaction was observed ($F_{1,28} = 1.385,\ p > 0.05,\ \eta_G^2 = 0.009$). Hence, the observed representational differences between phonemes and prosody below are unlikely to stem from differences in perceptual or processing demands. Moreover, the point of subjective equality (PSE), although different between tasks ($F_{1,28} = 39.852,\ p < 0.001,\ \eta_G^2 = 0.266$), did not differ between speakers ($F_{1,28} = 0.766,\ p > 0.05,\ \eta_G^2 = 0.000$; no task x speaker interaction: $F_{1,28} = 0.485,\ p > 0.05,\ \eta_G^2 = 0.004$), allowing us to pool data across speakers in the remaining analyses.

Categorical perception of phonemes and prosody. Next, we confirmed that the stimuli were perceived categorically. To do so, we compared the goodness of fit between linear and psychometric functions fitted to the behavioural responses (Fig. 1c-e). The psychometric functions provided a significantly better fit than the linear function ($F_{1,28} = 109.733,\ p < 0.001,\ \eta_G^2 = 0.173$; Fig. 1e), similarly in both tasks (no main effect of task: $F_{1,28} = 1.182,\ p > 0.05,\ \eta_G^2 = 0.014$; no task x function interaction: $F_{1,28} = 1.012,\ p > 0.05,\ \eta_G^2 = 0.003$). These findings indicate that both phonemes and prosody are perceived categorically; that is, continuous acoustic cues are mapped onto discrete categorical representations, similarly along both of our stimulus dimensions.

Phonemic and prosodic representations in ventral and dorsal stream regions. To first identify and compare where and when phonemic and prosodic information is generally

139 represented in the brain, we applied time-resolved RSA to MEG source activity pooled
140 across both tasks[32,33]. In this analysis, time-varying dissimilarities of neural activity patterns
141 evoked by the five phoneme or prosody levels were compared with modelled dissimilarities
142 of both stimulus acoustics and perceived categories (Fig. 2).

143 As illustrated in Fig. 2a, neural dissimilarities were quantified in 10 predefined regions
144 of interest (ROIs) across the two hemispheres, including the inferior frontal gyri (IFG) and
145 premotor cortices (PMC) in the dorsal stream, as well as the primary auditory cortices (PAC)
146 and the posterior (pSTS) and anterior superior temporal sulci (aSTS) in the ventral stream.
147 In addition, we introduced a whole-brain searchlight approach to complement the ROI-based
148 analysis. In each ROI or searchlight, we generated time series of 5 x 5 representational
149 dissimilarity matrices (RDMs) based on the cross-validated squared Euclidean distance
150 between the local multivariate response patterns to the five morph levels[37], separately for
151 phonemes and prosody, in 24-ms sliding time windows (Fig. 2b,c).

152 Model RDMs were built based on stimulus acoustics and perceived categories of the
153 stimuli using squared Euclidean distance. Stimulus acoustics were coded as the original
154 morph step of each stimulus on the 61-step phoneme or prosody continua. Perceived
155 categories were soft-coded as the expected proportion of “Paar” or “question” responses at
156 each of the five morph levels derived from the psychometric functions fitted to individual
157 behaviour (Fig. 2d). Based on these values, four model RDMs were constructed for each
158 participant, coding the acoustic and categorical dissimilarity of all 5 x 5 combinations of
159 morph levels, separately for phonemes and prosody (Fig 2e).

160 To assess neural representations of phonemes and prosody, the corresponding
161 model RDMs were used to predict each time step of the neural RDMs in a given region. We
162 first measured the variance explained by a linear combination of both acoustic and
163 categorical RDMs over time to identify general representations, and later applied variance
164 partitioning to tease apart acoustic and categorical representations (Fig. 2f). As a control,
165 surrogate and noise ceiling analyses validated that the variability in the neural RDMs was
166 largely accounted for by the combined acoustic and categorical RDMs: The observed

167 explained variance reached the empirical noise ceiling in all ROIs, and no significant
168 representations were found when the five morph levels were permuted (Supplementary Fig.
169 4).

170 Fig. 3a shows that phonemic information was represented exclusively in ventral
171 stream ROIs in both hemispheres, while dorsal stream areas remained silent. Phonemic
172 representations in all six temporal ROIs emerged as early as 70 ms after stimulus onset, that
173 is, shortly after voice onset (26–62 ms), with the first peak occurring in the right PAC at 108
174 ms post-onset, followed by peaks in the other regions, all aligned around 160 ms post-onset
175 (Table 1). Additionally, the bilateral PAC and pSTS exhibited sustained phonemic
176 representations after 300 ms post-onset, with those in the bilateral PAC persisting beyond
177 stimulus offset (464–500 ms). Converging results were found in the whole-brain searchlight
178 analysis (Fig. 3a, Supplementary Fig. 5a). Overall, these results indicate a high redundancy
179 of phonemic representations across bilateral temporal areas and time, raising the question
180 whether and how phonemic information propagates along the ventral stream.

181 Similar to phonemes, prosodic information was generally represented in bilateral
182 ventral stream areas, but extended further into dorsal stream regions, including the bilateral
183 PMC and right IFG (Fig. 3b). Prosodic representations emerged around 300 ms after
184 stimulus onset, that is, after pitch contours started to clearly differ between the five prosody
185 levels (184-250 ms, see Supplementary Information 2). The searchlight analysis confirmed
186 both the extent and the timing of prosodic representations (Fig. 3b, Supplementary Fig. 5b).
187 Comparable to phonemes, prosodic representations were strongly redundant in space and
188 time, although there were subtle regional differences in when they emerged (first in the
189 bilateral PAC and pSTS, last in the right PMC/IFG and left pSTS) or peaked (first in the right
190 PMC and right aSTS; Table 1). These subtle dynamics call for further analysis how prosodic
191 information propagates within and between the ventral and dorsal streams.

192

193 **Dynamic transfer of phonemic and prosodic representations along ventral and dorsal**
194 **streams.** Next, to elucidate how phonemic and prosodic information flows between the

195 cortical areas identified above, we employed a novel approach using mTE[34,35], a multivariate
196 extension of transfer entropy, which allowed us to assess directed representational transfer
197 between regions. We focused on the subset of ROIs that represented either phonemic or
198 prosodic information. To avoid mixing information over time, we first re-generated neural
199 RDMs in these ROIs at each time sample from 0 to 800 ms after stimulus onset, that is, we
200 re-computed dissimilarities between MEG response patterns evoked by the five phoneme or
201 prosody levels. Then, for each pair of ROIs, we quantified mTE between their time-resolved
202 neural RDMs. We considered one ROI in a pair as the source and the other as the target at
203 a time, and estimated mTE from the source to the target by lagging the target RDM series
204 ranging from 4 to 300 ms. At each time lag, the measured mTE was normalised (mTEn) by
205 the mean and standard deviation of the surrogate mTE distribution[38], obtained by cyclically
206 time-shifting the source RDM series[39,40].

207 Fig. 4 shows representational transfer between ventral stream areas for phonemes
208 (Fig. 4a-c, Supplementary Fig. 7a), and between both ventral and dorsal stream areas for
209 prosody (Fig. 4d-f, Supplementary Fig. 7b). Propagation of phonemic information was
210 extensive within and between hemispheres. Within each hemisphere, the PAC and pSTS,
211 that is, posterior temporal areas, were densely interconnected and projected anteriorly to the
212 aSTS. The left (but not right) aSTS showed feedback connections to the ipsilateral PAC and
213 pSTS. Furthermore, the PAC and pSTS generally projected to all contralateral areas,
214 including their homologues. The only connection not identified was from the left PAC to the
215 right pSTS. The aSTS showed less interhemispheric connections, projecting only to its
216 homologue. Interactions between posterior temporal areas tended to occur with shorter time
217 lags (within about 50 ms), while their influence on the aSTS (especially, the left aSTS) lasted
218 longer (up to 108 ms). These dynamics are compatible with the established hierarchical
219 organisation of the ventral stream with posterior-to-anterior information flow[22–24].

220 Prosodic representations displayed partially similar ventral stream dynamics as
221 phonemes, with some differences. Similar to phonemes, the PAC and pSTS reciprocally
222 exchanged prosodic information within each hemisphere, and the PAC projected anteriorly to

223 the aSTS in both hemispheres. However, pSTS-to-aSTS projections were confined to the left
224 hemisphere, and no anterior-to-posterior feedback connections were found. Moreover,
225 interhemispheric connections were relatively sparse: Exchange between posterior temporal
226 areas was confined to one projection from the right PAC to its left homologue, and only the
227 bilateral PAC (and the left pSTS) projected anteriorly to the contralateral aSTS. These
228 dynamics are, again, in line with posterior-to-anterior information transfer along the ventral
229 stream[22–24], but seem to highlight the relative importance of the PAC (compared to the pSTS)
230 as source.

231 In contrast to phonemes, prosodic representations were also transferred to dorsal
232 stream areas. Especially posterior temporal areas, including the bilateral PAC and left pSTS,
233 projected dorsally to the bilateral PMC and right IFG, without recurrent connections. The left
234 and right PMC communicated with each other, and both received top-down projections from
235 the right IFG. Interestingly, the PMC and aSTS, major target areas in the dorsal and ventral
236 streams, were bidirectionally connected within each hemisphere, and unidirectionally from
237 right to left between hemispheres. The bilateral aSTS also received input from the right IFG.
238 The transfer of prosodic representations took longer between more distant regions, with
239 delays of up to about 90 ms, especially from posterior temporal areas in the ventral stream
240 to the right IFG in the dorsal stream. Overall, these results demonstrate a strong cross-talk
241 between the ventral and dorsal streams, with the aSTS and PMC serving as central hubs in
242 the exchange of prosodic information.

243

244 **From acoustic to categorical representations of phonemes and prosody.** After having
245 shown how phonemic and prosodic representations are transferred along the ventral and
246 dorsal streams, we examined whether and how these representations shift from acoustics to
247 categories along these pathways and over time. To do so, we returned to the results of the
248 time-varying RSA (Fig. 3) and partitioned the explained variance in each ROI, complemented
249 by a whole-brain searchlight analysis. Given the high correlation between the acoustic and
250 categorical RDMs (Supplementary Fig. 3), we employed nested model comparisons to

estimate the variance uniquely explained by each model RDM at a given time point, separately for phonemes and prosody (Fig. 2f).

Phonemes were represented acoustically in all ROIs along the ventral stream (Fig. 5a), while categorical representations were restricted to the bilateral PAC and left pSTS (Fig. 5b). Both representations emerged early, within the first 200 ms after word onset, and overlapped in time. However, acoustic representations peaked consistently earlier (Table 2), suggesting a transition from acoustic to categorical representations over time[14,15,17]. These results were confirmed by the searchlight analysis, showing that acoustic representations extended more broadly and emerged slightly earlier in time than categorical representations (Supplementary Fig. 6a).

Prosody, similar to phonemes, was represented acoustically in all temporal ROIs, except for the left aSTS (Fig. 5c). However, categorical representations were spatially more widespread for prosody than phonemes, including all ventral stream areas as well as the right PMC (Fig. 5d). Notably, categorical representations also peaked nominally later (Table 2), similar to phonemes, and lasted longer than acoustic representations, in line with a gradual transfer from acoustics to categories[14,15,17], although acoustic and categorical representations showed considerable overlap in time. Interestingly, the right PMC exhibited only categorical, not acoustic, representations of prosody, in two time windows: one before and the other after stimulus offset. Again, all results were confirmed by the searchlight analysis (Supplementary Fig. 6b).

Lastly, to explore the lateralisation of phonemic and prosodic representations, we compared the time-varying variance uniquely explained by acoustic or categorical RDMs between homologous ROIs (Supplementary Fig. 8,9). None of the representations was significantly lateralised, neither for phonemes nor prosody (all cluster $ps > 0.05$).

Altogether, both the phoneme and prosody data suggest a fast transformation of acoustics into categories in the temporal lobe, although both abstraction levels co-existed in time, in line with recent studies showing preserved lower-level representations during higher-level processing stages[15,29]. However, the spatial dynamics of this transformation in the

ventral and dorsal streams differed between phonemes and prosody, suggesting partially different mechanisms of abstraction.

Task modulation effect on phonemic and prosodic representations. Next, we tested whether task demands to categorise phonemes or prosody intensified the respective neural representations. This idea builds on previous findings[9,41,42] showing, for example, that the task to identify phonemes or speakers in identical stimuli enhanced the encoding of task-relevant acoustic information[43]. To this end, we split the data by task (Fig. 1b), and estimated and compared the time-varying variance explained by the model RDMs in each ROI that had represented either phonemic or prosodic information (Fig. 3).

While the prosody task did not enhance any prosodic representations (all cluster $ps > 0.05$), the phoneme task enhanced the phoneme representations, selectively in the left pSTS, compared to the prosody task, around voice onset, peaking at 52 ms post-onset (Fig. 6a; $t_{28} = 3.334$, baseline-corrected Fisher Z variance contrast= 0.228, standard error of the mean (SEM)= 0.067). No such effect was found in other ROIs (all cluster $ps > 0.05$).

Next, we investigated which abstraction levels of phonemic representations in the left pSTS were modulated by task demands. Using variance partitioning, we compared the acoustic or categorical representations of phonemes in the left pSTS between the two tasks. While the acoustic representations were not modulated by task (Fig. 6b), the categorical phoneme representations were amplified during the phoneme task (Fig. 6c). This effect was evident again around voice onset (peak at 52 ms, $t_{28} = 3.345$, baseline-corrected Fisher Z unique variance contrast= 0.086, SEM= 0.025) and around 200 ms post-onset (peak at 200 ms, $t_{28} = 3.024$, baseline-corrected Fisher Z unique variance contrast= 0.081, SEM= 0.026).

Finally, we compared the task-modulation effect between the left and right pSTS. As depicted in Fig. 6d, the categorical representations of phonemes were enhanced by the phoneme task significantly more strongly in the left than right pSTS, again around voice onset, peaking at 52 ms ($t_{28} = 3.914$, baseline-corrected Fisher Z unique variance contrast=

306 0.138, SEM= 0.035). Overall, these findings highlight the special role of the left pSTS in
307 categorical phoneme perception and its modulation by task focus, in line with the
308 literature[23,31,41,43].

309 Discussion

310 Phonemes and prosody are fundamental building blocks of speech, perceived as abstract
311 categories arising from various acoustic cues within the dynamic input signal. Here, we
312 studied the neural implementation of this perceptual process by examining how phonemic
313 and prosodic information dynamically evolves from acoustic to categorical representations
314 throughout the brain and over time. Using a novel approach that combined behavioural
315 psychophysics and MEG with time-resolved RSA and mTE analysis, we uncovered three key
316 aspects of the representational dynamics of phonemes and prosody. First, the
317 spatiotemporal dynamics and information flow of phonemic and prosodic representations
318 replicated the dual-stream architecture for prosody and the involvement of the ventral stream
319 for phonemes. Specifically, both phonemic and prosodic information was represented along
320 the ventral stream, propagating from posterior to anterior temporal areas in both
321 hemispheres, while prosodic representations were additionally transferred along the dorsal
322 stream, including bilateral premotor and right inferior frontal regions. Second, both phonemic
323 and prosodic information was abstracted from acoustic to categorical representations that
324 overlapped, however, in space and time across different abstraction levels. Notably, while
325 acoustic representations of phonemes and prosody were distributed similarly across the
326 ventral stream, their categorical representations differed in spatial extent. Phoneme
327 categories were restricted to posterior temporal areas, whereas prosody categories were
328 more widespread, extending to anterior temporal and right premotor areas. Lastly, task
329 demands modulated the representations of phonemes but not prosody. Particularly, the
330 phoneme identification task enhanced the categorical representations of phonemes
331 selectively in the left posterior temporal region. These results establish a distributed

functional organisation of joint acoustic and categorical information processing along the ventral stream for both phonemes and prosody, and demonstrate an essential independent role of the dorsal stream in prosodic category processing.

Classical hierarchical models of speech perception assume that sound information is first received and acoustically analysed in primary areas, then transformed into more abstract linguistic representations towards non-primary, higher-order areas along a posterior-to-anterior gradient[8,9,23,24,28]. Despite the transfer of both phonemic and prosodic representations from posterior to anterior regions along the ventral stream, our results did not fully align with this serial hierarchical processing. First, we found that the PAC and pSTS represented information at similar latencies after relevant acoustic landmarks and exchanged this information bidirectionally, for both phonemes and prosody. This challenges the classical models of sequential information flow in the auditory cortex from primary to parabelt areas[9,22,24], and instead supports a recent view that these areas receive parallel input from subcortical regions and communicate with each another[31]. Second, the PAC represented categories of both phonemes and prosody along with their acoustic cues, suggesting that abstract linguistic representations can be formed at the earliest stage of cortical auditory processing[44,45]. This joint representation of acoustic and categorical information questions the traditional understanding of sequential hierarchical information abstraction. Indeed, evidence from functional neuroimaging with high magnetic field strength indicates that the superficial layers of primary areas perform more complex processes than simple auditory analysis, similar to non-primary areas[46], which may contribute to categorical representations of phonemes and prosody in the PAC. Third, acoustic representations were broadly distributed from posterior to anterior temporal areas, that is, well beyond primary regions, for both phonemes and prosody. These extended acoustic representations align with recent findings that acoustic-phonetic features are redundantly encoded throughout the temporal cortex alongside higher-level linguistic information[15], implying that acoustic details are not discarded after abstraction but are retained in higher-order regions. Such redundant

359 representations may benefit robust speech perception by facilitating the retrieval of acoustic
360 details from earlier elements for re-analysis[15,47].

361 Despite these similarities in successive, yet joint and distributed processing of
362 acoustic and categorical information, phonemes and prosody exhibited distinct
363 representational dynamics along the ventral stream. For example, we found that phonemes
364 and prosody differed in the spatial extent of their categorical representations. Phoneme
365 categories were represented locally in posterior temporal areas, including the bilateral PAC
366 and left pSTS. In contrast, representations of prosody categories extended more broadly
367 from posterior to anterior temporal areas, reaching bilaterally from the PAC and pSTS to the
368 aSTS. Given that phonemes are relatively short and often lack concrete meanings[48,49],
369 whereas prosody generally lasts longer and imparts meanings such as the speaker's
370 intention[3,4,50], this difference in spatial extent may be related to a gradient of temporal
371 integration windows along the temporal cortex which underlies meaning formation. Recent
372 intracranial studies have shown that neuronal populations in the temporal cortex are tuned to
373 auditory information at different time scales, with their temporal integration windows
374 enlarging along the posterior-anterior axis[48,51]. As these windows increase, neuronal
375 populations shift their encoding properties—from spectrotemporal cues to semantic
376 categories of sounds[48]. In light of this functional gradient, our findings suggest that
377 categorical representations of prosody extend to higher levels of the cortical hierarchy than
378 those of phonemes, reflecting differences in their duration and their association with specific
379 meanings.

380 Moreover, information flow between posterior and anterior temporal areas was
381 weighted differently between phonemes and prosody. While phonemic representations
382 transferred bilaterally from both the PAC and pSTS to the aSTS, the transfer of prosodic
383 representations was more pronounced from the PAC than from the pSTS to the aSTS,
384 indicating a relative dominance of the PAC in sending prosodic information. The PAC has
385 been shown to be sensitive to pitch information[52,53], which may underlie its stronger role in
386 transferring prosodic representations. Additionally, feedback connections were identified only

387 for phonemes, from the aSTS to the PAC and pSTS in the left hemisphere. Considering that
388 the left aSTS represents phoneme acoustics, these connections are thought to provide
389 acoustic feedback, facilitating the categorical representations in the left PAC and pSTS.

390 Phonemes and prosody also differed in their recruitment of the dorsal stream. Our
391 findings showed that phonemic representations were confined to temporal regions,
392 consistent with the perspective that the dorsal stream is not typically involved in sublexical
393 speech processing[9]. In contrast, representations of prosodic information extended to the
394 bilateral PMC and right IFG along the dorsal stream, replicating the dual-stream architecture
395 for prosody[10]. This difference in dorsal stream recruitment is unlikely to result from
396 differences in perceptual or processing demands[9], as the difficulty of identifying phonemes
397 and prosody was matched. More specifically, the right PMC represented prosody categories
398 in two distinct time windows: initially before and later after stimulus offset. This substantiates
399 the motor contribution to prosody categorisation[10], potentially involving multiple stages[54]. By
400 comparison, the right IFG displayed unspecific acoustic-categorical prosodic representations
401 after word offset, which raises the possibility that the IFG is involved in decision-making
402 relying on shared acoustic and categorical information[27,28].

403 At the same time, the dorsal and ventral stream extensively exchanged prosodic
404 information. The PMC received projections from posterior temporal areas and was
405 bidirectionally connected to the aSTS, which may reflect auditory-motor integration for the
406 categorical perception of prosody[54]. Similarly, the right IFG obtained afferent input from
407 posterior temporal areas but back-projected to the bilateral PMC and aSTS, possibly
408 modulating their representations[55]. These findings suggest that the PMC and aSTS serve as
409 central hubs for cross-stream interactions. Given that there is no direct anatomical
410 connection between these regions[8,10], future studies should explore whether the observed
411 functional connection is indirectly driven by common input from posterior temporal areas and
412 the right IFG[56], or possibly mediated by subcortical regions[57,58].

413 Our analysis further showed a disparity between phonemes and prosody regarding
414 the influence of task demands. Top-down efforts to identify phonemes enhanced their

415 categorical representation, selectively in the left pSTS and promptly after voice onset. This
416 may reflect an increased gain of neuronal populations in the left pSTS, which receive direct
417 input from subcortical areas and are non-linearly tuned to phonetic features[18,31,59]. Moreover,
418 this selective effect in the left pSTS aligns with a recent finding that disrupting the left
419 superior temporal area alone can impair spoken word recognition[31], indicating its crucial role
420 in phoneme perception. In contrast, no effect of task demands was observed on prosodic
421 representations, despite evidence that efforts to perceive prosody modulate cortical
422 activity[10,42]. This could be due to the less transient nature of acoustic prosodic cues, making
423 it less critical to capture a specific acoustic landmark in time. Prosodic information remained
424 informative after the pitch divergence point, allowing participants to accumulate sufficient
425 evidence over time, possibly irrespective of task. Correspondingly, neural responses to
426 prosody are likely to be less time-locked to a specific landmark, potentially further blurring
427 task effects. Future work should aim to replicate this result, such as with vowels, where
428 phonemic and prosodic cues overlap over time.

429 Interestingly, despite previously reported representational asymmetries between the
430 hemispheres[10–12,19,23], our results did not support the lateralised encoding of phonemic and
431 prosodic information at any level of abstraction. This may be ascribed to extensive cross-
432 hemispheric transfer of phonemic and prosodic representations, consistent with largely
433 bilateral neural involvement during speech processing[60].

434 Before closing, several caveats regarding our findings should be noted. First, we
435 considered only a single phonemic and prosodic cue among various distinctive features:
436 voice onset time and pitch contour. Generalising our findings to other phonemic and prosodic
437 contrasts requires further investigation in future research. Second, our novel mTE analysis
438 assessed the overall information flow of phonemic and prosodic representations, leaving
439 unresolved which levels of abstraction are transferred between regions over time. Moreover,
440 some connections might have arisen from common sources or reflected secondary links via
441 intermediate areas[56]. Future studies should address these potential indirect connections to
442 clarify the transfer of phonemic and prosodic representations, considering different levels of

abstraction. Lastly, we examined phonemes and prosody separately. While this approach allowed us to compare their representational dynamics, it offered limited insight into how different linguistic elements integrate, which is crucial for speech comprehension. Future work would benefit from exploring the interaction between phonemic and prosodic information in the brain to form a unified representation of speech.

Taken together, our findings reveal both shared and distinct representational dynamics of phonemes and prosody along the ventral and dorsal speech streams. The similarities along the ventral stream implicate general principles of representational abstraction of various acoustic cues into discrete categories. Notably, our results refine traditional models of speech processing[8,9,28] by showing joint and distributed instead of serially ordered and spatially segregated processing of acoustic and categorical information, potentially promoting robust speech perception. The different dynamics between phonemes and prosody along the ventral and dorsal streams indicate partially distinct mechanisms of abstraction, tailored to the respective cues and their associated linguistic functions. This flexible and complementary adaptation of regional weights and information flow to transient phonemic and pitch-based prosodic cues is likely to underlie listeners' ability to concurrently process and grasp the richness of meanings conveyed by different linguistic elements in the continuous speech stream.

Methods

Participants. 34 native German speakers (17 females, mean age= 25.62, SD= 3.81) were recruited and provided written informed consent prior to participation. All participants self-reported normal hearing with no history of neurological or psychiatric disorders, and were assessed as right-handed by the Edinburgh Handedness Inventory (mean laterality quotient= 92.38, SD= 8.57)[61]. Five participants (two females) were excluded due to poor behavioural performance or low MEG signal quality, leaving 29 participants for the analyses. This study adhered to the ethical standards of the Declaration of Helsinki and was approved by the Ethics Committee of the Medical Faculty, Leipzig University (403/14-ff).

Stimuli. Stimuli consisted of the German words “Bar” and “Paar”, which gradually varied in both phonemic and prosodic cues. Mono-syllabic words were used to minimise the involvement of compositional linguistic processing, such as syntactic and semantic processes[10]. For stimulus development, we first recorded “Bar” and “Paar” with both statement and question intonations, produced by two German native speakers (one female) in a soundproof booth using a Røde NT55 microphone and Audacity® (version 2.0.5, sampling at 16 kHz, 16-bit, mono). We then segmented the recordings into word-initial phonemes (/b/ and /p/) and word stems, and used them to separately morph the VOT and pitch contour for each speaker in MATLAB (R2014a, MathWorks).

We first morphed the pitch contour of the word stems from statement to question using STRAIGHT[62]. For this, we placed anchor points on corresponding time-frequency landmarks in the two utterances (see ref.[10]), and decomposed the sound signals into five features (F_0, frequency structure, duration, spectrotemporal density, and aperiodicity). We then re-synthesised new word stems by interpolating these features in 2% steps on linear (duration and aperiodicity) and logarithmic (F_0, frequency structure, and spectrotemporal density) scales. The resynthesis continued up to 120% of the original question by extrapolating the landmarks, resulting in 61 word stems with a duration of 430 ms. Next, we

488 morphed the VOT by manually splicing incrementally longer segments (1-ms steps) of
489 aspiration noise from the original /p/ between the plosive (/b/) and the word stems in 61
490 steps, starting from the original /b/ (0-ms VOT) reaching up to 60-ms VOT. The splicing
491 produced 61 VOT morphs. Applied to each of the 61 word stems, this resulted in 3,721
492 stimuli for each speaker (7,442 stimuli in total), varying along 61 x 61-step orthogonal
493 morphing continua of VOT and pitch contour (middle panel of Fig. 1). Additionally, we
494 inserted a 10-ms silence at the beginning of the stimuli and applied a 10-ms fadeout at the
495 end to prevent unwanted clicks. All stimuli were normalized in root-mean-square amplitude.
496 Out of the 61 x 61 steps, 5 x 5 morph levels were selected for each participant based
497 on two behavioural pre-tests (see Supplementary Information 1), leading to a subset of 25
498 stimuli per speaker (50 stimuli in total) presented during the MEG experiment. The first pre-
499 test used an adaptive-staircase method to identify five levels of VOT and pitch contour
500 centred on individual category boundaries. The distance between the five levels was
501 selected to ensure that the difficulty of categorising phonemes and prosody was matched.
502 Stimulus selection was validated in the second pre-test and adjusted, if necessary.
503
504 **Experimental procedure during the MEG experiment.** Stimuli were delivered via air-
505 conduction earphones (ER3-14A/B, Etymotic Research) at an average of 74 dB SPL. The
506 volume was adjusted based on each participant's hearing threshold, measured before the
507 experiment proper using a stimulus that was not presented later.
508 The MEG experiment consisted of six runs, each of which contained four blocks.
509 Across the four blocks, participants were alternately asked to identify either the word
510 ("Bar"/"Paar") or the prosody ("statement"/"question") of the stimuli (Fig. 1b). The stimuli of
511 the same speaker were presented in two blocks in a row per run, but with different tasks. For
512 example, all stimuli of the female speaker were first presented with a phoneme, then with a
513 prosody task, followed by a phoneme and a prosody block with male stimuli. Order of task
514 and speaker per run was randomised and counterbalanced across participants.

Each block presented 29 stimuli in separate trials, that is, the 25 stimuli of the speaker preceded by four practice trials. These practice trials consisted of stimuli with extreme levels of phonemes and prosody allowing participants to adapt swiftly to a given task. They were not considered in the analyses. The 25 stimuli per block (i.e., 5 x 5 morph levels) were presented pseudorandomly following the principles of a type-1 index-1 sequence to prevent carry-over and position effects[63].

Participants were asked to carefully listen to each word and mentally categorise whether the speaker said “Bar” or “Paar” or uttered “statement” or “question”, depending on the task. The task was indicated on screen at the beginning of each block and the colour of the fixation cross during the block reminded participants of the type of task (green for phoneme and red for prosody). After one-sixth of the stimuli (i.e., twice per stimulus and task in random order), participants were asked to indicate their response by pressing a button on a response box. To avoid confounding categories and preparatory motor activity, response buttons were randomly assigned in each response trial by presenting the labels “Bar” and ”Paar” (or “statement” and ”question”) on the left or right side of the screen after stimulus offset (balanced across stimuli, tasks, and speakers). Stimulus onset asynchrony was 2800 ms with a random jitter between -500 and 500 ms. Response trials included an additional 2000-ms timeout for the button press, and button presses were followed by a 1000-ms pause before the onset of the next stimulus. The entire procedure was controlled by Presentation® software (version 18.0, Neurobehavioral Systems) and lasted about two hours, including optional breaks between the runs.

Behavioural analysis. For each participant, we computed the proportions of “Paar” and “question” responses at the five phoneme and prosody levels for each speaker across the six runs. Four participants in whom the difference between minimal and maximal proportion was smaller than 0.6 in at least one of the tasks were excluded from the analysis, as they either perceived no clear question/Paar or no clear statement/Bar.

The proportions at the five morph levels were then used to fit psychometric curves to each participant's data per task and speaker using the following expression:

$$\hat{\boldsymbol{y}} = \frac{1}{1 + \exp[(a - \boldsymbol{x})\, b]}. \tag{1}$$

$\boldsymbol{x}$ and $\hat{\boldsymbol{y}}$ are column vectors, representing the five levels and the estimated proportions, respectively. a is the point where the curve predicts equal proportions for both categories, and b is the slope of the curve at a. The fitted 50%-point a and slope b were taken as the individual estimates of PSE and perceptual difficulty. The PSE and perceptual difficulty were separately compared between tasks and speakers at the group level using repeated measures analyses of variance (rmANOVAs). As no speaker-related effects were found for both measures, we averaged the behavioural responses across speakers for the subsequent analyses.

A further analysis was conducted to verify the categorical perception of phonemes and prosody. For each task, we fitted a linear function $\hat{\boldsymbol{y}} = a\boldsymbol{x} + b$, and a psychometric curve, as in (1), to each participant's responses, and measured the adjusted R^2 as a metric for their goodness of fit. The adjusted R^2 between the linear and curve fits was compared at the group level using an rmANOVA, with function (linear vs. curve) and task (phoneme vs. prosody) as factors, to assess whether the increase of response proportions across the five levels better followed a psychometric curve rather than a linear function for both phonemes and prosody.

MEG data acquisition and preprocessing. MEG data were collected during each run using a Neuromag Vectorview system (Elekta) with 306 sensors (102 magnetometers and 204 gradiometers). The MEG signals were sampled at a rate of 1,000 Hz, with a passband from DC to 330 Hz, while participants' head positions were tracked using five head coils.

The acquired MEG data were preprocessed using MNE-Python (version 1.1.0). For each run, external interference was suppressed by spatiotemporal signal-space separation, including correction for head movements and semi-automatically identified noisy channels

569 (mean= 3.74 sensors, SD= 2.69). The signals were then lowpass-filtered at 80 Hz using a
570 finite impulse response (FIR) filter with a 10-Hz transition bandwidth (one-pass, zero-phase
571 shift, non-causal, Hann window), followed by the reduction of line noise (50 Hz) and its
572 harmonics using ZapLine[64]. The low-frequency range below 0.3 Hz was further attenuated
573 with a highpass filter (a one-pass, zero-phase-shift, non-causal FIR filter using a Hann
574 window with a transition bandwidth of 0.3 Hz) that was designed to replace the need for
575 baseline correction[65]. The filtered signals were downsampled to 250 Hz and subjected to
576 independent component analysis (ICA) to remove artefacts related to physiological noise,
577 such as eye movements and heartbeats. The ICA solution was derived from the data
578 concatenated across all six runs for each participant. After the artefact removal, the signals
579 were segmented into epochs from -200 to 800 ms around stimulus onset. Epochs with
580 excessive noise were discarded by Autoreject[66], after which one participant was excluded for
581 having more than 20% of epochs rejected. On average, 1,191 epochs (SD= 44.28) were
582 retained for each participant across the six runs.
583
584 **Source reconstructions.** Source activity was reconstructed from the preprocessed MEG
585 signals with anatomical constraints, derived from high-resolution individual T1-weighted
586 images (1 x 1 x 1-mm^3 voxel size) obtained using a 3-T magnetic resonance imaging (MRI)
587 scanner (Siemens). Anatomical images were segmented using FreeSurfer (version 7.3.2) to
588 construct a volume conduction model and source space for each participant. The volume
589 conduction was estimated by a single-layer boundary element model created with the
590 watershed algorithm applied to the inner skull. The source space comprised 10,242 vertices
591 per hemisphere along the cortical surface, where dipole orientations were defined as the
592 surface normal with a loose constraint. The anatomical image was aligned with the sensor
593 space by coregistering the individual MEG and MRI data through iterative procedures based
594 on the fiducials and digitised points on the head model. The sensor signals were then
595 projected onto the source space using eLORETA[67] (regularisation parameter= $\frac{1}{9}$), while

being whitened by a covariance matrix of sensor noise. The noise covariance matrix was estimated from the prestimulus period (-200 to 0 ms around stimulus onset) of all epochs and regularised by the Ledoit-Wolf method with cross-validation[68]. Source reconstructions were performed on the evoked responses to each of the five phoneme or prosody levels, averaged across the corresponding epochs for each speaker and run when estimating task-agnostic source activity, as well as for each task when estimating task-specific source activity. The estimated source activity in the individual source space was finally morphed into the common *fsaverage* space for subsequent analyses.

Time-resolved RSA. Time-resolved RSA was employed to identify where and when phonemes and prosody were represented in the brain and at which level of abstraction, that is, in terms of their acoustics or categories[32,33]. This analysis involved constructing RDMs based on multivariate neural dissimilarities in specific regions and across time, reflecting the relationship between stimuli in the local representational space. To delineate the characteristics of the representational structure, neural RDMs were compared with model RDMs built on acoustic and perceptual dissimilarities of the stimuli, respectively. We performed time-resolved RSA in individual participants and analysed phoneme and prosody separately to compare their representational dynamics in space and time, as well as between different abstraction levels.

Neural RDMs were primarily constructed from source activity in 10 predefined ROIs, but a whole-brain searchlight analysis complemented the ROI-based approach (Fig. 2a). The ROIs included the bilateral IFG, PMC, PAC, pSTS, and aSTS, covering areas previously shown to be involved in processing phonemes and prosody[10,12,30]. We defined the ROIs to be comparable in size and spatially distinct to minimise cross-talk, based on a multimodal cortical parcellation[69]. A searchlight with 10-mm radius was shifted across all vertices in the source space. For both approaches, sliding time windows of 24 ms were used to generate time series of neural RDMs in specific regions. Time windows advanced in 4-ms steps for the ROIs, while half-window-sized steps were applied in the searchlight analysis to reduce

computational load. In each ROI or searchlight location, we extracted task-agnostic MEG source activity patterns evoked by the five phoneme or prosody levels within a given time window for each speaker and run, and measured the squared Euclidean distance between all level pairs within each speaker, cross-validated across the six runs[37] (Fig. 2b). The measured neural dissimilarities were then averaged across speakers, yielding a time series of 5 x 5 neural RDMs per region and participant, separately for phonemes and prosody (Fig. 2c).

Meanwhile, model RDMs were built for stimulus acoustics and perceived categories (Fig. 2d). Stimulus acoustics were coded as the original morph steps of the stimuli along the 61-step phoneme or prosody continuum for each speaker, which preserved the statistics of stimulus acoustic features (see Supplementary Information 3). Perceived categories were modelled as the expected proportion of “Paar” and “question” responses at each of the five morph levels, derived from the psychometric curves fitted to individual behaviour pooled across speakers. Both acoustic and categorical dissimilarities between the five phoneme or prosody levels were estimated by calculating the squared Euclidean distance between the respective values, producing 5 x 5 RDMs, each normalised to a range between 0 and 1. Finally, we averaged two RDMs encoding acoustic dissimilarities from the different speakers, resulting in one acoustic and one categorical RDM per participant for both phonemes and prosody (Fig. 2e).

We quantified neural representations of phonemes and prosody by comparing the corresponding neural and model RDMs using hierarchical linear modelling. Specifically, three types of nested linear models were built to predict each time step of the neural RDMs in a given ROI or searchlight, as follows:

- Full model: neural RDM ~ intercept + acoustic RDM + categorical RDM
- Reduced model: neural RDM ~ intercept + (acoustic OR categorical) RDM
- Null model: neural RDM ~ intercept

Before entering the linear models, all RDMs were vectorised based on the lower triangular entries, as they were symmetric. The model fits (R^2) were estimated by fitting formula (2) below using non-negative least squares (NNLS), which prevented the negative weights for that negative distances in the neural RDM could arise from cross-validation:

$$\boldsymbol{W}\boldsymbol{d} = \boldsymbol{W}\boldsymbol{M}\boldsymbol{\beta} + \boldsymbol{e}. \tag{2}$$

$\boldsymbol{W}$ is a pre-whitening matrix multiplied to both the neural RDM column vector $\boldsymbol{d}$, and the predictor matrix $\boldsymbol{M}$. $\boldsymbol{\beta}$ is a column vector of weights to be estimated, corresponding to the number of predictors. $\boldsymbol{e}$ is the error term. The pre-whitening matrix $\boldsymbol{W}$ was introduced to mitigate the problem that the distance estimates in the RDMs were correlated due to the shared morph levels among them, leading to sub-optimal comparisons between the neural and model RDMs[37]. Thus, we constructed $\boldsymbol{W}$ from a noise covariance matrix along the five morph levels in a given region, following a recent proposal[37].

The estimated R^2 from the nested models was used to assess general representations, as well as unique acoustic and categorical representations (Fig. 2f), as shown below:

$$\hat{R}_G^2 = \max\left(\hat{R}_{full}^2 - \max\left(\hat{R}_{null}^2, 0\right), 0\right), \tag{3}$$

$$\hat{R}_U^2 = \max\left(\hat{R}_{full}^2 - \max\left(\hat{R}_{reduced}^2, 0\right), 0\right). \tag{4}$$

The $\max(\cdot, 0)$ operator zeros any negative R^2 that could result from NNLS fitting. While R^2, in principle, cannot be negative, NNLS can yield a value below zero when the predictors explain less variance in the dependent variable than its mean, due to the non-negative weight constraint. In our case, this occurs particularly when the mean of distance estimates in a given neural RDM is negative. $\hat{R}_G^2$ is a measure of general representations, estimating the variance explained by both model RDMs. $\hat{R}_U^2$ represents the unique variance explained by the acoustic or categorical RDM that is not included in the Reduced model. The nested model comparison, following (4), enabled us to partition the variance explained by the highly correlated predictor RDMs $\hat{R}_G^2$, into their individual unique contributions (Supplementary Fig. 4). The estimated $\hat{R}_G^2$ and $\hat{R}_U^2$ for each acoustic and categorial RDM were square-rooted and

then Fisher-Z-transformed for visualisation and group-level inference, followed by baseline correction per region and participant to account for bogus effects driven by free parameters.

Additionally, time-resolved RSA was performed on surrogate data to validate the observed general representations of phonemes and prosody in each ROI (Supplementary Fig. 4). The surrogate data were created by permuting the corresponding five morph levels for each participant. All possible permutations of the five levels were considered, except for the original order ($5! - 1 = 119$). We measured the variance of a level-permuted neural RDM explained by both model RDMs at each time step, as described above, and took the median of the explained variance across permutations as the estimate of surrogate general representations.

Lastly, we estimated the noise ceiling along the dimension of either phonemes or prosody in each ROI by calculating the lower and upper bounds of the structured variance in the corresponding neural RDM at each time step. The lower and upper bounds were computed for each participant following (3), while the Full models were defined as follows:

$$\boldsymbol{d}_i \sim 1 + \overline{\boldsymbol{d}}_{-i}, \tag{5}$$

$$\boldsymbol{d}_i \sim 1 + \overline{\boldsymbol{d}}, \tag{6}$$

where $\boldsymbol{d}_i$ represents the neural RDM for the i-th participant. The lower bound was estimated by predicting $\boldsymbol{d}_i$ using the intercept and $\overline{\boldsymbol{d}}_{-i}$, the neural RDM averaged across participants excluding the i-th participant as in (5). The upper bound was derived by predicting $\boldsymbol{d}_i$ using the intercept and $\overline{\boldsymbol{d}}$, the grand-averaged neural RDM as in (6). The computed bounds were then averaged across participants, yielding a time-varying range of explainable variance in a given ROI (Supplementary Fig. 4).

mTE analysis. To investigate how phonemic and prosodic information propagates between the involved ROIs, we applied a novel mTE analysis to time-resolved neural RDMs. This analysis employed a recently proposed framework that combines Rényi's α-order entropy, a generalisation of Shannon entropy, with a kernel trick to measure directed information

transfer between two multivariate time series without estimating their individual or joint probability density functions[34,35].

In this framework, the entropy of an arbitrary-dimensional time series $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T\}$ is measured as follows[34]:

$$H_\alpha(\boldsymbol{A}) = \frac{1}{1-\alpha}\log_2\left(tr(\boldsymbol{A}^\alpha)\right) = \frac{1}{1-\alpha}\log_2\left[\sum_{i=1}^{T}\lambda_i(\boldsymbol{A})^\alpha\right], \tag{7}$$

$$\boldsymbol{A} = \frac{\boldsymbol{K}}{tr(\boldsymbol{K})}, \tag{8}$$

where $\boldsymbol{K}$ is the Gram matrix derived from the Gaussian kernel function $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right)$. $\boldsymbol{A}$ is the normalised Gram matrix, and $\lambda_i(\boldsymbol{A})$ is the i-th eigenvalue of $\boldsymbol{A}$. $H_\alpha(\boldsymbol{A})$ is a matrix-based analogue to Rényi's α-order entropy, which approaches Shannon entropy as $\alpha \to 1$. Given two multidimensional time series $\boldsymbol{X}$ and $\boldsymbol{Y}$ of length T, the matrix-based Rényi's α-order joint entropy can be written as[34]:

$$H_\alpha(\boldsymbol{A_X}, \boldsymbol{A_Y}) = H_\alpha\left(\frac{\boldsymbol{A_X} \odot \boldsymbol{A_Y}}{tr(\boldsymbol{A_X} \odot \boldsymbol{A_Y})}\right), \tag{9}$$

where $\boldsymbol{A_X}$ and $\boldsymbol{A_Y}$ are the normalised Gram matrices by applying (8) to the Gram matrices $(\boldsymbol{K_X})_{ij} = \kappa_{\boldsymbol{X}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $(\boldsymbol{K_Y})_{ij} = \kappa_{\boldsymbol{Y}}(\boldsymbol{y}_i, \boldsymbol{y}_j)$, respectively. The symbol $\odot$ represents the Hadamard product. Based on (7) and (9), mTE can be defined as[35]:

$$mTE(\boldsymbol{X} \to \boldsymbol{Y}) = H_\alpha(\boldsymbol{A_{X_t}}, \boldsymbol{A_{Y_t}}) - H_\alpha(\boldsymbol{A_{X_t}}, \boldsymbol{A_{Y_t}}, \boldsymbol{A_{Y_{t+l}}}) + H_\alpha(\boldsymbol{A_{Y_t}}, \boldsymbol{A_{Y_{t+l}}}) - H_\alpha(\boldsymbol{A_{Y_t}}), \tag{10}$$

where $\boldsymbol{A_{X_t}}$ and $\boldsymbol{A_{Y_t}}$ are the normalised Gram matrices of multivariate time series $\boldsymbol{X}$ and $\boldsymbol{Y}$ at time t, and $\boldsymbol{A_{Y_{t+l}}}$ is the one based on $\boldsymbol{Y}$ at future time $t + l$.

Using (10), we assessed how informative the past RDMs in a source ROI $\boldsymbol{X_t}$ are for predicting the future RDMs in a target ROI $\boldsymbol{Y_{t+l}}$ beyond its own past RDMs $\boldsymbol{Y_t}$. This approach is similar to Granger causality but can also capture non-linear interactions, including representational transformations from acoustics to categories. Here, we set the bandwidth of the Gaussian kernel function σ as the median distance between the time samples[70], and used $\alpha = 2$ for neutral weighting of rare events[34].

728 mTE analysis was performed on a subset of ROIs that showed significant general
729 representations of either phonemes or prosody (Fig. 3), depending on the analysis of
730 interest. For each of these ROIs, we first re-generated neural RDMs at each time sample
731 from 0 to 800 ms after stimulus onset to avoid mixing information over time. Neural RDMs
732 were built as described in *Time-resolved RSA*, by computing neural pattern dissimilarities
733 between all pairs of the five morph levels. These RDMs were vectorised and adaptively
734 whitened, yielding a multivariate time series per ROI and participant, separately for
735 phonemes and prosody. Next, we measured mTE between the time-resolved neural RDMs
736 for pairs of ROIs, with each ROI alternatively serving as the source and target. The target
737 RDM series was lagged by l to introduce time asymmetry, with the lag l varying from 4 to
738 300 ms to cover representational transfer across a broad time range. mTE was estimated at
739 each time lag while ensuring that the number of time samples used remained equal.
740 Notably, the measured mTE values are non-negatively biased and differ in scales, as
741 their upper bounds are determined by the entropy of the future RDMs in a target ROI
742 conditioned on its own past: $H_\alpha(\boldsymbol{A}_{\boldsymbol{Y}_{t+l}}|\boldsymbol{A}_{\boldsymbol{Y}_t}) = H_\alpha(\boldsymbol{A}_{\boldsymbol{Y}_t}, \boldsymbol{A}_{\boldsymbol{Y}_{t+l}}) - H_\alpha(\boldsymbol{A}_{\boldsymbol{Y}_t})$ in (10)[38]. These
743 properties complicate the comparison of mTE values and statistical inference at the group
744 level. Therefore, we normalised the mTE values by the mean and standard deviation of the
745 surrogate mTE distribution, obtained by cyclically time-shifting the source RDM series based
746 on a point exhaustively drawn from the interval between the 5th and 95th percentiles of the
747 time samples used[39,40]. This procedure produced the normalised mTE (mTEn) value for each
748 time lag, directed connection, and participant, separately for phonemes and prosody.
749
750 **Lateralisation analysis.** Lateralisation analysis was performed to determine whether
751 phonemic and prosodic representations at different abstraction levels were lateralised along
752 ventral and dorsal stream regions. Specifically, we calculated the difference in the time
753 series of baseline-corrected Fisher-Z variance uniquely explained by either acoustic or

categorical RDM between homologous ROIs for each participant, separately for phonemes and prosody.

Analysis of task-modulation effect. Here, we assessed whether phonemic and prosodic representations were enhanced by relevant task demands. First, we examined the task-modulation effect on the general representations of either phonemes or prosody within the subset of ROIs identified in the previous analysis (Fig. 3). For these ROIs, we ran time-resolved RSA on task-specific MEG source activity patterns to estimate the time-varying variance explained by both model RDMs during the phoneme and prosody identification tasks. We used the task-agnostic whitening matrix $\boldsymbol{W}$ for both tasks, as there was no reason to assume a difference in the noise covariance structure across them. We then computed the difference between the tasks for each ROI and participant, separately for phonemic and prosodic representations.

Next, we examined the task-modulation effect at different abstraction levels, focusing on phonemic representations in the left pSTS, as no other effects were observed in the previous step. For each task, we estimated the time-varying unique variance explained by each acoustic and categorical RDM using variance partitioning. We then calculated the difference between the tasks for each model RDM and participant.

Finally, we tested the lateralisation of the task-modulation effect on the categorical representations of phonemes in the left pSTS, as no effect was observed for the acoustic representations. We calculated the difference between the tasks for the categorical RDM in the right pSTS and contrasted this difference with that in the left pSTS for each participant.

The (unique) variance explained, measured in the analysis, was converted to the Fisher-Z scale and baseline-corrected before being compared between the tasks.

Statistical inference. For group-level inference in all MEG data analyses, statistical significance was determined using cluster-based permutation tests. Specifically, either temporal or spatiotemporal clusters were formed by one-sided t-tests against zero with a

threshold of $p < 0.05$, and family-wise error rate (FWER) was controlled at $p < 0.05$ across time points per ROI, across time lags per directed connection in the mTE analysis, or across time points and vertices in the searchlight analysis. To assess the difference between the two tasks, we performed one-sided instead of two-sided t-tests as we anticipated that task demands would enhance relevant representations. Two-sided t-tests were used to compare homologous ROIs when testing for lateralisation. Permutation distributions were generated by randomly shuffling signs or condition labels 100,000 times in the ROI-based, and 10,000 times in the searchlight analyses. Comparisons across multiple ROIs or directed connections were controlled at a false discovery rate (FDR) of $q < 0.05$, based on the minimum cluster p-value for each contrast.

Data availability

The stimuli presented and the behaviour responses of individual participants collected during the MEG experiment, as well as the preprocessed neural data—including the neural RDMs and time-resolved RSA and mTE results based on the ROIs—are publicly available and can be downloaded via the following repository: https://github.com/SeungCheolBaek/representation_dynamics_phonemes_prosody.

Code availability

The code to replicate the ROI-based RSA and mTE analysis, as well as the main figures, is publicly available in the following repository: https://github.com/SeungCheolBaek/representation_dynamics_phonemes_prosody.

References

1. Hellbernd, N. & Sammler, D. Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *J Mem Lang* **88**, 70–86 (2016).
2. Pisoni, D. B. & Luce, P. A. Acoustic-phonetic representations in word recognition. *Cognition* **25**, 21–52 (1987).
3. Hirschberg, J. & Pierrehumbert, J. The intonational structuring of discourse. in *24th Annual Meeting of the Association for Computational Linguistics* 136–144 (1986).
4. Pierrehumbert, J. & Hirschberg, J. *The Meaning of Intonational Contours in the Interpretation of Discourse*. *Intentions in Communication* (The MIT Press, 1990). doi:10.7551/mitpress/3839.001.0001.
5. Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. Perception of the speech code. *Psychol Rev* **74**, 431–461 (1967).
6. Xie, X., Buxó-Lugo, A. & Kurumada, C. Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition* **211**, (2021).
7. Belyk, M. & Brown, S. Perception of affective and linguistic prosody: An ALE meta-analysis of neuroimaging studies. *Soc Cogn Affect Neurosci* **9**, 1395–1403 (2013).
8. Friederici, A. D. The brain basis of language processing: from structure to function. *Physiol Rev* **91**, 1357–1392 (2011).
9. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat Rev Neurosci* **8**, 393–402 (2007).
10. Sammler, D., Grosbras, M. H., Anwander, A., Bestelmeyer, P. E. G. & Belin, P. Dorsal and ventral pathways for prosody. *Current Biology* **25**, 3079–3085 (2015).
11. Saur, D. *et al. Ventral and Dorsal Pathways for Language*. www.pnas.org/cgi/content/full/ (2008).
12. Gong, X. L. *et al.* Phonemic segmentation of narrative speech in human cerebral cortex. *Nat Commun* **14**, (2023).
13. de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L. & Theunissen, F. E. The hierarchical cortical organization of human speech processing. *Journal of Neuroscience* **37**, 6539–6557 (2017).
14. Brodbeck, C., Hong, L. E. & Simon, J. Z. Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current Biology* **28**, 3976-3983.e5 (2018).
15. Keshishian, M. *et al.* Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. *Nat Hum Behav* **7**, 740–753 (2023).
16. Khalighinejad, B., da Silva, G. C. & Mesgarani, N. Dynamic encoding of acoustic features in neural responses to continuous speech. *Journal of Neuroscience* **37**, 2176–2185 (2017).
17. Di Liberto, G. M., O'Sullivan, J. A. & Lalor, E. C. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology* **25**, 2457–2465 (2015).
18. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. *Phonetic Feature Encoding in Human Superior Temporal Gyrus*. *28 FEBRUARY* vol. 343 www.sciencemag.org (2014).
19. Brodbeck, C. & Simon, J. Z. Cortical tracking of voice pitch in the presence of multiple speakers depends on selective attention. *Front Neurosci* **16**, (2022).
20. Llanos, F., German, J. S., Gnanateja, G. N. & Chandrasekaran, B. The neural processing of pitch accents in continuous speech. *Neuropsychologia* **158**, (2021).

21. Tang, C., Hamilton, L. S. & Chang, E. F. Intonational speech prosody encoding in the human auditory cortex. *Science (1979)* **357**, 797–801 (2017).
22. Chevillet, M., Riesenhuber, M. & Rauschecker, J. P. Functional correlates of the anterolateral processing hierarchy in human auditory cortex. *Journal of Neuroscience* **31**, 9345–9352 (2011).
23. DeWitt, I. & Rauschecker, J. P. Phoneme and word recognition in the auditory ventral stream. *Proc Natl Acad Sci U S A* **109**, (2012).
24. Leaver, A. M. & Rauschecker, J. P. Cortical representation of natural complex sounds: Effects of acoustic features and auditory object category. *Journal of Neuroscience* **30**, 7604–7612 (2010).
25. Chevillet, M. A., Jiang, X., Rauschecker, J. P. & Riesenhuber, M. Automatic phoneme category selectivity in the dorsal auditory stream. *Journal of Neuroscience* **33**, 5208–5215 (2013).
26. Evans, S. & Davis, M. H. Hierarchical organization of auditory and motor representations in speech perception: Evidence from searchlight similarity analysis. *Cerebral Cortex* **25**, 4772–4788 (2015).
27. Myers, E. B., Blumstein, S. E., Walsh, E. & Eliassen, J. Inferior Frontal Regions Underlie the Perception of Phonetic Category Invariance. *Psychol Sci* **20**, 895–903 (2009).
28. Schirmer, A. & Kotz, S. A. Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences* vol. 10 24–30 Preprint at https://doi.org/10.1016/j.tics.2005.11.009 (2006).
29. Brodbeck, C. *et al.* Parallel processing in speech perception with local and global representations of linguistic context. *Elife* **11**, (2022).
30. Preisig, B. C., Riecke, L. & Hervais-Adelman, A. Speech sound categorization: The contribution of non-auditory and auditory cortical regions. *Neuroimage* **258**, (2022).
31. Hamilton, L. S., Oganian, Y., Hall, J. & Chang, E. F. Parallel and distributed encoding of speech across human auditory cortex. *Cell* **184**, 4626-4639.e13 (2021).
32. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci* **2**, (2008).
33. Su, L., Fonteneau, E., Marslen-Wilson, W. & Kriegeskorte, N. Spatiotemporal searchlight representational similarity analysis in EMEG source space. in *Proceedings - 2012 2nd International Workshop on Pattern Recognition in NeuroImaging, PRNI 2012* 97–100 (2012). doi:10.1109/PRNI.2012.26.
34. Yu, S., Giraldo, L. G. S., Jenssen, R. & Principe, J. C. Multivariate extension of matrix-based rényi's α-order entropy functional. *IEEE Trans Pattern Anal Mach Intell* **42**, 2960–2966 (2020).
35. Zhou, W., Yu, S. & Chen, B. Causality detection with matrix-based transfer entropy. *Inf Sci (N Y)* **613**, 357–375 (2022).
36. Kronrod, Y., Coppess, E. & Feldman, N. H. A unified account of categorical effects in phonetic perception. *Psychon Bull Rev* **23**, 1681–1712 (2016).
37. Diedrichsen, J. *et al.* Comparing representational geometries using whitened unbiased-distance-matrix similarity. *Neuron Behav Data Anal Theory* **5**, 1–31 (2020).
38. Boba, P. *et al.* Efficient computation and statistical assessment of transfer entropy. *Front Phys* **3**, (2015).
39. Diks, C. & Fang, H. Transfer entropy for nonparametric granger causality detection: An evaluation of different resampling methods. *Entropy* **19**, (2017).

40. Quian Quiroga, R., Kraskov, A., Kreuz, T. & Grassberger, P. Performance of different synchronization measures in real data: A case study on electroencephalographic signals. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **65**, 14 (2002).
41. Hugdahl, K., Thomsen, T., Ersland, L., Rimol, L. M. & Niemi, J. The effects of attention on speech perception: An fMRI study. *Brain Lang* **85**, 37–48 (2003).
42. Kreitewolf, J., Friederici, A. D. & von Kriegstein, K. Hemispheric lateralization of linguistic prosody recognition in comparison to speech and speaker recognition. *Neuroimage* **102**, 332–344 (2014).
43. Rutten, S., Santoro, R., Hervais-Adelman, A., Formisano, E. & Golestani, N. Cortical encoding of speech enhances task-relevant acoustic information. *Nat Hum Behav* **3**, 974–987 (2019).
44. Kilian-Hütten, N., Valente, G., Vroomen, J. & Formisano, E. Auditory cortex encodes the perceptual interpretation of ambiguous sound. *Journal of Neuroscience* **31**, 1715–1720 (2011).
45. Levy, D. F. & Wilson, S. M. Categorical Encoding of Vowels in Primary Auditory Cortex. *Cerebral Cortex* **30**, 618–627 (2020).
46. Moerel, M., De Martino, F., Uğurbil, K., Yacoub, E. & Formisano, E. Processing complexity increases in superficial layers of human primary auditory cortex. *Sci Rep* **9**, (2019).
47. Gwilliams, L., Linzen, T., Poeppel, D. & Marantz, A. In spoken word recognition, the future predicts the past. *Journal of Neuroscience* **38**, 7585–7599 (2018).
48. Norman-Haignere, S. V. *et al.* Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nat Hum Behav* **6**, 455–469 (2022).
49. Pierrehumbert, J. *Phonological and Phonetic Representation*. *Journal of Phonetics* vol. 18 (1990).
50. Inbar, M., Grossman, E. & Landau, A. N. Sequences of Intonation Units form a ~ 1 Hz rhythm. *Sci Rep* **10**, (2020).
51. Cusinato, R. *et al.* Intrinsic Neural Timescales in the Temporal Lobe Support an Auditory Processing Hierarchy. *Journal of Neuroscience* **43**, 3696–3707 (2023).
52. Norman-Haignere, S., Kanwisher, N. & McDermott, J. H. Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *Journal of Neuroscience* **33**, 19451–19469 (2013).
53. Kim, S. G. *et al.* MEG correlates of temporal regularity relevant to pitch perception in human auditory cortex. *Neuroimage* **249**, (2022).
54. Liebenthal, E. & Möttönen, R. An interactive model of auditory-motor speech perception. *Brain and Language* vol. 187 33–40 Preprint at https://doi.org/10.1016/j.bandl.2017.12.004 (2018).
55. Cope, T. E. *et al.* Temporal lobe perceptual predictions for speech are instantiated in motor cortex and reconciled by inferior frontal cortex. *Cell Rep* **42**, (2023).
56. Karimi-Rouzbahani, H., Woolgar, A., Henson, R. & Nili, H. Caveats and Nuances of Model-Based and Model-Free Representational Connectivity Analysis. *Front Neurosci* **16**, (2022).
57. David, O., Maess, B., Eckstein, K. & Friederici, A. D. Dynamic causal modeling of subcortical connectivity of language. *Journal of Neuroscience* **31**, 2712–2717 (2011).
58. Saur, D. Beyond ventral and dorsal streams: thalamo-cortical connections for subcortical language integration. *Brain* vol. 147 1927–1928 Preprint at https://doi.org/10.1093/brain/awae152 (2024).

59. Atiani, S., Elhilali, M., David, S. V., Fritz, J. B. & Shamma, S. A. Task Difficulty and Performance Induce Diverse Adaptive Patterns in Gain and Shape of Primary Auditory Cortical Receptive Fields. *Neuron* **61**, 467–480 (2009).
60. Cogan, G. B. *et al.* Sensory-motor transformations for speech occur bilaterally. *Nature* **507**, 94–98 (2014).
61. Oldfield, R. C. THE ASSESSMENT AND ANALYSIS OF HANDEDNESS: THE EDINBURGH INVENTORY. *Neuropsychologia* **9**, 97–113 (1971).
62. Kawahara, H. & Mutsui, H. Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)* (2003).
63. Nonyane, B. A. S. & Theobald, C. M. Design sequences for sensory studies: Achieving balance for carry-over and position effects. *British Journal of Mathematical and Statistical Psychology* **60**, 339–349 (2007).
64. de Cheveigné, A. ZapLine: A simple and effective method to remove power line artifacts. *Neuroimage* **207**, (2020).
65. Widmann, A., Schröger, E. & Maess, B. Digital filter design for electrophysiological data - a practical approach. *J Neurosci Methods* **250**, 34–46 (2015).
66. Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F. & Gramfort, A. Autoreject: Automated artifact rejection for MEG and EEG data. *Neuroimage* **159**, 417–429 (2017).
67. Pascual-Marqui, R. D. *et al.* Assessing interactions in the brain with exact low-resolution electromagnetic tomography. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **369**, 3768–3784 (2011).
68. Ledoit, O. & Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *J Multivar Anal* **88**, 365–411 (2004).
69. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
70. De La Pava Panche, I., Alvarez-Meza, A. M. & Orozco-Gutierrez, A. A Data-Driven Measure of Effective Connectivity Based on Renyi's α-Entropy. *Front Neurosci* **13**, (2019).

Acknowledgements

We thank Florian Scharf for extensive efforts in stimulus generation, and express our gratitude to Yvonne Wolff-Rosier for assistance in conducting the experiment and collecting the data. This work was supported by the Max Planck Society. The funder had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

Conceptualization: S-C.B., M.G., and D.S. Methodology: S-C.B., S-G.K., M.G., and D.S. Software: S-C.B., B.M., M.G., and D.S. Validation: S-C.B., S-G.K., B.M., M.G., and D.S. Formal analysis: S-C.B., M.G., and D.S. Data curation: S-C.B., B.M., M.G., and D.S. Writing–original draft: S-C.B. and D.S. Writing–review and editing: S-C.B., S-G.K., B.M., M.G., and D.S. Visualization: S-C.B., S-G.K., and D.S. Supervision: D.S. Project administration: D.S. Funding acquisition: D.S. All authors have read, commented on and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Daniela Sammler.

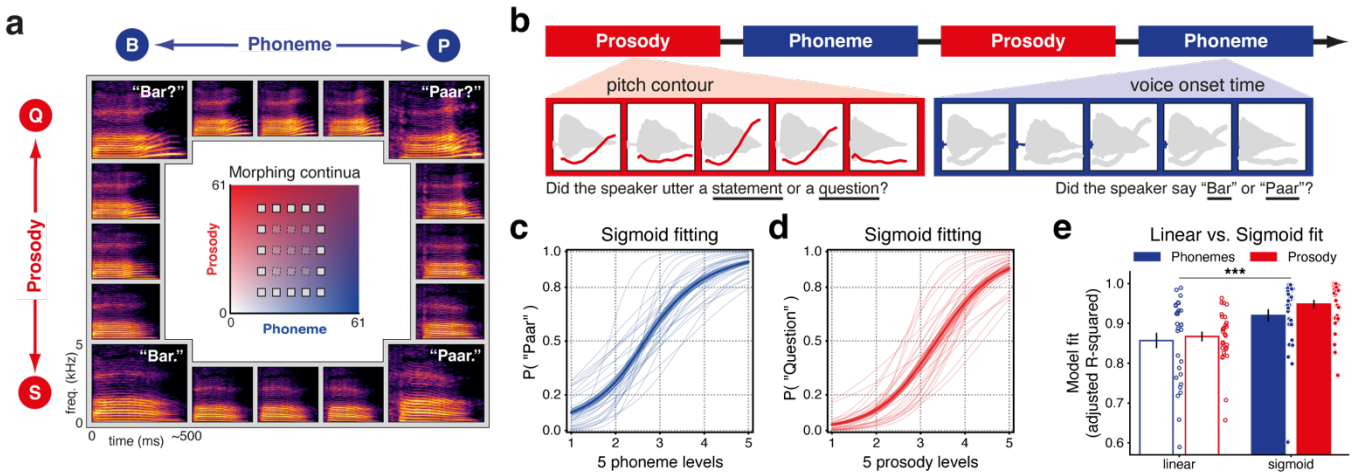

Fig. 1| Phoneme and prosody identification tasks and behavioural results. a, Auditory stimuli consisted of single words that varied in word-initial phonemes (horizontal axis) and prosody (vertical axis). The stimuli were generated via audio morphing applied to natural recordings of two German words, "Bar" (engl., "pub") and "Paar" (engl., "pair), uttered as both statement and question. Morphing resulted in a stimulus space of 61 x 61-step orthogonal continua (middle panel) with gradual transitions from /b/ to /p/ and from statement to question. 5 x 5 levels in this space, exemplified by the small squares, were selected for each participant (see also Fig. 2d). Stimulus examples are depicted as spectrograms in the surrounding panels (corresponding to the white squares in the stimulus space). **b**, Schematic of one out of six experimental runs. Each run contained two phoneme and two prosody blocks in alternating order, in which participants were asked to identify the word ("Bar"/"Paar") or the prosody ("statement"/"question"), respectively. The same set of stimuli was presented during both tasks. **c**,**d**, Sigmoid functions fitted to behavioural responses of "Paar" across five phoneme levels (**c**) and of "question" across five prosody levels (**d**). Shaded areas indicate the standard error of the mean (SEM). Thin lines represent individual participants. **e**, Comparisons of linear and sigmoid fits (adjusted R-squared) across the two identification tasks. Irrespective of the tasks, behavioural responses across five morph levels better followed a sigmoid function than a linear function (repeated measures analysis of variance (rmANOVA), main effect of function, $p < 0.001$; ***), confirming the categorial perception of both phonemes and prosody. Error bars denote SEM. Linear (empty) and sigmoid (filled) fits for each participant are depicted as circles.

993

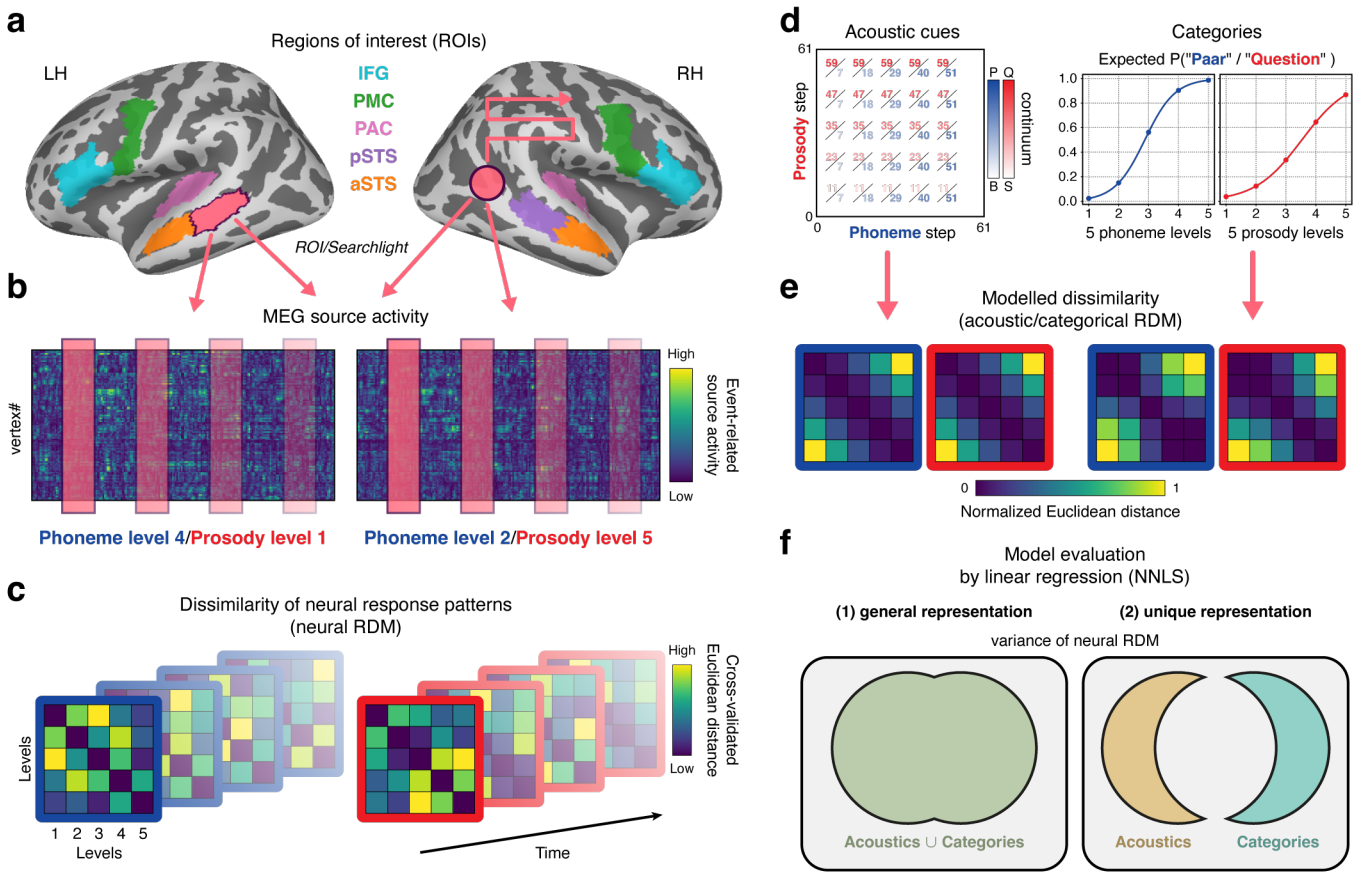

Fig. 2| Workflow of time-resolved representational similarity analysis (RSA). a, Multivariate magnetoencephalography (MEG) source activity evoked by the five phoneme or prosody levels was extracted from 10 predefined regions of interest (ROIs) across both hemispheres for each participant. The ROI-based approach was complemented by a whole-brain searchlight analysis (searchlight radius= 10 mm; pink circle). IFG, inferior frontal gyrus; PMC, premotor cortex; PAC, primary auditory cortex; p/aSTS, posterior/anterior superior temporal sulcus. **b**, For each ROI or searchlight location, cross-validated squared Euclidean distance (over six experimental runs) was calculated between the local neural response patterns to the five morph levels in 24-ms sliding time windows (step size: ROI= 4 ms, searchlight= 12 ms). **c**, This yielded a time series of 5 x 5 neural representational dissimilarity matrices (RDMs) for each participant, separately for phonemes (blue) and prosody (red). **d**, Neural representational characteristics were probed in terms of both stimulus acoustics and perceived categories. Stimulus acoustics (left) were coded as the original morph step of the stimuli on the 61-step phoneme or prosody continua. Perceived categories (right) were soft-coded as the expected proportion of “Paar” or “question” responses at each of the five morph levels, derived from the sigmoid functions fitted to individual behaviour. **e**, Stimulus acoustics and perceived categories across the five levels were compared separately using squared Euclidean distance, normalised between 0 and 1. This resulted in one acoustic and one categorical RDM, each with a size of 5 x 5, per participant, for phonemes and for prosody. **f**, To assess general representations of phonemes and prosody for each participant (left), each time step of the corresponding neural RDMs in a given region was predicted by a linear combination of acoustic and categorial RDMs using non-negative least squares (NNLS) fitting. Explained variance was further partitioned to examine unique acoustic and categorical representations (right).

994

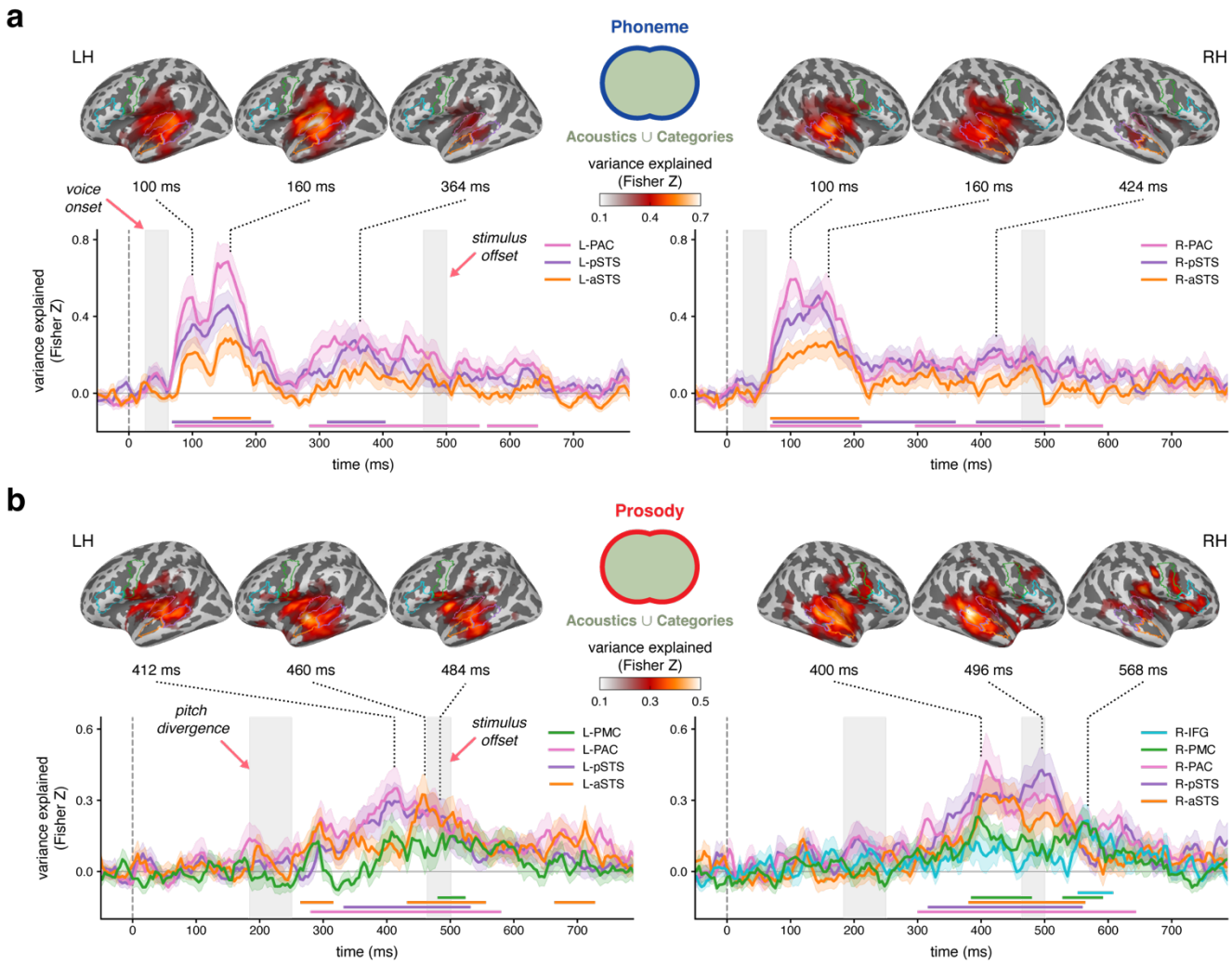

Fig. 3| Phonemic and prosodic representations in ventral and dorsal stream regions. a,b, Time-varying variance explained by both acoustic and categorical RDMs (transformed into Fisher-Z scale and corrected by the average baseline from -200 to 0 ms around stimulus onset) in the ROIs that exhibit significant representations of phonemes (**a**) or prosody (**b**) in the left (LH; left panel) and right (RH; right panel) hemispheres. Statistical significance was determined by cluster-based permutation tests (one-sided t_{28}-test against zero, cluster-forming threshold (CFT) at $p < 0.05$, 100,000 permutations, family-wise error rate (FWER)= 0.05 across time points per ROI, ROI-wise false discovery rate (FDR) $q < 0.05$). ROI colour coding is the same as in Fig. 2a (L, left; R, right). Shaded areas indicate SEM. Horizontal lines above the x-axis show significant time intervals (ROI-wise cluster and peak statistics are summarised in Table 1). At selected time points around the peaks of the ROIs, cortical variance maps (Fisher-Z scale, baseline-corrected) of significant representations from the whole-brain searchlight analysis are shown at the top (cluster-based permutation tests, one-sided t_{28}-test against zero, CFT at $p < 0.05$, 10,000 permutations, FWER= 0.05 across vertices and time points), with the ROIs outlined in the corresponding colours for comparison. **a**, Phonemic representations emerged after voice onset (left grey rectangle) across bilateral temporal areas and persisted beyond stimulus offset (right grey rectangle), particularly in posterior temporal areas. **b**, Prosodic representations occurred after pitch divergence (left grey rectangle) and extended from bilateral temporal areas to bilateral premotor and right inferior frontal regions around and after stimulus offset (right grey rectangle).

995

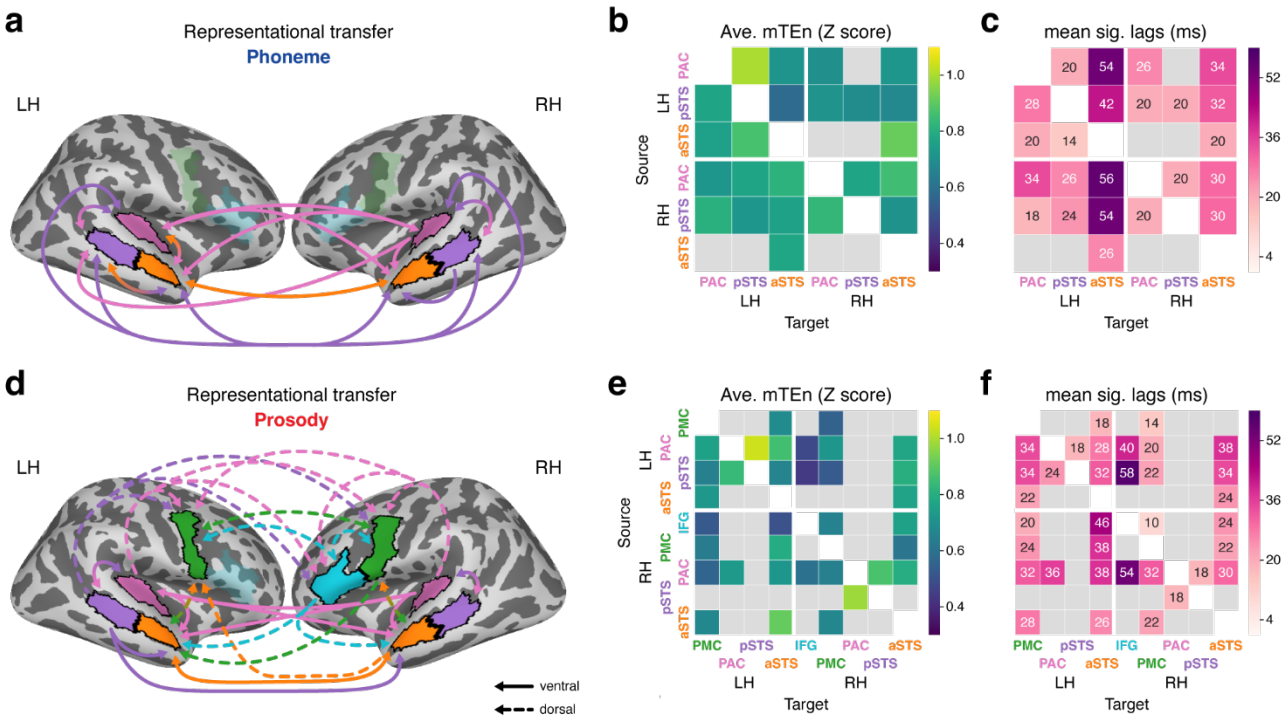

Fig. 4| Transfer of phonemic and prosodic representations along ventral and dorsal streams. **a,d**, Directed representational transfer of phonemes (**a**) and prosody (**d**) between the ROIs that displayed significant corresponding representations (Fig. 3) in the left (LH) and right (RH) hemispheres (cluster-based permutation tests, one-sided t_{28}-test on the normalised multivariate transfer entropy (mTEn) against zero, CFT at $p < 0.05$, 100,000 permutations, FWER= 0.05 across time lags per directed connection, directed-connection-wise FDR $q < 0.05$). Arrows indicate directed representational transfer from the source to the target ROIs, colour-coded according to the source ROIs. **b,c,e,f**, Averaged mTEn (**b**,**e**) and mean time lags (**c**,**f**) across significant time lags, corresponding to **a** (**b**,**c**) and **d** (**e**,**f**). Rows and columns show the source and target ROIs, respectively (mTEn and significant intervals across time lags are shown in Supplementary Fig. 7). **a**, Phonemic representations were transferred along the ventral stream, from posterior to anterior temporal ROIs, with feedback connections from the left aSTS to the ipsilateral PAC and pSTS. **d**, Prosodic representations propagated from posterior temporal ROIs to the aSTS along the ventral stream (solid) and to the bilateral PMC and right IFG along the dorsal stream (dashed). Note the relative dominance of the PAC as a source compared to the pSTS (almost no projection from the right pSTS), and the cross-stream interaction between aSTS and PMC, two major target areas in the ventral and dorsal stream.

996

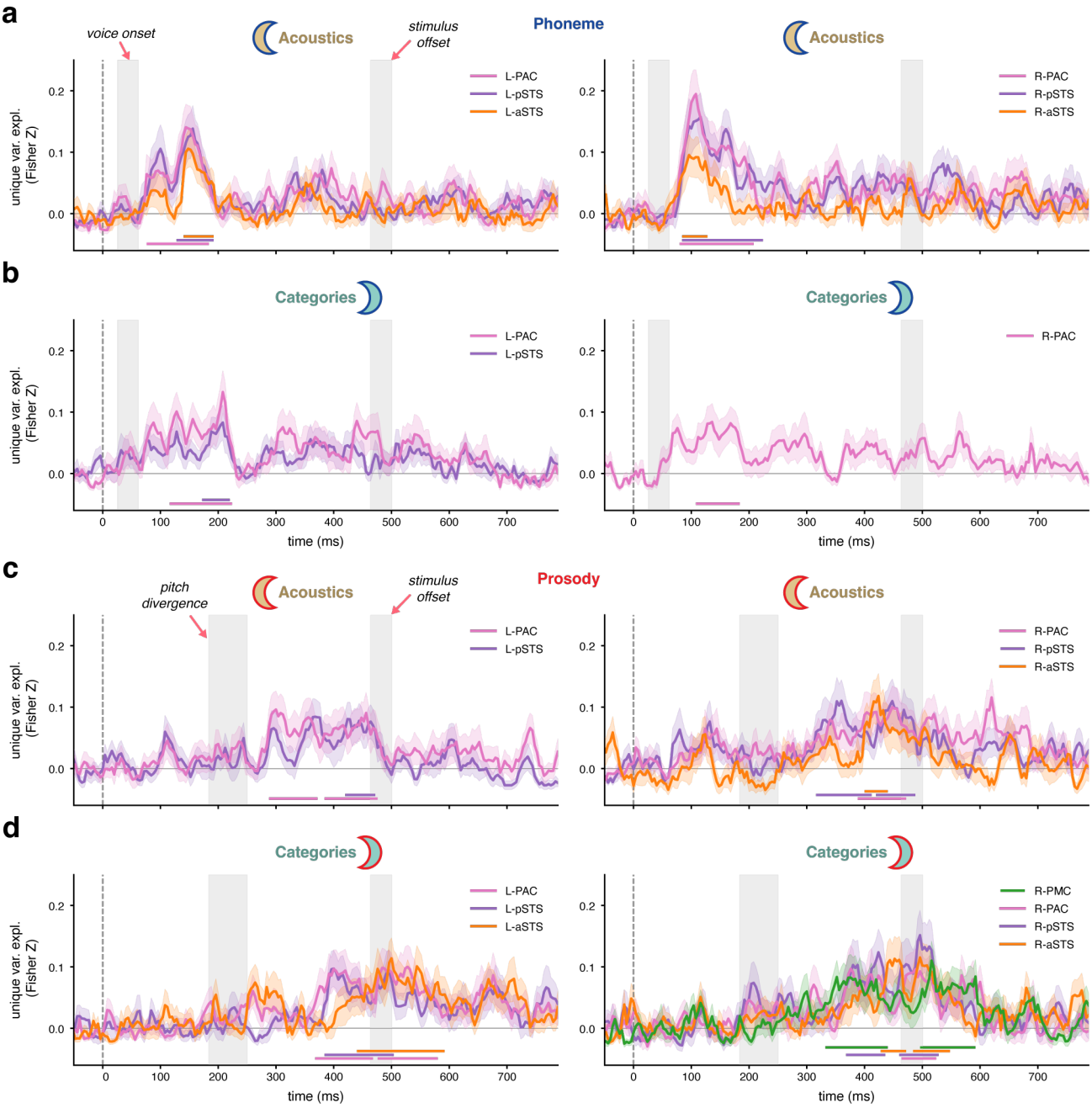

Fig. 5| Acoustic and categorical representations for phonemes and prosody in ventral and dorsal stream regions. a-d, Time-varying variance uniquely explained by the acoustic (**a**,**c**) or categorical RDM (**b**,**d**) (Fisher-Z scale, baseline-corrected) in the ROIs that exhibit significant unique representations of phonemes (**a**,**b**) or prosody (**c**,**d**) in the left (left panel) or right (right panel) hemisphere. Statistical significance was determined by cluster-based permutation tests (one-sided t_{28}-test against zero, CFT at $p < 0.05$, 100,000 permutations, FWER= 0.05 across time points per ROI, ROI-wise FDR $q < 0.05$). Shaded areas indicate SEM. Horizontal lines above the x-axis show significant time intervals (ROI-wise cluster and peak statistics are summarised in Table 2). **a**, Acoustic representations of phonemes emerged after voice onset (left grey rectangle) and extended across bilateral temporal ROIs. **b**, Phoneme categories were represented slightly later, but overlapped considerably in time with the acoustic representations (**a**). Notably, the categorical representations were distributed less broadly than acoustics, limited to the bilateral PAC and left pSTS. **c**, Prosody acoustics were represented after pitch divergence (left grey rectangle) in all bilateral temporal ROIs except for the left aSTS. **d**, Successive categorical representations of prosody overlapped with the acoustic representations (**c**) in time but lasted longer beyond stimulus offset (right grey rectangle). Note that categorical representations of prosody were spatially broader than those of phonemes (**b**), including the bilateral aSTS and right PMC.

997

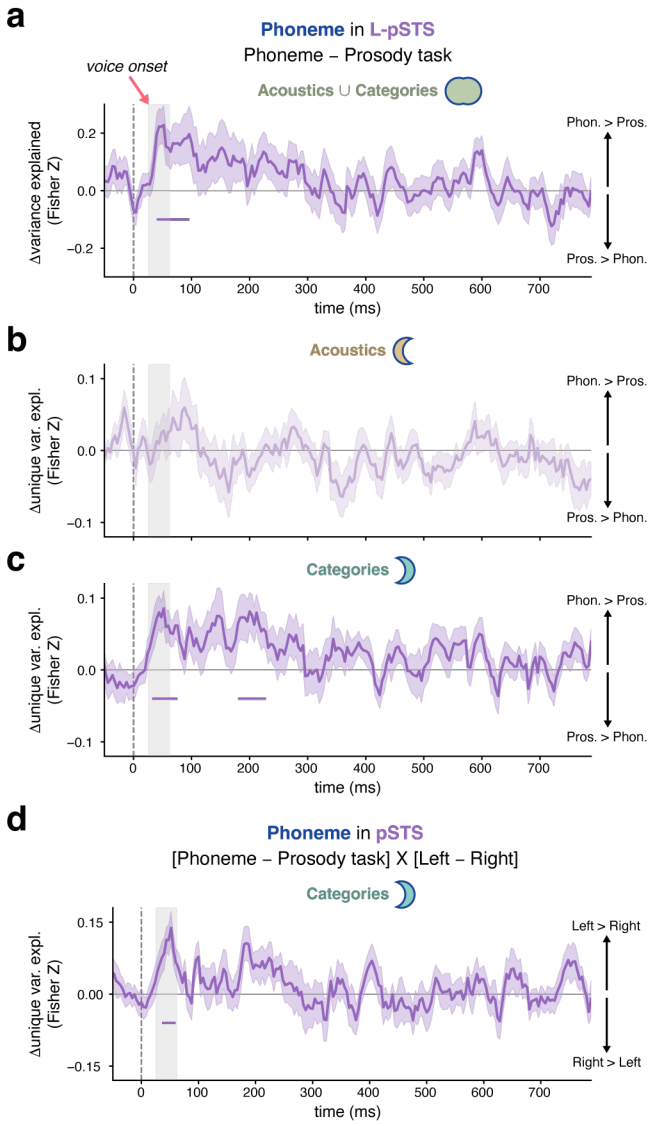

Fig. 6| Task modulation effect on phonemic representations in the left pSTS. a-c, Contrasts of time-varying variance (Fisher-Z, baseline-corrected) explained by both acoustic and categorical RDMs for phonemes (**a**) or uniquely explained by either RDM (**b**,**c**) between the phoneme and prosody identification tasks in the left pSTS. Shaded areas indicate SEM. Horizontal lines show significant time intervals (cluster-based permutation tests, one-sided t_{28}-test, CFT at $p < 0.05$, 100,000 permutations, FWER= 0.05 across time points). **a**, Phonemic representations were enhanced during the phoneme tasks (correction for multiple ROI testing at FDR $q < 0.05$) around voice onset (grey rectangle). **b,c**, The phoneme task intensified the categorical representations (**c**), but not acoustic representations (**b**, pale). **d**, Contrast of task-modulated unique variance for the categorical RDM (Fisher-Z) between the left and right pSTS over time (statistical inference and plotting details as above, but with two-sided t_{28}-test). Categorical representations of phonemes were selectively enhanced in the left pSTS.

998

Table 1| Cluster and peak statistics for phonemic and prosodic representations in the ROIs.

Linguistic element	ROI	Cluster statistics		Peak statistics		
		Extent (ms)	p-value	Latency (ms)	t_{28}	Effect size (SEM)
Phoneme	L-PAC	72-228	<0.001	156*	7.456	0.685 (0.090)
		284-552	<0.001	368	3.419	0.301 (0.087)
		564-644	0.0203	608	2.538	0.177 (0.069)
	L-pSTS	68-224	<0.001	156*	5.641	0.457 (0.080)
		312-404	0.0026	356	2.945	0.275 (0.091)
	L-aSTS	132-192	0.0018	160*	3.773	0.284 (0.074)
	R-PAC	68-212	<0.001	108*	6.667	0.595 (0.088)
		296-524	<0.001	468	3.643	0.242 (0.065)
		532-592	0.0214	568	3.048	0.177 (0.057)
	R-pSTS	72-360	<0.001	144*	4.824	0.508 (0.103)
		392-500	0.0028	420	3.203	0.235 (0.072)
	R-aSTS	68-208	<0.001	164*	3.592	0.269 (0.074)
Prosody	L-PAC	280-580	<0.001	416*	4.147	0.351 (0.083)
	L-pSTS	332-532	<0.001	412*	3.662	0.297 (0.080)
	L-aSTS	264-316	0.0273	296	3.192	0.206 (0.064)
		432-556	<0.001	456*	3.754	0.324 (0.085)
		664-728	0.0132	712	3.015	0.211 (0.069)
	L-PMC	480-524	0.0297	488*	2.669	0.159 (0.059)
	R-PAC	300-644	<0.001	408*	3.845	0.466 (0.119)
	R-pSTS	316-560	<0.001	496*	4.638	0.428 (0.091)
	R-aSTS	380-564	<0.001	408*	3.855	0.326 (0.083)
	R-PMC	384-480	0.001	396*	3.118	0.231 (0.073)
		528-592	0.004	560	3.186	0.207 (0.064)
	R-IFG	552-608	0.011	568*	3.409	0.215 (0.062)

This table lists statistics of significant clusters for phonemic and prosodic representations in the ROIs, along with their peak statistics. L, left; R, right; effect size, baseline-corrected Fisher-Z variance explained by both acoustic and categorical RDMs; *, global peak of effect size in a given ROI.

Table 2| Cluster and peak statistics for acoustic and categorical representations of phonemes or prosody in the ROIs.

Linguistic element	ROI	Abstraction level	Cluster statistics		Peak statistics		
			Extent (ms)	p-value	Latency (ms)	t_{28}	Effect size (SEM)
Phoneme	L-PAC	Acoustics	76-184	<0.001	144*	3.693	0.141 (0.037)
		Categories	116-224	<0.001	208*	3.794	0.133 (0.034)
	L-pSTS	Acoustics	128-192	<0.001	156*	3.667	0.139 (0.037)
		Categories	172-220	0.0083	208*	2.896	0.083 (0.028)
	L-aSTS	Acoustics	140-192	0.0020	148*	3.419	0.106 (0.030)
	R-PAC	Acoustics	80-208	<0.001	108*	4.872	0.195 (0.039)
		Categories	108-184	<0.001	136*	2.773	0.084 (0.030)
	R-pSTS	Acoustics	84-224	<0.001	116*	3.963	0.158 (0.039)
	L-aSTS	Acoustics	140-192	0.0067	148*	3.419	0.106 (0.030)
Prosody	L-PAC	Acoustics	288-372	<0.001	300*	3.166	0.096 (0.030)
			384-476	<0.001	456	2.683	0.090 (0.033)
		Categories	368-486	<0.001	400	3.088	0.097 (0.031)
			476-580	<0.001	528*	2.812	0.100 (0.035)
	L-pSTS	Acoustics	420-472	0.0116	424*	2.896	0.081 (0.027)
		Categories	384-504	<0.001	400*	3.257	0.097 (0.029)
	L-aSTS	Categories	440-592	<0.001	500*	3.370	0.114 (0.033)
	R-PAC	Acoustics	388-472	0.0024	448*	2.831	0.100 (0.035)
		Categories	464-524	0.0091	488*	2.841	0.096 (0.033)
	R-pSTS	Acoustics	316-412	0.0012	360*	3.174	0.110 (0.034)
			420-488	0.0029	448	3.015	0.110 (0.036)
		Categories	368-436	0.007	424	3.238	0.124 (0.038)
			460-528	0.0043	496*	3.673	0.152 (0.041)
	R-aSTS	Acoustics	400-440	0.0163	424*	3.205	0.118 (0.036)
		Categories	428-472	0.0132	452	3.058	0.113 (0.036)
			484-548	0.0035	496*	3.053	0.115 (0.037)
	R-PMC	Categories	332-440	<0.001	352	3.029	0.084 (0.027)
			496-592	<0.001	516*	3.410	0.110 (0.032)

This table lists statistics of significant clusters for acoustic and categorical representations of phonemes or prosody in the ROIs, along with their peak statistics. L, left; R, right; effect size, baseline-corrected Fisher-Z variance uniquely explained by each acoustic or categorical RDM; *, global peak of effect size in a given ROI.