Title: Speech Fine Structure Contains Critical Temporal Cues to Support Speech Segmentation

Running title: Fine Structure Contains Temporal Cues

Authors: Xiangbin Teng[a], Gregory B. Cogan[b], David Poeppel[a, c]

Affiliation:

[a]Department of Neuroscience, Max-Planck-Institute for Empirical Aesthetics, Frankfurt, Germany 60322

[b]Department of Neurosurgery, Duke University, Durham, North Carolina, USA 27710

[c]Department of Psychology, New York University, New York, NY, USA 10003

Submitting author: Xiangbin Teng
Corresponding author:
Xiangbin Teng
Department of Neuroscience
Max-Planck-Institute for Empirical Aesthetics
Grüneburgweg 14
60322, Frankfurt am Main
xiangbin.teng@gmail.com
Phone number: +49 69 8300479-329

Number of pages: 43
Number of figures: 5
Number of words for Abstract: 196

Conflict of Interest: The authors declare no competing financial interests.

Author Contributions:
Conceived and designed the experiments: XT, GBC & DP
Performed the experiments: XT
Analyzed the data: XT
Contributed reagents/materials/analysis tools: XT
Wrote the paper: XT, GBC, DP

1

1 **Abstract**

2 Segmenting the continuous speech stream into units for further perceptual and linguistic

3 analyses is fundamental to speech recognition. The speech amplitude envelope (SE) has

4 long been considered a fundamental temporal cue for segmenting speech. Does the

5 temporal fine structure (TFS), a significant part of speech signals often considered to

6 contain primarily spectral information, contribute to speech segmentation? Using

7 magnetoencephalography, we show that the TFS entrains cortical oscillatory responses

8 between 3-6 Hz and demonstrate, using mutual information analysis, that (i) the temporal

9 information in the TFS can be reconstructed from a measure of frame-to-frame spectral

10 change and correlates with the SE and (ii) that spectral resolution is key to the extraction

11 of such temporal information. Furthermore, we show behavioural evidence that, when the

12 SE is temporally distorted, the TFS provides cues for speech segmentation and aids

13 speech recognition significantly. Our findings show that it is insufficient to investigate

14 solely the SE to understand temporal speech segmentation, as the SE and the TFS derived

15 from a band-filtering method convey comparable, if not inseparable, temporal

16 information. We argue for a more synthetic view of speech segmentation – the auditory

17 system groups speech signals coherently in both temporal and spectral domains.

18

19 Keywords: speech segmentation, cortical entrainment, spectral correlation, spectro-

20 temporal

2

1 Introduction

2   Parsing the continuous speech stream into appropriate units for subsequent perceptual

3   and linguistic analyses serves as the basis for recognition (Poeppel 2003; Ghitza 2012;

4   Giraud and Poeppel 2012). Speech is typically argued to be comprised of slower

5   amplitude fluctuations, called the speech amplitude envelope (SE) and faster temporal

6   and frequency events, called the temporal fine structure (TFS). These distinct aspects of

7   speech signals have been investigated in different experiments to elucidate their relative

8   contributions (Smith et al. 2002; Xu and Pfingst 2003; Zeng et al. 2004; Lorenzi et al.

9   2006; Moore 2008; Shamma and Lorenzi 2013; Ewert et al. 2018). Evidence suggests

10 that the human auditory system relies on temporal information supplied by the SE for

11 grouping speech information (Ghitza and Greenberg 2009; Ghitza 2012; Ding and Simon

12 2014). Behavioural data demonstrate that four bands of noise modulated by the SE

13 suffice for speech recognition (Shannon et al. 1995). Neurophysiological evidence shows

14 that cortical entrainment to the SE shows a high correlation with speech intelligibility

15 (Luo and Poeppel 2007; Ding and Simon 2012; Giraud and Poeppel 2012; Peelle et al.

16 2013; Doelling et al. 2014). Speech, though can be decomposed into two parts (SE and

17 TFS) using a 'filterbank' method (Shannon et al. 1995; Smith et al. 2002), containing

18 coherent amplitude fluctuations and spectral cues. Does the auditory system primarily

19 exploit the SE for segmentation or use additional attributes of the speech signal when

20 segmenting the incoming speech stream?

21   The TFS is argued to convey distinct cues not present in the SE (Gilbert et al. 2007)

22 and to play an important role in mediating speech perception in challenging backgrounds

23 (Hopkins et al. 2008; Moore 2008; Swaminathan et al. 2016). It was demonstrated in

24 neurophysiological studies that the TFS contributes to robust cortical entrainment for

25 speech in noise (Ding and Simon 2014). Indeed, after eliminating amplitude fluctuations,

26 the processed speech signal can still entrain cortical oscillatory responses (Zoefel and

27 VanRullen 2015; 2016). These studies argue for an important role played by the TFS,

28 which complements the traditional view that the SE is the dominant cue for speech

29 segmentation. However, it remains unclear how the TFS contributes to robust cortical

30 entrainment to speech and whether the TFS and SE play different roles in speech

31 segmentation.

32 Various studies suggest that the envelope can be 'recovered' from the TFS through

33 cochlear processing (Ghitza 2001; Zeng et al. 2004; Shamma and Lorenzi 2013). From a

34 neurophysiological perspective, it is argued that cues in the spectral structure of speech

35 (without amplitude modulations) entrain cortical oscillations and provide temporal

36 information (Zoefel and VanRullen 2015). The TFS provides acoustic cues to help form

37 auditory objects for grouping and entrainment (Ding and Simon 2012; Ding et al. 2014).

38 But the cues, whether acoustic or high-level, are not clearly specified in previous studies,

39 and it is not well understood how these cues are extracted from the TFS.

40 Here we first take a traditional approach by separating the SE and the TFS using the

41 'filterbank' method and test whether the human auditory system can capitalize on the

42 TFS to temporally segment speech signals. If the TFS and the SE represent different

43 aspects of speech signals and the auditory system primarily extracts temporal information

44 from the SE for segmenting speech, as indicated by the previous studies (Luo and

45 Poeppel 2007; Ghitza and Greenberg 2009; Ding and Simon 2012; Ghitza 2012; Giraud

46 and Poeppel 2012; Peelle et al. 2013; Ding and Simon 2014; Doelling et al. 2014), we

4

47   would expect that the TFS alone cannot entrain cortical oscillatory responses (to the same

48   level) as the SE does. In contrast, if we do find that the TFS supplies sufficient temporal

49   information and robustly entrains cortical oscillatory responses, we aim to determine how

50   temporal information is extracted from the TFS, as well as the nature of this temporal

51   information. We then test behaviourally whether the TFS will provide temporal cues and

52   help increase intelligibility when the SE is disrupted temporally.

53   We show through neurophysiological results that the TFS, similar to the SE, robustly

54   entrains cortical oscillatory responses. The patterns of cortical entrainment differentiate

55   between the TFS of different sentences. To better understand what in the TFS elicits

56   robust cortical entrainment, we modified a method - cochlear scaled correlation (Stilp and

57   Kluender 2010) - to derive temporal information from the TFS. We compute the mutual

58   information between neurophysiological responses evoked by the TFS and 1) the original

59   SE, 2) the recovered envelope from the TFS, and 3) the derived temporal information

60   through spectral correlation in the TFS. We determine that the temporal information in

61   the TFS is highly relevant to the SE, and is contributed by the spectral correlation of the

62   TFS as well as by the recovered envelope. We further show that spectral resolution

63   strongly affects the ability to extract temporal information from the TFS. Next, we

64   temporally distort speech by using a widely cited but not often-used manipulation, locally

65   reversing speech segments (Saberi and Perrott 1999; Kiss et al. 2008; Stilp et al. 2010).

66   We demonstrate that the TFS helps restore critical temporal information and significantly

67   improves intelligibility of temporally distorted speech.

68   Our results demonstrate that the TFS provides significant temporal information for

69   segmenting speech. The TFS and the SE convey comparable, if not inseparable, temporal

70   information for speech segmentation. We conclude that speech segmentation and cortical

71   entrainment to speech are a result of tracking both the temporal and spectral structure of

72   speech.

Materials and Methods

73

Ethics statement.

74

75   The study was approved by the New York University Institutional Review Board (IRB#

76   10–7277) and conducted in conformity with the 45 Code of Federal Regulations (CFR)

77   part 46 and the principles of the Belmont Report. All the participants in Experiment 1 and

78   2 gave informed written consent.

79   **Participants.**

80   Fifteen Chinese native speakers from New York University took part in Experiment 1.

81   The data from three participants were excluded from MEG analysis because trial orders

82   were not recorded during neurophysiological recording. One participant was further

83   excluded from the mutual information analysis because part of data from this subject was

84   removed after preprocessing due to a noise issue, which resulted in the loss of the trial

85   order for the mutual information analysis. Therefore, in Experiment 1, the analysis

86   included the neurophysiological data from 12 participants for inter-trial phase coherence

87   and single-trial classification (6 females; age ranging from 21 to 32; right-handed) and

88   from 11 participants for the mutual information analysis (6 females; age ranging from 21

89   to 32; right-handed). Handedness was determined using the Edinburgh Handedness

90   Inventory (Oldfield 1971).

91   Twenty-One Chinese native speakers studying at New York University took part in

92   Experiment 2: ten in Experiment 2A (5 females; age ranging from 23 to 35; all self-

93   reported right-handed) and eleven in Experiment 2B (8 females; age ranging from 22 to

94   28; all self-reported right-handed). One participant was excluded from Experiment 2B

7

95   because of poor performance. All participants had normal hearing and no neurological

96   deficits according to their self-report.

97   **Stimuli.**

98   One hundred Chinese sentences from the Mandarin Hearing in Noise Test were used in

99   the present study (Fu et al. 2011; Zhu et al. 2014). All sentences are composed of 7

100   syllables each and have similar grammatical structure and are spoken by a female

101   speaker. The overall power of all sound files was normalized to the same value before

102   further acoustic processing. We first decomposed the sentences into envelopes and TFS

103   (see detailed method also in (Smith et al. 2002)). We filtered the speech signal into 16

104   bands using cochlear filter banks spanning from 80 to 8820 Hz and created analytic

105   signals for each frequency band through a Hilbert transformation (see Figure 1A for an

106   illustration). The envelope is computed as the magnitude of the analytic signal and TFS

107   was reconstructed by applying a cosine function to the phase series of the analytic signal

108   for each frequency band. We selectively chose 16 bands for decomposing the speech

109   signal because previous studies show that the envelope cannot be recovered from the

110   TFS, or the recovered envelope from TFS is no longer beneficial for speech recognition

111   when 16 bands are used (Gilbert and Lorenzi 2006; Sheft et al. 2008), which enables us

112   to further investigate what else in the TFS, besides the recovered envelope, can contribute

113   to speech segmentation.

114   In Experiment 1, we generated 28 TFS stimuli reconstructed from twenty-eight

115   randomly selected sentences by averaging TFS of each frequency band across 16 bands

116   for each sentence.

117   In Experiment 2, we processed the sentences and created four types of reversed

118   speech: directly reversed speech (R), envelope reversed speech (ER), fine structure

119   reversed speech (FSR), and envelope reversed noise-vocoded speech (ERNV). To

120   generate R sentences, we cut each sentence into short segments using a rectangular

121   window (e.g. 50 ms) and reversed each segment temporally. Then we concatenated the

122   reversed segments of each sentence in the original order to form a new sentence, whose

123   local segments were temporally reversed. The R sentences were generated in the same

124   way as in previous studies (Saberi and Perrott 1999; Stilp et al. 2010).

125   Figure 3A illustrates procedures to generate ER, FSR, and ERNV sentences. We cut

126   the envelope of each band into segments using a fixed window size and reversed each

127   segment temporally, and then concatenated them to form a reversed envelope for each

128   band. The new reversed envelope was then used to modulate the intact TFS of the

129   corresponding band to generate the ER sentences. We used the new reversed envelopes to

130   modulate narrow band noise of corresponding frequency bands to get the ERNV

131   sentences. We kept the envelopes intact while cutting TFS into segments and reversing

132   TFS segments to form reversed TFS. The intact envelopes and the reversed TFS were put

133   together to form FSR sentences.

134   In Experiment 2A, R, ER, and FSR sentences were used to test intelligibility on each

135   type of speech. Six window sizes were used to cut speech into segments for the R

136   sentences: 30, 50, 70, 80, 90, and 120 ms; six window sizes for the ER sentences: 30, 70,

137   90, 120, 150, and 200 ms; three window sizes for the FSR sentences: 30, 150, and 300

138   ms. In Experiment 2B, ER and ERNV sentences were used. Six window sizes were

139   chosen for the ERNV sentences: 30, 50, 70, 80, 90, and 120 ms; six window sizes for the

140   ER sentences: 30, 70, 90, 120, 150, and 200 ms.

141   All stimuli used in Experiment 1 and 2 were normalized to ~65 dB SPL. The stimuli

142   were delivered through plastic air tubes connected to foam ear pieces (E-A-R Tone Gold

143   3A Insert earphones, Aearo Technologies Auditory Systems) in Experiment 1 and

144   through Sennheisser 370 headphones in Experiment 2.

145   **Experiment 1: MEG procedure**

146   We selected TFS stimuli of 28 total sentences. 25 TFS stimuli of different sentences were

147   presented once each and 3 TFS stimuli of 3 different sentences were presented 25 times

148   each. All TFS stimuli were pseudo-randomly presented in one block during MEG

149   recordings. To keep subjects alert, after hearing each TFS stimulus participants were

150   prompted to make a judgment via a button box on whether TFS stimuli sounded like

151   speech or not. The behavioral responses were not analyzed because all participants

152   reported that TFS stimuli did not sound like speech. This could be because the TFS was

153   derived using 16 bands and the envelope cues cannot be recovered from the TFS (Smith

154   et al. 2002; Hopkins et al. 2010). The participants had no prior knowledge on the TFS

155   stimuli or on the sentences from which the TFS stimuli were derived. The inter-trial

156   interval (ISI) of 1.5 – 2 s began after the key press. The ISI was used as a baseline for

157   MEG analysis.

158   **MEG recording and preprocessing.**

159   MEG signals were measured with participants in a supine position, in a magnetically

160   shielded room using a 157-channel whole-head axial gradiometer system (KIT,

10

161   Kanazawa Institute of Technology, Japan). A sampling rate of 1000 Hz was used with an

162   online 1-200 Hz analog band-pass filter and a notch filter centered around 60 Hz. After

163   the main experiment, participants were presented with 1 kHz tone beeps of 50 ms

164   duration as a localizer to determine their M100 evoked responses, which is a canonical

165   auditory response (Roberts et al. 2000). 20 channels with the largest M100 response in

166   both hemispheres (10 channels in each hemisphere) were selected as auditory channels

167   for each participant individually. Further analysis was conducted only on the selected

168   channels.

169   MEG data analysis was conducted in MATLAB using the Fieldtrip toolbox

170   (Oostenveld et al. 2011) and the wavelet toolbox. Raw MEG data were noise-reduced

171   offline using the time-shifted PCA (de Cheveigné and Simon 2007) and sensor noise

172   suppression (de Cheveigné and Simon 2008) methods. A low-pass filter with cutoff

173   frequency of 100 Hz was applied offline on the de-noised data in the MEG160 software

174   (Yokogawa Electric Corporation and Eagle Technology Corporation, Tokyo, Japan) and

175   the preprocessed MEG data was then downsampled to 500 Hz. Trials were visually

176   inspected, and those with artifacts such as channel jumps and large fluctuations were

177   discarded. An independent component analysis was used to correct for eye blink-, eye

178   movement-, heartbeat-related and system-related artifacts. Each trial was divided into 6s

179   epoch, with 2s pre-stimulus period and 4s post-stimulus period. The variable baseline was

180   corrected for in each trial by subtracting out the mean of the whole trial before doing

181   further analyses.

182   To extract instantaneous phase information, single-trial data in each MEG channel

183   were transformed using a Morlet wavelet function embedded in the Fieldtrip toolbox,

184   with a frequency ranged from 1 to 60 Hz in steps of 1 Hz. To balance spectral and

185   temporal resolution of the time-frequency transformation, from 1 to 20 Hz, the window

186   length increased linearly from 1.5 cycles/frequency to 7 cycles/frequency, and was kept

187   constant at 7 cycles/frequency above 20 Hz. The analysis windows for each trial was

188   from 1s pre-stimulus period to 3s post-stimulus period with a temporal step of 10 ms.

189   Phase and power response (squared absolute value) were extracted from the wavelet

190   transform output at each time-frequency point for classification and mutual information

191   analysis.

192   **Inter-trial phase coherence (ITPC).**

193   In Experiment 1, the 'inter-trial phase coherence' (ITPC) was calculated on each time-

194   frequency point (details as in (Lachaux et al. 1999)). ITPC is a measure of consistency of

195   phase-locked neural activity entrained by stimuli across trials. ITPC of different

196   frequency bands reflects phase tracking of cortical oscillations to temporally modulated

197   stimuli. ITPC was computed across 20 trials for each of 3 TFS stimuli that were

198   presented repeatedly and across 20 of 25 different TFS stimuli that were presented once

199   each. As 1 – 5 trials were removed for each stimulus during preprocessing, to avoid bias

200   of estimating ITPC caused by unequal numbers of trials across different stimuli, we only

201   selected 20 trials for calculating ITPC.

202   **Single-trial classification.**

203   A single-trial classification analysis of 3 repeated TFS stimuli was carried out to

204   investigate whether cortical oscillations entrained by TFS can differentiate the TFS from

205   specific sentences. This classification analysis was described in detail in (Ng et al. 2013)

206   as well as in (Luo and Poeppel 2007; Cogan and Poeppel 2011; Herrmann et al. 2013).

207   For 20 trials of each repeated TFS stimulus, one trial was left out, and then a template

208   was created by averaging phase series using the circular mean across the remaining trials

209   for each of the TFS stimuli. There were three repeated TFS stimuli; so three templates of

210   TFS stimuli were created. The circular distances between each template and each left-out

211   trial from each TFS stimulus was computed. The circular distance was applied for phase

212   classification by taking the circular mean during the period of 300 ms to 1500 ms after

213   the stimulus onset, and on each frequency. A trial was given one template's label if the

214   distance between this trial and the template was the smallest among the three templates.

215   The classification analysis was conducted on each frequency within the frequency

216   range that showed robust ITPC for 3 TFS stimuli compared with 25 TFS stimuli that

217   were present once each.

218   A confusion matrix of classification scores was constructed for each trial of each

219   stimulus type on each auditory channel. Then, classification performance was measured

220   in a signal detection framework: correctly labelling the target stimulus was count as a

221   'hit' while labelling the other two stimuli as the target stimulus was counted as 'false

222   alarm'; d' was calculated based on hit rates and false alarm rates and averaged across all

223   auditory channels. Classification accuracy using the phase of each frequency was

224   indicated by the mean of d' over the three TFS stimuli, which was compared to the total

225   d' of the identification task which indicates participants' sensitivity in the behavioral

226   study (Macmillan and Creelman 2004)

227   **Spectral correlation and cochlear-scaled spectral correlation.**

228   The TFS preserves rich spectral information in speech, which may confer temporal

229   information through the change of spectral content along time. To quantify this spectral

230   change, inspired from cochlear-scaled entropy (Stilp et al. 2010), we created two indices

231   – spectral correlation (SC) and cochlear-scaled spectral correlation (CSC). We first used

232   a short-time Fourier transformation to generate spectral profiles of acoustic segments of

233   20 ms and then computed Pearson's correlation coefficients of the spectral contents

234   between adjacent temporal segments. This correlation, constrained by sampling rate and

235   the number of samples in the sound files, is indicated as SC. Similar to the cochlear-

236   scaled entropy, the spectral correlation computed after binning frequencies according to

237   cochlear bands was indicated as CSC. We computed CSC using 32, 18, 8 and 4 bands

238   separately to evaluate the effect of the number of cochlear bands on resolving temporal

239   information from TFS.

240   **Recovered envelope from TFS.**

241   Previous studies have shown that speech amplitude envelopes can be recovered from TFS

242   through cochlear processing (Ghitza 2001; Zeng et al. 2004). To measure how the

243   recovered envelope from the TSF provides temporal information, we filtered 28 TFS

244   stimuli in neurophysiological recording using Gammatone filterbanks of 32, 18, 8 and 4

245   bands, separately. The envelope of each cochlear band was extracted by using the Hilbert

246   transform on each band and taking the absolute value (Glasberg and Moore 1990;

247   Søndergaard and Majdak 2013). We then averaged the envelopes across all bands to get

248   the recovered envelope from TFS.

249   **Mutual information (MI) analysis.**

250   To investigate what temporal information in TFS entrains neurophysiological responses

251   and distinguishes different TFS sounds, we used the framework of MI to quantify shared

252   information between MEG signals and acoustic properties in the stimuli (Quian Quiroga

14

253   and Panzeri 2009; Panzeri et al. 2010). MI was calculated using the Information

254   Breakdown Toolbox in MATLAB (Pola et al. 2003; Magri et al. 2009). We computed the

255   MI between the phase series of each frequency (1 – 60 Hz) extracted from the time-

256   frequency analysis described above and the acoustic properties of the stimuli in

257   Experiment 1 - SC, CSC, original envelopes of sentences from which TFS stimuli were

258   extracted, and the recovered envelopes from TFS (Cogan and Poeppel 2011; Gross et al.

259   2013; Ng et al. 2013; Kayser et al. 2015). The MI value of each frequency was

260   calculated for each subject, for each condition, and for each auditory channel across

261   trials.

262   In the present study, the acoustic properties of each stimulus are simply the values at

263   each time point. For each frequency of the neurophysiological response, the phase

264   distribution was composed of six equally spaced bins: 0 to pi/3, pi/3 to pi * 2/3, pi * 2/3

265   to pi, pi to pi * 4/3, pi * 4/3 to pi * 5/3, and pi * 5/3 to pi * 2. By choosing 6 bins for

266   phase information, we ensured that there is enough temporal resolution to capture

267   acoustic dynamics (Cogan and Poeppel 2011). The acoustic properties were first

268   normalized within each sentence by dividing their maximum value and then grouped

269   within each condition using 8 bins equally spaced from the minimum value to the

270   maximum value. Eight bins were chosen because we wanted to have enough discrete

271   precision to capture changes in acoustic properties while making sure that each bin has

272   sufficient counts for MI analysis, since the greater number of bins would lead to zero

273   counts in certain bins.

274   The estimation of MI is subject to bias caused by finite sampling of the probability

275   distributions because limited data was supplied in the present study. Therefore, a

276   quadratic extrapolation embedded in the Information Breakdown Toolbox was applied to

277   correct bias. MI is computed on the data set of each condition. A quadratic function is

278   then fit to the data points and the actual MI is taken to be the zero-crossing value. This

279   new value reflects the estimated MI for an infinite number of trials and greatly reduces

280   the finite sampling bias (Montemurro et al. 2007; Panzeri et al. 2007).

281   A shuffling procedure of sentence labels was applied to determine the baseline for MI

282   analysis. 28 TFS sentences used in Experiment 1 were randomly assigned to trials in

283   each condition to create a new date set and the same MI analysis was implemented on

284   each frequency. This procedure was repeated 1000 times to get a distribution of MI

285   values, from which a 99% one-side threshold was derived. By doing this, we preserved

286   the structure of time series in data. Therefore, be comparing the baseline with the results,

287   we ensured that MI results were not due to noise and specific processing procedures.

288   **Experiment 2: Behavioral measurement**

289   All participants were sitting in a soundbooth while doing the tasks. In Experiment 2A, 10

290   R, 10 ER, and 10 FSR sentences for each window size were presented. We presented

291   each type of sentence in separate blocks, such that three blocks (R block, ER block, and

292   FSR block) were presented in Experiment 2A and there were 60 different R sentences

293   (six window sizes $\times$ 10 sentences), 60 ER sentences (six window sizes $\times$ 10 sentences),

294   and 30 FSR sentences (three window sizes $\times$ 10 sentences) total. Since we only have 100

295   different sentences in the materials, 10 sentences were shared between the R block and

296   ER block. These shared ten sentences were in the largest window size condition (120 and

297   200 ms). Because intelligibility is very low for R and ER blocks under the largest

16

298 window size, we can avoid the confound of hearing an intelligible sentence a second time

299 which could prime listeners to better understand the second instance of the sentence.

300   Thirty different FSR sentences were selected from the sentences of the largest window

301 size used in the R block and ER block, 15 from each block. The FSR block was always

302 presented at the end, because, in our preliminary testing, we found that the FSR sentences

303 were highly intelligible. We first presented the R block and then the ER block, instead of

304 counter-balancing the order of two blocks. By doing this, we tried to avoid the confound

305 that ER sentences contain more speech cues - as the intact TFS was preserved in ER

306 sentences, participants could in theory adapt to the cues in ER sentences and improve

307 performance in the R block.

308   In Experiment 2B, 10 ER and 10 ERNV sentences for each window size were

309 presented and there were 60 ER sentences and 60 ERNV sentences total in two separate

310 blocks. 10 sentences were shared between the ER block and the ERNV block. We set

311 these ten shared sentences in the condition of the largest window size. The order of ER

312 block and ERNV block was counter-balanced between participants.

313   Using MATLAB, the participants were presented with one sentence on each trial and

314 were required to type in an Excel sheet what they heard after each sentence was presented

315 (10 second limit). After the participants finished typing, they pushed a key on a message

316 dialogue box to start next sentence. The input method for Chinese characters was

317 Microsoft Pinyin IME 2003 without autocomplete. We treated each character of a

318 sentence as one response and used the total number of the correct characters typed by the

319 participants over 10 sentences divided by the total number of characters (70) as the

320 intelligibility score for each window size.

321   **Psychometric function fitting.**

322   For the R and ER blocks in Experiment 2A and the ER block and ERNV block in

323   Experiment 2B, we fitted a psychometric curve to each participant's intelligibility score

324   for each block using a Weibull function in the Palademes toolbox 1.5.2 (Prins and

325   Kingdom 2009). The 50 percent intelligibility threshold and the slope of the

326   psychometric curves were derived for each participant and later used for further analysis.

327   Because intelligibility in FSR stayed at a ceiling level across all window sizes, we did not

328   fit a psychometric curve to the data of FSR block. For the purpose of illustration, we

329   averaged the intelligibility scores across subjects and fit a psychometric function to the

330   averaged scores. These psychometric functions using group averaged scores were not

331   used in any analysis.

332   **Results**

333   **TFS robustly entrains low frequency cortical oscillatory responses**

334   We used magnetoencephalography (MEG) recordings to measure neurophysiological

335   responses to the TFS in Experiment 1. As cortical oscillatory responses entrained by

336   auditory signals reflect how the auditory system both extracts temporal information and

337   parses the incoming acoustic stream, studying cortical entrainment evoked by the TFS

338   can reveal what information in TFS is extracted (Giraud and Poeppel 2012; Henry and

339   Obleser 2012; Kayser et al. 2012; Luo and Poeppel 2012; Doelling et al. 2014; Henry et

340   al. 2014; Kayser et al. 2015; Di Liberto et al. 2016; Teng et al. 2017).

341   We presented the 28 TFS stimuli in one block to 12 Mandarin Chinese native speakers

342   while recording their neurophysiological responses. Three of the 28 TFS stimuli were

343   repeatedly presented (25 times), and the remaining 25 TFS stimuli were presented once.

344   We selected 20 MEG channels using a tone localizer (see Methods) and computed inter-

345   trial phase coherence (ITPC) across 20 trials for each of three repeated TFS stimuli and

346   across TFS stimuli from 20 different sentences. We conducted a one-way repeated

347   measures ANOVA on ITPC from 1 to 60 Hz, with four sentence types (three repeated

348   sentences and one group of different sentences) as the main factor, and found a

349   significant main effect from $1 - 8$ Hz ($p < 0.05$, False Discovery Rate (FDR) corrected

350   (Benjamini and Hochberg 1995)). We then grouped ITPC values from 1 to 8 Hz and

351   conducted pairwise comparisons to examine differences between the four sentence types.

352   The ITPC results show that cortical responses between 1 to 8 Hz are robustly entrained

353   by the three repeated TFS stimuli (compared to the ITPC from the 20 distinct sentences;

354   Fig 1B) ($p < 0.05$, paired t test, Bonferroni corrected). The topographies of ITPC show

355   response patterns with an auditory origin, which is consistent with the hypothesis that the

356   TFS evoked robust cortical entrainment (Fig. 1D).
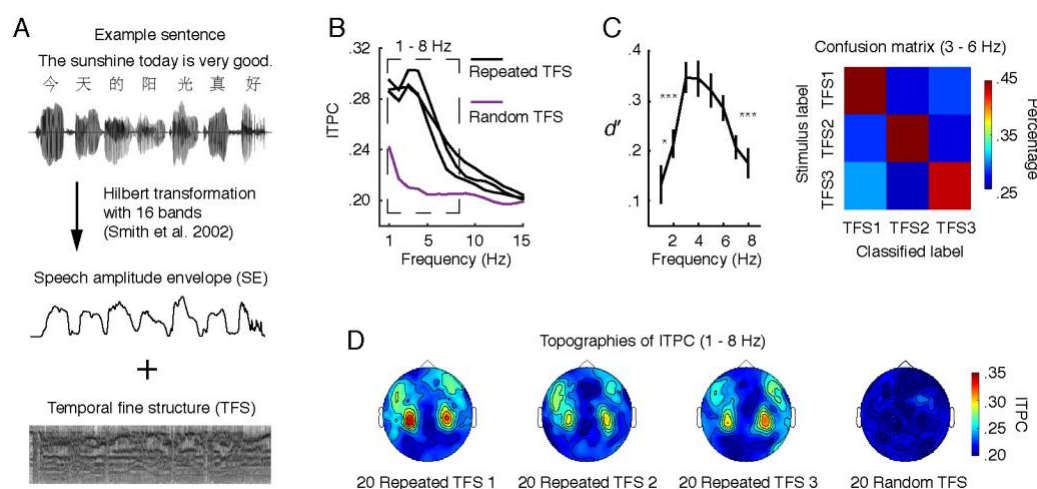
357

358   **Figure 1.**

359   Cortical entrainment to TFS. (A) Speech signals were decomposed using 16 bands into the speech
360   amplitude envelope (SE) and the temporal fine structure (TFS). (B) The repeated TFS stimuli evoke robust
361   cortical entrainment from 1 to 8 Hz. X-axis: frequency of neural response. Y-axis: inter-trial phase
362   coherence (ITPC). Black lines show entrainment to the three repeated TFS stimuli, separately. The violet
363   line shows cortical entrainment to 20 random TFS sentences, used as a baseline for the ITPC. The dashed
364   box indicates frequencies where ITPC of the repeated TFS stimuli is significantly larger than ITPC of the
365   20 random TFS stimuli ($p < 0.05$, paired t test, Bonferroni corrected). (C) Left panel, classification analysis
366   conducted on the phase series within each frequency band. The results show that the neural signals in
367   frequencies between 3 and 6 Hz are the most informative to separate out different repeated TFS stimuli.
368   Right panel: the group-averaged confusion matrix between 3 and 6 Hz. Each repeated TFS stimulus can be
369   robustly classified. (D) Topographies for ITPC between 1 and 8 Hz confirm that the cortical entrainment to
370   repeated TFS stimuli is of auditory origin. The error bars represent +/- SEM over subjects. Asterisks show
371   significant level (***, $p < 0.001$; *, $p < 0.05$).

372   To determine whether cortical oscillatory responses track distinct temporal structures

373   reflected in the TFS, we employed a single-trial classifier to classify trials of the three

374   repeated TFS stimuli between 1 and 8 Hz, using the MEG phase series (Fig. 1*C*) (Cogan

375   and Poeppel 2011; Herrmann et al. 2013; Ng et al. 2013). We used a signal detection

376   paradigm and converted the results of the classifier into the total *d'*, which indicates the

377   classification accuracy across all of the three repeated TFS stimuli (See Methods). We

378   found that the total *d'* was significantly above the zero value (*d'* for 33 percent correct

379   classification) from 1 to 8 Hz ($p < 0.05$, one sample t test, Bonferroni corrected) (Fig. 1C,

380   left panel). We then compared the total d' between different frequencies and found that

381   the phase patterns between 3 to 6 Hz were the most informative to the classification

382   analysis ($p < 0.05$, paired t test, Bonferroni corrected).

383   These data on cortical entrainment to the TFS bear a strong resemblance to previous

384   studies in which the SE was found to entrain neural responses below 10 Hz

385   (Luo and Poeppel 2007; Kerlin et al. 2010; Cogan and Poeppel 2011; Ding and Simon

386   2012; Peelle et al. 2013; Zion Golumbic et al. 2013). As prominent temporal information

387   in speech between 3 and 6 Hz is carried by the SE (Ding et al. 2017), the results suggest

388   that the temporal information relevant to the SE is perhaps read-out from the TFS by the

389   auditory system.

390   **Temporal information extracted from TFS correlates with original envelope and**

391   **can be reconstructed from spectral correlation.**

392   To investigate the nature of the temporal information extracted from the TFS, we

393   measured how the SE could explain phase patterns evoked by the corresponding TFS

394   using a mutual information (MI) framework (Quian Quiroga and Panzeri 2009; Panzeri et

395   al. 2010). Analyses were conducted on the data from 11 of 12 Mandarin Chinese native

396   speakers used in the previous analysis. One subject was excluded because of noise issues.

397   We computed MI between the SE of 20 different sentences and the phase series evoked

398   by the TFS derived from these 20 sentences. To set a baseline for MI values, we

399   generated a null distribution of MI values by shuffling the labels between the sentences

400   and computed MI values from 1000 shuffled datasets. We found that phase patterns of

401   the entrained oscillations by TFS share a significant amount of information with the SE

402   from 1 to 8 Hz ($p < 0.01$) (Fig. 2*A*, left panel), which aligns with the findings above on

403   cortical entrainment (Fig. 1*A*). This result demonstrates that the temporal information

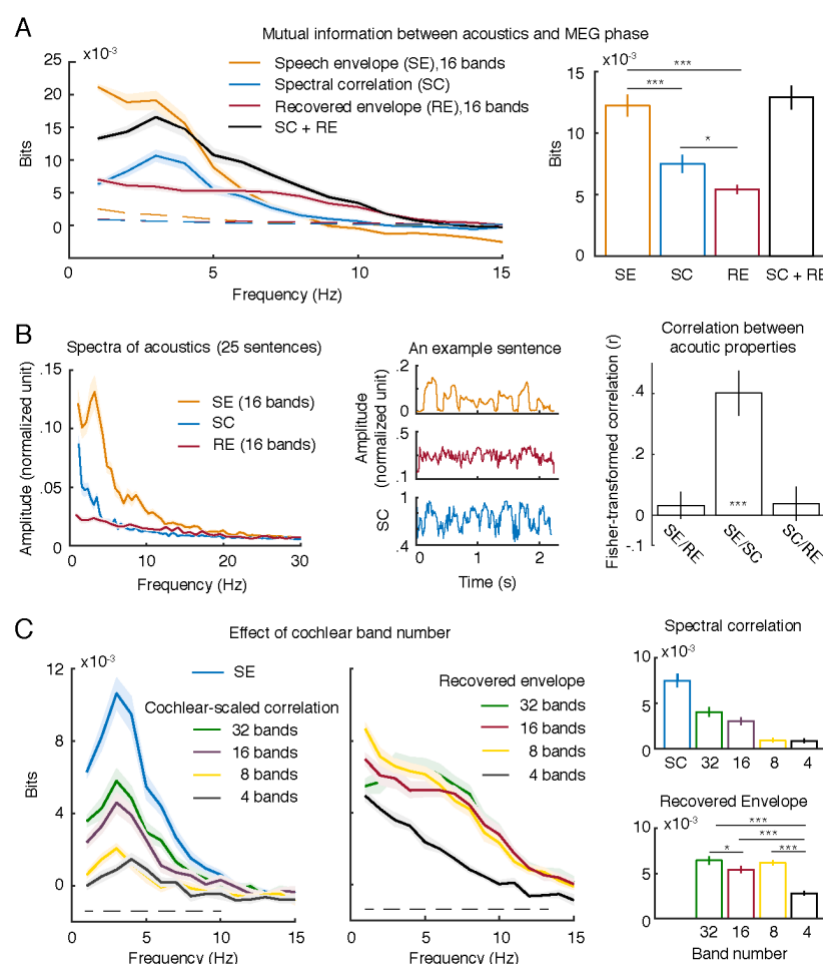404   extracted from the TFS strongly correlates with the envelope information.

405

Figure 2.

407   Results of mutual information (MI) analysis. (*A*) MI between MEG phase series evoked by the TFS and the
408   speech amplitude envelope (SE), spectral correlation (SC), and recovered envelope (RE). The speech
409   amplitude envelope and the recovered envelope were computed using 16 bands. The left panel shows MI
410   results from 1 to 15 Hz. The dashed lines are significance thresholds with a one-sided alpha level of 0.01
411   for each acoustic property, which were derived from a permutation method (see Methods). The right panel
412   shows average MI between 3 and 6 Hz where the phase series is most informative to discriminate the TFS
413   from different sentences. The results show clearly that the TFS contains temporal information which
414   correlates with the SE. The temporal information of the TFS can be derived from the spectral correlation
415   and the recovered envelope. (*B*) Illustration of acoustics. The left panel shows the averaged spectra of the
416   three acoustic properties used in the MI computation. The results were computed using the same 25
417   sentences for all the acoustic properties. The middle panel shows an example time series of each acoustic
418   property for the same sentence. The color code is as in (*A*). The right panel shows Fisher-transformed
419   Pearson correlation coefficients between three acoustic properties. It can be seen that the SE and the SC are

420   highly correlated ($p < 0.001$). (*C*) Effects of spectral resolution on MI results. The left panel shows MI

421   results of the spectral correlation and cochlear-scaled correlation. The dashed line indicates frequencies

422   where the main effect of band number or spectral resolution is significant ($p < 0.05$, FDR corrected). The

423   middle panel shows MI results of the recovered envelope computed using different number of frequency

424   bands. The dashed line indicates frequencies where the main effect of band number is significant ($p < 0.05$,

425   FDR corrected). The right panel shows averaged MI values between 3 and 6 Hz where the phase series is

426   most informative. The results indicate that the spectral resolution or band number significantly affects the

427   extraction of temporal information through spectral correlation - but not much through recovered envelope.

428   This suggests that for cochlear implant users or people with hearing loss, the inability to use the TFS for

429   speech perception could be due to degraded spectral resolution. The shaded area indicates +/- SEM over

430   subjects. Asterisks show significance levels (***, $p < 0.001$; *, $p < 0.05$).

431   Next we tried to determine how temporal information is extracted from the TFS. It has

432 been suggested that the SE can be recovered from TFS through cochlear processing

433 (Ghitza 2001; Zeng et al. 2004; Shamma and Lorenzi 2013). We filtered TFS stimuli

434 using a gammatone filter bank of 16 bands to simulate cochlear processing (Patterson

435 1976; Patterson et al. 1987) and obtained the recovered envelopes from the gammatone

436 filter outputs (See Methods for details). The averaged spectrum of recovered envelopes

437 over the 25 sentences and an example of a time series can be seen in Figure 2*B* (red line).

438 We then computed MI between the recovered envelope and the phase series evoked by

439 the TFS. Although we found significant MI values across a wide range of frequencies (1

440 – 15 Hz) ($p < 0.01$, FDR corrected), the amount of MI between the recovered envelope

441 and the phase series was much smaller than the MI between the SE and the phase series

442 (Fig. 2*A*, red line). This result suggests that the recovered envelope is not solely what the

443 auditory system uses to extract temporal information from TFS (Sheft et al. 2008). Other

444 processes may be in play. Nonetheless, the result confirms the previous finding from a

445 neurophysiological perspective, namely that when TFS is extracted using more than 8

446 bands, as in our present study, the recovered envelopes are no longer beneficial for

447 speech recognition (Gilbert and Lorenzi 2006).

448   As TFS contains rich spectral information (Moore 2008; Shamma and Lorenzi 2013)

449   and many studies have shown that cortical responses can be entrained by frequency

450   modulations (Henry and Obleser 2012; Herrmann et al. 2013; Henry et al. 2014; Teng,

451   Tian, Doelling, et al. 2017; Teng, Tian, Rowland, et al. 2017), the cortical entrainment

452   elicited by the TFS may be caused by the spectral structure of the TFS. To reconstruct

453   temporal information from the spectral structure of the TFS, we modified the cochlear-

454   scaled entropy paradigm (Stilp and Kluender 2010) and computed spectral correlation

455   using a short moving temporal window (20 ms) (see Methods for details). The average

456   spectrum of spectral correlation of 25 sentences and an example of time series for one

457   sentence can be seen in Figure 2B (blue line), which showed similar dynamics to the SE

458   (Fig. 2B, orange line). The SE and the SC are significantly correlated over the 25

459   sentences used (one-sample t-test against zeros, $t(24) = 12.16$, $p < 0.001$, Bonferroni

460   corrected), but no significant correlation was found between the RE and the SE ($t(24) =$

461   1.58, $p = 0.384$, Bonferroni corrected) as well as between the RE and the SC ($t(24) =$

462   1.45, $p = 0.480$, Bonferroni corrected) (Fig. 2B, right panel). We computed MI between

463   the spectral correlation and the phase series evoked by TFS. MI values were found to be

464   significant from 1 to 10 Hz ($p < 0.01$, FDR corrected) and were larger than the MI

465   between the recovered envelope and the phase series from 2 to 4 Hz, which coincides

466   with the peak of spectra of the SE (~ 3 Hz) (Fig. 2A, orange line). The results

467   demonstrate that the auditory system extracts temporal information relevant to the SE

468   from TFS using a procedure modelled by spectral correlation.

469   We combined MI computed from both spectral correlation and recovered envelope and

470   summarized the MI values over 3 to 6 Hz, the most informative frequency range found in

471   the classification analysis (Fig. 2*A*, right panel). We found that there was no significant

472   difference between the MI values computed using SE and the MI values computed using

473   both the recovered envelope and the spectral correlation ($t(10) = 1.18$, $p = 1$, $d = 0.37$,

474   Bonferroni corrected).

475   In summary, the results suggest that temporal information can be extracted from TFS

476   which correlates with the temporal information carried by the SE. A process that

477   computes the spectral correlation between adjacent temporal frames could be used by the

478   auditory system to extract the envelope information from TFS. The recovered envelope

479   from cochlear processes also provides temporal information but does not seem to play a

480   prominent role (Sheft et al. 2008).

481   **Extraction of temporal information from TFS depends on the number of frequency**

482   **bands (spectral resolution).**

483   As cochlear implants often provide poor spectral resolution (Oxenham and Kreft 2014)

484   and persons with hearing loss manifest degraded spectral sensitivity (Hopkins and Moore

485   2011), we tested whether the number of frequency bands (spectral resolution) affects

486   extracting temporal information from TFS, which may explain the inability of the

487   cochlear implant users to effectively use TFS.

488   We computed the cochlear-scaled correlation (CSC) by binning frequencies into 32,

489   16, 8 and 4 cochlear bands separately (see Methods) and then calculated MI between

490   MEG phase series and the CSC of the different band numbers (Fig. 2*C*, left panel). We

491   conducted a one-way repeated measures ANOVA from 1 to 60 Hz, with spectral

492   resolution as the main factor (five levels: SC, CSC of 32, 16, 8 and 4 bands), and found a

493   significant main effect of the spectral resolution from 1 to 10 Hz ($p < 0.05$, FDR

494   corrected). We then averaged MI from 3 to 6 Hz and found a downward linear trend with

495   decreased spectral resolution ($F(1,10) = 67.76$, $p < 0.001$, $\eta_P^2 = .871$) (Fig. 2C, right

496   panel). The results provide compelling evidence that spectral resolution significantly

497   affects extraction of temporal information from the TFS.

498   We next examined the effect of the band number on the recovered envelope. We

499   computed recovered envelopes using 32, 16, 8, and 4 cochlear bands separately and

500   calculated MI between MEG phase series and the recovered envelopes of different band

501   numbers (Fig. 2C, middle panel). We conducted a one-way repeated measures ANOVA

502   from 1 to 60 Hz, with the number of cochlear bands as the main factor (four levels: 32,

503   16, 8 and 4 bands), and found a significant main effect of the band number from 1 to 13

504   Hz ($p < 0.05$, FDR corrected). We then averaged MI over 3 to 6 Hz and found a

505   downward linear trend with decreased band number ($F(1,10) = 61.93$, $p < 0.001$, η_P^2

506   = .861). However, in a post-hoc test, we found that MI computed using 32 bands was

507   larger than using 16 bands ($t(10) = 4.18$, $p = .011$, $d = 1.32$) but not more than 8 bands

508   ($t(10) = 0.80$, $p = 1$, $d = 0.25$). The MI computed using 4 bands is lower than all the other

509   bands (32 bands: $t(10) = 9.89$, $p < .001$, $d = 3.13$; 16 bands: $t(10) = 6.86$, $p < .001$, $d =$

510   2.17; 8 bands: $t(10) = 10.89$, $p < .001$, $d = 3.44$). Bonferroni correction was applied.

511   The results show that the number of bands does not affect recovering the envelope

512   from the TFS as the frequency bands decrease from 32 to 8 bands. In contrast, we found a

513   significant effect of the band number for spectral correlation. This is consistent with

514   previous findings that spectral resolution modulates the efficiency of extracting temporal

515   information from the TFS (Léger et al. 2015; Oxenham 2018).

516 **TFS significantly compensates for SE temporal information compromised in**

517 **temporally disrupted speech.**

518 We tested whether the TFS compensates for lost temporal information when the SE is

519 disrupted in Experiment 2. We used a 'reversed speech' paradigm in which the temporal

520 structure of speech signals is compromised by temporally reversing local segments and

521 then tested speech intelligibility (Saberi and Perrott 1999; Stilp et al. 2010). The

522 rationale is that the reversing procedure disrupts the modulation phase of the SE and, as

523 the reversed temporal window becomes larger, the modulation phase gets distorted more

524 severely and provides incorrect cues on the temporal structure of speech. This reversing

525 procedure, therefore, renders the temporal cues carried by the reversed SE unavailable for

526 segmenting speech signals. Then we can test whether the auditory system extracts certain

527 cues provided by the TFS to segment speech signals. Figure 3*A* show a schematic

528 illustration of stimulus generation. By comparing speech intelligibility for the reversed

529 speech when TFS is intact with the conditions when the SE and TFS are both disrupted,

530 we determine whether the TFS, similar to the SE, provides critical temporal information
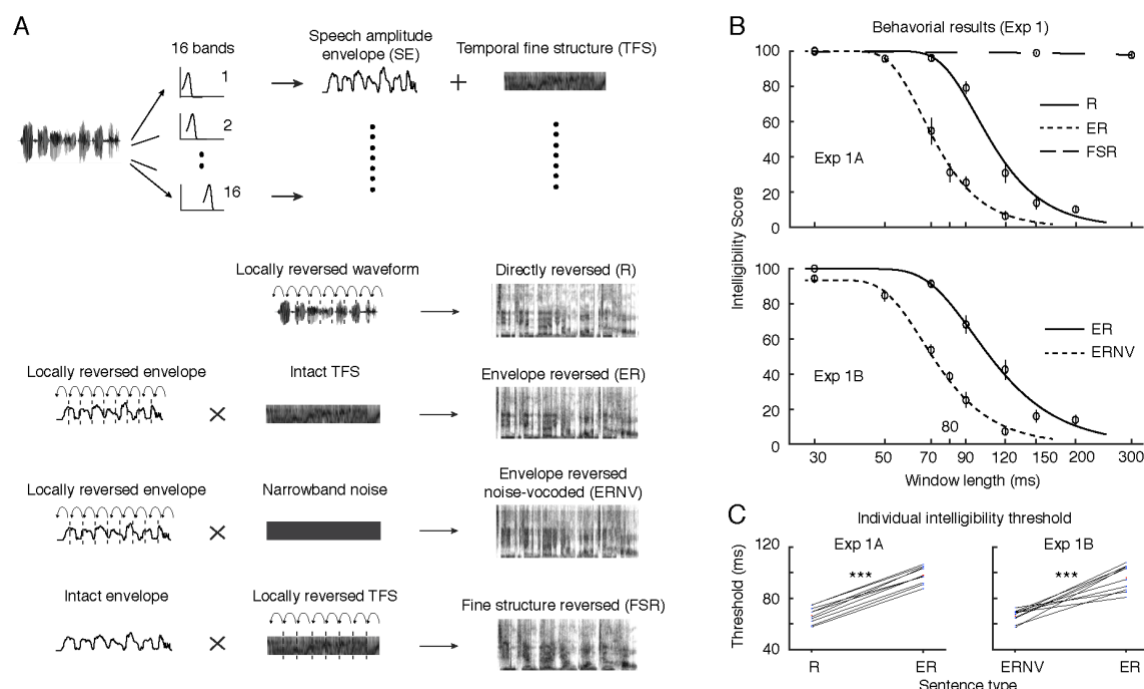
531 for speech segmentation.

Figure 3.

Illustration of stimulus manipulations and behavioral results. (A) Speech signals were first decomposed into the amplitude envelope (SE) and the temporal fine structure (TFS). Rectangular windows with various lengths were used to segment speech signals and then locally reverse each speech segment in time. Depending on which part of the speech signal was locally reversed, four types of reversed sentences were generated: directly reversed (R) – raw broadband speech signals were locally reversed; envelope reversed (ER) – we reversed only the envelope and then used it to modulate the intact TFS; envelope reversed noise-vocoded (ERNV) – we reversed the envelope and used it to modulate narrow band noise; fine structure reversed (FSR) – we reversed the TFS and used the intact envelope to modulate the reversed TFS. (B) Behavioral results of Intelligibility of Experiments 2A and 2B. The top panels show group-averaged results of Experiment 2A and Experiment 2B and the psychometric functions fit to the data. The x-axis represents window length, the y-axis represents the intelligibility score. It can be clearly seen that ER sentences with intact TFS show significantly higher intelligibility than sentences without intact TFS (R in Expt. 2A and ERNV in Expt. 2B). The lower panel shows 50% thresholds of psychometric function fits to individual data. Each black line represents each subject's individual threshold. The x-axis shows sentence type and the y-axis the threshold. Asterisks show significance levels (***, $p < 0.001$). The error bars are +/- SEM over subjects.

In Experiment 2A, we directly reversed the speech segments and varied the segment size to generate directly reversed (R) sentences – i.e. SE and TFS were both reversed.

Next, we locally reversed the SE and used this reversed SE to modulate the intact TFS to generate envelope reversed (ER) sentences. For fine structure reversed (FSR) sentences, we used the intact SE to modulate the reversed TFS. An illustration of the stimulus

28

555   generation and spectrograms of one example sentence is shown in Fig 3A (See Methods

556   for details). We recruited 10 Chinese native speakers and tested speech intelligibility for

557   these sentences (ten sentences for each segment size). For Experiment 1B, we created

558   envelope reversed noise-vocoded (ERNV) sentences by using the reversed SE to

559   modulate noise (Fig 3A). By comparing intelligibility of 10 Chinese native speakers for

560   the ER sentences with the ERNV sentences, we further evaluated the contribution of the

561   TFS. The behavioral results are shown in Fig 3B.

562   We fit each participant's intelligibility scores to a psychometric function (see

563   Methods) and conducted a two-way mixed effect ANOVA on the thresholds of the

564   psychometric functions (fifty percent correct threshold). We treated Experiment

565   (Experiment 2A and Experiment 2B) as the between-subjects factor and TFS (with intact

566   TFS: ER, or without intact TFS: R and ERNV) as the within-subject factor. We found a

567   significant main effect of TFS ($F(1,18) = 227.14$, $p < 0.001$, $\eta_p^2 = .927$), but not for

568   Experiment ($F(1,18) = .233$, $p = 0.635$, $\eta_p^2 = .013$). In a post-hoc test, we found in

569   Experiment 1A that thresholds (which are best interpreted as temporal tolerance with

570   respect to the reversal distortion) in the ER block were significantly larger than in the R

571   block (paired sample t-test: $t(9) = 26.41$, $p < .001$, $d = 8.35$, 95% CI [33.83, 29.49]); in

572   Experiment 1B, thresholds in the ER block were significantly larger than the ERNV

573   block (paired sample t-test: $t(9) = 7.71$, $p < .001$, $d = 2.44$, 95% CI [38.65, 21.12]).

574   The behavioral results for R sentences replicate previous findings that the

575   intelligibility of locally reversed speech degrades with increased segment size (Saberi and

576   Perrott 1999; Kiss et al. 2008; Stilp et al. 2010). Although Mandarin is a tone language,

577   and one might have expected tone-related differences in behavioral thresholds, the results

29

578   of R sentences are comparable to the results of previous studies using sentences from

579   non-tonal languages (Saberi and Perrott 1999; Stilp et al. 2010).

580   Next we turn to the novel manipulations. In Experiment 2A, speech intelligibility for

581   the ER sentences was significantly higher than for the R sentences. As the key difference

582   between the R and ER sentences is only whether the TFS is disrupted, this result suggests

583   that the intact TFS in the ER sentences aids in restoring the temporal information in the

584   reversed speech by increasing intelligibility. In Experiment 2B, we also found that the

585   intelligibility threshold for the ER sentences was significantly larger than for the ERNV

586   sentences, and that the intelligibility difference between ERNV and ER sentences was

587   comparable to the difference between the R and ER sentences. This second result lends

588   strong support to the hypothesis that the auditory system extracts temporal information

589   from the TFS to compensate for the disrupted SE.

590   Speech intelligibility for the FSR sentences did not change across different segment

591   sizes. This could be because the intact SE supplied sufficient information for speech

592   segmentation even though TFS was compromised. This could be considered a similar

593   case as Shannon et al. (1995), where noise modulated by four bands of the SE was

594   sufficient for speech perception. The data invite the hypothesis that the temporal

595   information in the TFS only starts to become a significant cue when the SE is disrupted.

596   The gain in speech intelligibility by adding the intact TFS may be caused by an

597   interaction at the acoustic level between TFS and the reversed envelope, such that the

598   original envelope is recovered (Kates 2011; Shamma and Lorenzi 2013). Admittedly,

599   while the shape of the envelope can be changed depending on its carrier, we argue that

600   this contributes little to the gain in intelligibility. In Figure 4, we show that the spectra of

601   the envelopes of the ER and ERNV sentences do not significantly differ across different

602   segment sizes. We first generated the ER and ERNV sentences from 20 raw sentences

603   and then extracted their envelopes using 16 bands. We computed the average spectra of

604   the ER and ERNV envelopes (Fig. 4*A*, upper panel) and then the differential spectra

605   between the ER and ERNV envelopes (Fig. 4*A*, lower panel). To quantify whether the

606   differences between the ER and ERNV envelope spectra were significant, we computed a

607   bootstrap threshold with a two-sided alpha level of 0.05: we randomly sampled from the

608   ER and ERNV sentences to form two new groups of sentences and then computed a

609   differential spectrum. We repeated this procedure 1000 times to generate thresholds for

610   the differential spectra between ER and ERNV. For window lengths of 70 and 90 ms,

611   speech intelligibility differed significantly between ER and ERNV sentences (Fig. 3*C*),

612   but there were no significant differences of the spectra of the envelopes (Fig. *4B*).
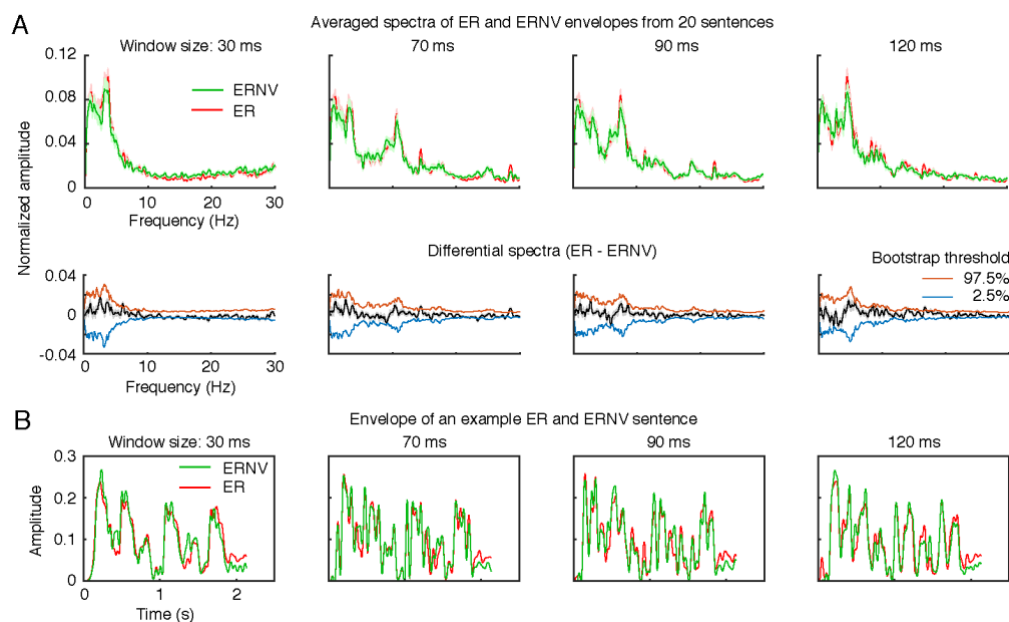
613

614

615   **Figure 4.**

616   (A) The upper panels: averaged spectra of the envelopes from 20 ER and ERNV sentences across four
617   window lengths used in Experiment 1. The green line represents the ERNV sentences; the red line
618   represents the ER sentences. The x-axis is frequency and the y-axis shows normalized amplitude of the
619   spectra. The lower panel depicts the differential spectra computed by subtracting the spectra of ERNV
620   sentences from the spectra of ER sentences. The black line represents the differential spectra and the red
621   and blue lines show bootstrap thresholds. The envelopes of the ER and ERNV sentences do not
622   significantly differ. Note that this is also true for the window lengths of 70 ms and 90 ms, where speech
623   intelligibility for the ER sentences was significantly higher than the ERNV sentences. (B) Example
624   envelopes of an ER and ERNV sentence across different window lengths. The shaded area represents +/-
625   SEM over sentences.

626

627 **Discussion**

628 Previous studies that principally focused on the speech amplitude envelope revealed only

629 part of the mechanism for segmentation. The SE and the TFS convey comparable

630 temporal information to elicit cortical entrainment. The auditory system relies on various

631 cues in speech signals, both temporal and spectral, for segmentation. The data we

632 presented demonstrate that the TFS contains temporal information that can be used for

633 speech segmentation by entraining cortical oscillatory responses. Using MEG

634 experiments, we first demonstrated that the TFS could entrain cortical oscillatory

635 responses, with phase patterns specific to a particular TFS. We then evaluated

636 contributions to temporal information from the recovered envelope and the spectral

637 dynamics in the TFS and found that the temporal information of the TFS comes primarily

638 from the dynamics of its spectral structure, which can be captured using spectral

639 correlation. A further analysis on the effect of the number of frequency bands showed

640 that spectral resolution plays a major role in extracting temporal information from TFS.

641 Using behavioral measurements, we next showed that the TFS contains critical acoustic

642 cues for the auditory system to restore the temporal structure of speech and aids speech

643 recognition.

644 **The data explain previous findings on the benefit of TFS in challenging listening**

645 **environments.**

646 Our behavioral results showed that TFS helps the auditory system restore temporal

647 information when the SE is disrupted. This result echoes previous findings that TFS helps

648 increase speech intelligibility under challenging environments (Hopkins et al. 2008;

649 Moore 2008) and further suggests that the gain from the TFS is a result of

650   complementary temporal information contained in the TFS. The auditory system can

651   monitor spectral changes to recover temporal information lost due to a disrupted

652   envelope.

653   Our results involving cortical entrainment evoked by the TFS could explain the

654   finding that, when the SE is smoothed by added noise, robust cortical entrainment can

655   still be found (Zoefel and VanRullen 2015). The smoothing procedure used in Zoefel et

656   al. (2015) operated on each frequency band and could have left the TFS intact and

657   therefore, the auditory system could still extract sufficient temporal information from the

658   spectral structure of the processed speech signals. Other data show that the TFS helps the

659   speech amplitude envelope entrain cortical oscillatory responses in noise (Ding et al.

660   2014), which is presumably because the TFS contains temporal information similar to the

661   SE and can provide comparable temporal information to entrain cortical oscillatory

662   responses when the envelope is disrupted by white noise.

663   **The auditory system extracts temporal information from both the TFS and the SE.**

664   Previous studies (Ghitza 2001; Zeng et al. 2004) suggest that the auditory system extracts

665   temporal information from the TFS through the recovered envelope. Our data support this

666   general point. Furthermore, we also found that the recovered envelope by itself cannot

667   fully explain the amount of temporal information in the TFS (Fig. 2*A*). Our MI analysis

668   between SC and the MEG phase series reveals that the auditory system monitors spectral

669   changes in the TFS to extract temporal information, and that the dynamics of the spectral

670   structure in the TFS explains a larger amount of variance, indicated by MI analysis, in the

671   MEG phase series than the recovered envelope.

672   The TFS preserves the rich spectral information of speech and carries information

673   relevant to pitch (Moore and Moore 2003; Moore et al. 2006), lexical tone (Xu and

674   Pfingst 2003; Zeng et al. 2005), and the acoustic transitions between consonants and

675   vowels (Rosen 1992). All this information in the spectral domain of speech provides

676   dynamic cues for speech segmentation. We used SC and CSC to extract temporal

677   information from TFS, which we aim to reveal in the TFS how spectral information

678   changes along time and where temporal information exists. This does not necessarily

679   mean that SC and CSC represent how exactly the auditory system extracts information

680   from the TFS. SC and CSC can be viewed as one of many algorithms that the auditory

681   system implements to process the TFS, and other algorithms may also achieve a similar

682   computational goal of recovering temporal information from the TFS (Shamma and

683   Lorenzi 2013; Ewert et al. 2018).

684   **Spectral resolution plays a key role in processing the TFS.**

685   We found that the number of frequency bands strongly modulates the extraction of

686   temporal information from the TFS (Fig. 2*C*). Although temporal acuity, which is

687   important for processing temporal fluctuations in speech, is relatively preserved in

688   cochlear implant users, deficits of perceiving speech in challenging environments among

689   people with hearing loss still exists (Oxenham and Kreft 2014). It has been argued that

690   the reduced spectral resolution smears spectral details of speech and maskers (Fu and

691   Nogaki 2005; Oxenham and Kreft 2014). Our results suggest that the reduced spectral

692   resolution smears the spectral details of TFS which results in poor extraction of temporal

693   information from the spectral structure in speech. This smearing effect prevents the

35

694 auditory system from parsing the speech stream using spectral contents and therefore

695 leads to degraded intelligibility.

696 The correlation between spectral resolution and speech intelligibility (Oxenham and

697 Kreft 2014) could be due to the way the auditory system groups speech information in

698 terms of spectral and temporal structure. Modulated noise often reduces listeners'

699 sensitivity to the TFS (Hopkins and Moore 2011), which could be because modulated

700 noise contains dynamic changes in spectral contents and interferes with the spectral

701 structure preserved in TFS. Reduced spectral resolution prevents the auditory system both

702 from tracking spectral changes in speech signals and from separating maskers from target

703 speech based on spectral details.

704 **The relationship between the envelope and fine structure.**

705 Speech is often investigated as an envelope and a fine structure. As the envelope

706 manifests characteristic temporal structures of speech signals, such as syllable and

707 phoneme structures (Di Liberto et al. 2016), mechanisms revealed by studies on the

708 envelope are often argued to be related to temporal coding. Therefore, many studies on

709 speech segmentation mostly focus on the SE and separate between the SE and the TFS.

710 The present study indicates a blurred line between the envelope and TFS in speech

711 segmentation, as our results show that the auditory system can rely on the TFS to

712 temporally group speech information. The common practice of using band-filtering

713 methods to separate envelope and TFS may not be an ideal way to extract critical features

714 from speech signals, as previous finding shows that the TFS and the SE code comparable

715 information in the early/periphery auditory system and therefore cannot be fully

716 separated (Shamma and Lorenzi 2013). Speech segmentation could rely on local

717   temporal-spectral operations, such as SC and CSC used in the present study and spectro-

718   temporal receptive fields found in many studies (Theunissen et al. 2000; Eggermont

719   2001; Machens et al. 2004; Ding and Simon 2013; Mesgarani et al. 2014).

720   We calculated the modulation power spectra of raw speech signals and the TFS, using

721   different numbers of frequency bands (Singh and Theunissen 2003; Elliott and

722   Theunissen 2009). We created time-frequency representations of the speech signals and

723   TFS using the log amplitude of their spectrograms obtained with Gaussian windows. We

724   then applied the 2D Fourier Transform to the spectrograms and created modulation power

725   spectra by taking the amplitude squared as a function of the Fourier pairs of the time and

726   frequency axes. We illustrate in Figure 5 that the TFS represents the residual signals left

727   by envelope extraction - the higher the number of frequency bands used, the less

728   modulation power was left in the TFS. This demonstration further argues that TFS and

729   envelope cannot be fully separated and the auditory system segments speech signals in a

730   'holistic' or 'synthetic' manner, concurrently and cooperatively using both temporal and

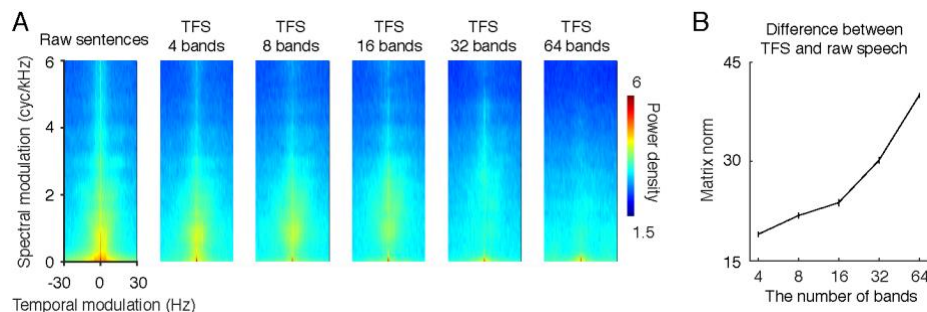731   spectral information.

732

733   **Figure 5.**

734   Averaged modulation power spectrum of speech signals and TFS. (*A*)We selected 20 sentences used in the
735   present study for analysis and averaged the results across 20 sentences. The x-axis of each plot represents
736   temporal modulation and the y-axis spectral modulation. From left to right, the modulation power spectrum
737   was calculated for raw sentences and the TFS reconstructed using different numbers of frequency bands.
738   The modulation power decreases as more frequency bands are used to extract the envelope. (*B*) We

739   quantified the difference of the modulation power spectrum between the raw sentences and the TFS by first
740   calculating subtraction between the modulation power spectrum of the raw sentences and the TFS and then
741   taking the norm. It can be seen that, as the number of frequency bands increases, the TFS preserves less
742   modulation information of the raw sentences (the matrix norm increases).
743

744   **Conclusion: speech segmentation from a synthetic perspective.**

745   The envelope and the temporal fine structure of speech are often studied separately in the

746   context of cortical entrainment to speech and speech segmentation. Our study

747   demonstrates that the auditory system treats speech as a whole, and that temporal speech

748   segmentation involves processing both the temporal and spectral contents of speech. This

749   view provides new interpretations to previous findings and also generates new

750   hypotheses for the future study of the neural basis of speech segmentation.

751

752   References

753   Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and
754   powerful approach to multiple testing. J Roy Stat Soc B Met. 57:289–300.
755   Cogan GB, Poeppel D. 2011. A mutual information analysis of neural coding of speech
756   by low-frequency MEG phase information. J Neurophysiol. 106:554–563.
757   de Cheveigné A, Simon JZ. 2007. Denoising based on time-shift PCA. J Neurosci Meth.
758   165:297–305.
759   de Cheveigné A, Simon JZ. 2008. Sensor noise suppression. J Neurosci Meth. 168:195–
760   202.
761   Di Liberto GM, O'Sullivan JA, Lalor EC. 2016. Low-Frequency Cortical Entrainment to
762   Speech Reflects Phoneme-Level Processing. Curr Biol. 25:2457–2465.
763   Ding N, Chatterjee M, Simon JZ. 2014. Robust cortical entrainment to the speech
764   envelope relies on the spectro-temporal fine structure. NeuroImage. 88 IS -:41–46.
765   Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D. 2017. Temporal modulations in
766   speech and music. Neurosci Biobehav R. 81:181–187.
767   Ding N, Simon JZ. 2012. Emergence of neural encoding of auditory objects while
768   listening to competing speakers. Proc Natl Acad Sci U S A. 109:11854–11859.
769   Ding N, Simon JZ. 2013. Adaptive Temporal Encoding Leads to a Background-
770   Insensitive Cortical Representation of Speech. J Neurosci. 33:5728–5735.
771   Ding N, Simon JZ. 2014. Cortical entrainment to continuous speech: functional roles and
772   interpretations. Front Hum Neurosci. 8:1–7.
773   Doelling KB, Arnal LH, Ghitza O, Poeppel D. 2014. Acoustic landmarks drive delta–
774   theta oscillations to enable speech comprehension by facilitating perceptual parsing.
775   NeuroImage. 85, Part 2 IS -:761–768.
776   Eggermont JJ. 2001. Temporal modulation transfer functions in cat primary auditory
777   cortex: separating stimulus effects from neural mechanisms. J Neurophysiol. 87:305–
778   321.
779   Elliott TM, Theunissen FE. 2009. The Modulation Transfer Function for Speech
780   Intelligibility. PLoS Comput Biol. 5:1–14.
781   Ewert SD, Paraouty N, Lorenzi C. 2018. A two-path model of auditory modulation
782   detection using temporal fine-structure and envelope cues. European Journal of
783   Neuroscience.
784   Fu Q-J, Nogaki G. 2005. Noise Susceptibility of Cochlear Implant Users: The Role of
785   Spectral Resolution and Smearing. JARO. 6:19–27.
786   Fu Q-J, Zhu M, Wang X. 2011. Development and validation of the Mandarin speech
787   perception test. J Acoust Soc Am. 129:EL267–EL273.
788   Ghitza O. 2001. On the upper cutoff frequency of the auditory critical-band envelope
789   detectors in the context of speech perception. J Acoust Soc Am. 110:1628–1640.
790   Ghitza O. 2012. On the Role of Theta-Driven Syllabic Parsing in Decoding Speech:
791   Intelligibility of Speech with a Manipulated Modulation Spectrum. Front Psychol. 3.
792   Ghitza O, Greenberg S. 2009. On the Possible Role of Brain Rhythms in Speech
793   Perception: Intelligibility of Time-Compressed Speech with Periodic and Aperiodic
794   Insertions of Silence. Phonetica. 66:113–126.

795 Gilbert G, Bergeras I, Voillery D, Lorenzi C. 2007. Effects of periodic interruptions on
796   the intelligibility of speech based on temporal fine-structure or envelope cues. J
797   Acoust Soc Am. 122:1336–1339.
798 Gilbert G, Lorenzi C. 2006. The ability of listeners to use recovered envelope cues from
799   speech fine structure. J Acoust Soc Am. 119:2438–2444.
800 Giraud A-L, Poeppel D. 2012. Cortical oscillations and speech processing: emerging
801   computational principles and operations. Nat Neurosci. 15:511–517.
802 Glasberg BR, Moore BCJ. 1990. Derivation of auditory filter shapes from notched-noise
803   data. Hear Res. 47:103–138.
804 Gross J, Hoogenboom N, Thut G, Schyns P, Panzeri S, Belin P, Garrod S. 2013. Speech
805   rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol.
806   11:e1001752.
807 Henry MJ, Herrmann B, Obleser J. 2014. Entrained neural oscillations in multiple
808   frequency bands comodulate behavior. Proc Natl Acad Sci U S A. 111:14935–14940.
809 Henry MJ, Obleser J. 2012. Frequency modulation entrains slow neural oscillations and
810   optimizes human listening behavior. Proc Natl Acad Sci U S A. 109:20095–20100.
811 Herrmann B, Henry MJ, Grigutsch M, Obleser J. 2013. Oscillatory Phase Dynamics in
812   Neural Entrainment Underpin Illusory Percepts of Time. J Neurosci. 33:15799–
813   15809.
814 Hopkins K, Moore BCJ. 2011. The effects of age and cochlear hearing loss on temporal
815   fine structure sensitivity, frequency selectivity, and speech reception in noise. J
816   Acoust Soc Am. 130:334–349.
817 Hopkins K, Moore BCJ, Stone MA. 2008. Effects of moderate cochlear hearing loss on
818   the ability to benefit from temporal fine structure information in speech. J Acoust Soc
819   Am. 123:1140–1153.
820 Hopkins K, Moore BCJ, Stone MA. 2010. The effects of the addition of low-level, low-
821   noise noise on the intelligibility of sentences processed to remove temporal envelope
822   information. J Acoust Soc Am. 128:2150–2161.
823 Kates JM. 2011. Spectro-temporal envelope changes caused by temporal fine structure
824   modification. J Acoust Soc Am. 129:3981–3990.
825 Kayser C, Ince RAA, Panzeri S. 2012. Analysis of slow (theta) oscillations as a potential
826   temporal reference frame for information coding in sensory cortices. PLoS Comput
827   Biol. 8:e1002717.
828 Kayser SJ, Ince RAA, Gross J, Kayser C. 2015. Irregular Speech Rate Dissociates
829   Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha. J Neurosci.
830   35:14691–14701.
831 Kerlin JR, Shahin AJ, Miller LM. 2010. Attentional Gain Control of Ongoing Cortical
832   Speech Representations in a "Cocktail Party." J Neurosci. 30:620–628.
833 Kiss M, Cristescu T, Fink M, Wittmann M. 2008. Auditory language comprehension of
834   temporally reversed speech signals in native and non-native speakers. Acta
835   neurobiologiae experimentalis. 68:204.
836 Lachaux J-P, Rodriguez E, Martinerie J, Varela FJ. 1999. Measuring phase synchrony in
837   brain signals. Hum Brain Mapp. 8:194–208.
838 Léger AC, Reed CM, Desloge JG, Swaminathan J, Braida LD. 2015. Consonant
839   identification in noise using Hilbert-transform temporal fine-structure speech and

840   recovered-envelope speech for listeners with normal and impaired hearinga). J
841   Acoust Soc Am. 138:389–403.

842 Lorenzi C, Gilbert G, Carn H, Garnier S, Moore BCJ. 2006. Speech perception problems
843   of the hearing impaired reflect inability to use temporal fine structure. Proc Natl
844   Acad Sci U S A. 103:18866–18869.

845 Luo H, Poeppel D. 2007. Phase patterns of neuronal responses reliably discriminate
846   speech in human auditory cortex. Neuron. 54:1001–1010.

847 Luo H, Poeppel D. 2012. Cortical oscillations in auditory perception and speech:
848   evidence for two temporal windows in human auditory cortex. Front Psychol. 3:170.

849 Machens CK, Wehr MS, Zador AM. 2004. Linearity of Cortical Receptive Fields
850   Measured with Natural Sounds. J Neurosci. 24:1089–1100.

851 Macmillan NA, Creelman CD. 2004. Detection Theory: A User's Guide. Taylor &
852   Francis.

853 Magri C, Whittingstall K, Singh V, Logothetis NK, Panzeri S. 2009. A toolbox for the
854   fast information analysis of multiple-site LFP, EEG and spike train recordings. BMC
855   Neurosci. 10:81.

856 Mesgarani N, Cheung C, Johnson K, Chang EF. 2014. Phonetic Feature Encoding in
857   Human Superior Temporal Gyrus. Science. 343:1006–1010.

858 Montemurro MA, Senatore R, Panzeri S. 2007. Tight data-robust bounds to mutual
859   information combining shuffling and model selection techniques. Neural Comput.
860   19:2913–2957.

861 Moore BCJ. 2008. The Role of Temporal Fine Structure Processing in Pitch Perception,
862   Masking, and Speech Perception for Normal-Hearing and Hearing-Impaired People.
863   JARO. 9:399–406.

864 Moore BCJ, Glasberg BR, Flanagan HJ, Adams J. 2006. Frequency discrimination of
865   complex tones; assessing the role of component resolvability and temporal fine
866   structure. J Acoust Soc Am. 119:480–490.

867 Moore GA, Moore BCJ. 2003. Perception of the low pitch of frequency-shifted
868   complexes. J Acoust Soc Am. 113:977–985.

869 Ng BSW, Logothetis NK, Kayser C. 2013. EEG phase patterns reflect the selectivity of
870   neural firing. Cereb Cortex. 23:389–398.

871 Oldfield RC. 1971. The assessment and analysis of handedness: The Edinburgh
872   inventory. Neuropsychologia. 9:97–113.

873 Oostenveld R, Fries P, Maris E, Schoffelen J-M. 2011. Fieldtrip: open source software
874   for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput
875   Intell Neurosci. 2011:1–9.

876 Oxenham AJ. 2018. How We Hear: The Perception and Neural Coding of Sound. Annu
877   Rev Psychol. 69:27–50.

878 Oxenham AJ, Kreft HA. 2014. Speech Perception in Tones and Noise via Cochlear
879   Implants Reveals Influence of Spectral Resolution on Temporal Processing. Trends
880   in Hearing. 18:2331216514553783.

881 Panzeri S, Brunel N, Logothetis NK, Kayser C. 2010. Sensory neural codes using
882   multiplexed temporal scales. Trends Neurosci. 33:111–120.

883 Panzeri S, Senatore R, Montemurro MA, Petersen RS. 2007. Correcting for the Sampling
884   Bias Problem in Spike Train Information Measures. J Neurophysiol. 98:1064–1072.

885   Patterson RD. 1976. Auditory filter shapes derived with noise stimuli. J Acoust Soc Am.
886   59:640.
887   Patterson RD, Nimmo-Smith I, Holdsworth J, Rice P. 1987. An efficient auditory
888   filterbank based on the gammatone function. Ina meeting of the IOC Speech Group
889   on Auditory Modelling at RSRE. 2.
890   Peelle JE, Gross J, Davis MH. 2013. Phase-locked responses to speech in human auditory
891   cortex are enhanced during comprehension. Cerebral Cortex. 23:1378–1387.
892   Poeppel D. 2003. The analysis of speech in different temporal integration windows:
893   cerebral lateralization as "asymmetric sampling in time." Speech Commun. 41:245–
894   255.
895   Pola G, Thiele A, Hoffmann KP, Panzeri S. 2003. An exact method to quantify the
896   information transmitted by different mechanisms of correlational coding. Network.
897   14:35–60.
898   Prins N, Kingdom FAA. 2009. Palamedes: Matlab routines for analyzing psychophysical
899   data. http://www.palamedestoolbox.org.
900   Quian Quiroga R, Panzeri S. 2009. Extracting information from neuronal populations:
901   information theory and decoding approaches. Nat Rev Neurosci. 10:173–185.
902   Roberts TP, Ferrari P, Stufflebeam SM, Poeppel D. 2000. Latency of the auditory evoked
903   neuromagnetic field components: stimulus dependence and insights toward
904   perception. J Clin Neurophysiol. 17:114–129.
905   Rosen S. 1992. Temporal Information in Speech: Acoustic, Auditory and Linguistic
906   Aspects. Philosophical Transactions of the Royal Society of London B: Biological
907   Sciences. 336:367–373.
908   Saberi K, Perrott DR. 1999. Cognitive restoration of reversed speech : Abstract : Nature.
909   Nature. 398:760–760.
910   Shamma S, Lorenzi C. 2013. On the balance of envelope and temporal fine structure in
911   the encoding of speech in the early auditory system. J Acoust Soc Am. 133:2818–
912   2833.
913   Shannon RV, Zeng FG, Kamath V, Wygonski J. 1995. Speech Recognition with
914   Primarily Temporal Cues. Science.
915   Sheft S, Ardoint M, Lorenzi C. 2008. Speech identification based on temporal fine
916   structure cues. J Acoust Soc Am. 124:562–575.
917   Singh NC, Theunissen FE. 2003. Modulation spectra of natural sounds and ethological
918   theories of auditory processing. J Acoust Soc Am. 114:3394–3411.
919   Smith ZM, Delgutte B, Oxenham AJ. 2002. Chimaeric sounds reveal dichotomies in
920   auditory perception. Nature. 416:87–90.
921   Stilp CE, Kiefte M, Alexander JM, Kluender KR. 2010. Cochlea-scaled spectral entropy
922   predicts rate-invariant intelligibility of temporally distorted sentences. J Acoust Soc
923   Am. 128:2112–2126.
924   Stilp CE, Kluender KR. 2010. Cochlea-scaled entropy, not consonants, vowels, or time,
925   best predicts speech intelligibility. Proc Natl Acad Sci U S A. 107:12387–12392.
926   Swaminathan J, Mason CR, Streeter TM, Best V, Roverud E, Kidd G. 2016. Role of
927   Binaural Temporal Fine Structure and Envelope Cues in Cocktail-Party Listening. J
928   Neurosci. 36:8250–8257.

929   Søndergaard PL, Majdak P. 2013. The Auditory Modeling Toolbox. In: Blauert J editor.
930   The Technology of Binaural Listening. Berlin, Heidelberg: Springer Berlin
931   Heidelberg. p. 33–56.
932   Teng X, Tian X, Rowland J, Poeppel D. 2017. Concurrent temporal channels for auditory
933   processing: Oscillatory neural entrainment reveals segregation of function at different
934   scales. PLoS Biol. 15:e2000812.
935   Theunissen FE, Sen K, Doupe AJ. 2000. Spectral-temporal receptive fields of nonlinear
936   auditory neurons obtained using natural sounds. J Neurosci. 20:2315–2331.
937   Xu L, Pfingst BE. 2003. Relative importance of temporal envelope and fine structure in
938   lexical-tone perception (L). J Acoust Soc Am. 114:3024–3027.
939   Zeng F-G, Nie K, Liu S, Stickney G, Del Rio E, Kong Y-Y, Chen H. 2004. On the
940   dichotomy in auditory perception between temporal envelope and fine structure cues
941   (L). J Acoust Soc Am. 116:1351–1354.
942   Zeng F-G, Nie K, Stickney GS, Kong Y-Y, Vongphoe M, Bhargave A, Wei C, Cao K.
943   2005. Speech recognition with amplitude and frequency modulations. Proc Natl Acad
944   Sci U S A. 102:2293–2298.
945   Zhu S, Wong LLN, Chen F. 2014. Development and validation of a new Mandarin tone
946   identification test. Int J Pediatr Otorhinolaryngol. 78:2174–2182.
947   Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM,
948   Goodman RR, Emerson RG, Mehta AD, Simon JZ, Poeppel D, Schroeder CE. 2013.
949   Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a
950   "Cocktail Party." Neuron. 77:980–991.
951   Zoefel B, VanRullen R. 2015. Selective Perceptual Phase Entrainment to Speech Rhythm
952   in the Absence of Spectral Energy Fluctuations. J Neurosci. 35:1954–1964.
953   Zoefel B, VanRullen R. 2016. EEG oscillations entrain their phase to high-level features
954   of speech sound. NeuroImage. 124:16–23.
955