Research Report

Biological Sciences: Neuroscience

Cortical tracking of speakers' formant changes predicts selective listening

Francisco Cervantes Constantino[1,2*] **and Ángel A. Caputi**[1]

[1]Instituto de Investigaciones Biológicas Clemente Estable; [2]Universidad de la República

*For correspondence: fcervantes@psico.edu.uy

Present address: Centro de Investigación Básica en Psicología, Universidad de la República, Montevideo, 11.600 Uruguay

Keywords: formant | phase locking | selective attention

Abstract A social scene is particularly informative when people are distinguishable. To understand somebody amid a 'cocktail party' chatter, we automatically index their voice. This ability is underpinned by parallel processing of vocal spectral contours from speech sounds, but it has not yet been established how this occurs in the brain's cortex. We investigate single-trial neural tracking of slow modulations in speech formants using electroencephalography. Participants briefly listened to unfamiliar single speakers, and in addition, they performed a cocktail party comprehension task. Quantified through stimulus reconstruction methods, robust tracking was found in neural responses to slow (delta-theta range) modulations of the fourth and fifth formant band contours equivalent to the 3.5–5 KHz audible range. Instantaneous inter-formant spacing (ΔF), which also yields indexical information from the vocal tract, was similarly decodable. Moreover, EEG evidence of listeners' spectral tracking abilities predicted their chances of succeeding at selective listening when faced with two-speaker speech mixtures. In summary, the results indicate that the communicating brain can rely on locking of cortical rhythms to major changes led by upper resonances of the vocal tract. Their corresponding articulatory mechanics hence continuously issue a fundamental credential for listeners to target in real time.

Significance statement The human voice is acoustically fingerprinted, thanks in part to the shape and function of the larynx and oral cavities as they change from person to person, and from one instant to the next. This study shows that a substantial portion of these time-varying signatures are consistently traced by cortical networks. People who are more capable of tracking these variations neurally are more likely to grasp speakers of their choice despite masking by others. Thus, the changes can be used for the brain to tell what pieces of an unfamiliar voice fit together and which ones to put apart. Solving this inference in real time represents a cornerstone to understanding communication in everyday's social world.

When attending a cocktail party, you might want to know *who* to hear from before *what*. The vocal chord system produces a speaker's characteristic pitch, instantaneously subject to time-varying resonance effects due to the ever-changing articulatory manoeuvres of the supraglottal vocal tract. The resulting family of formant patterns jointly serves to cue and index speech from its vocal source. In many species, auditory processing of pitch and formants facilitates intraspecific recognition. These spectral cues often signal important anatomical and social features, including age, sex, body size, reproductive fitness, and social status or significance [1], [2], [3], [4], [5], [6], [7], [8]. An observer may infer the likely size of an unseen interlocutor given the

average spacing or dispersion between the speaker's formants, for instance. This versatility suggests that such spectral channels are routinely harnessed to broadcast and receive caller identity credentials through auditory space—a feat predating by far the emergence of language.

Recent work demonstrated that the central auditory system has the ability to track instantaneous pitch changes contained in natural speech [9], [10]. Evidence of cortical encoding of frequency modulations (FM) in the voice represents a promising step to unravel how the brain targets "who" speaks [11]. A natural limitation in these studies is of course that cortical auditory population responses weakly resolve frequency-following responses at pitch range (85-200 Hz). Cortical time-locking in the even faster formant rate domains has not been observed using neural tracking approaches. Yet, it is has been long clear that representing formant content matters for cortical systems that establish speaker identity [12], [13], [14], phonemic analyses, and so on.

We hypothesized that delta and theta band population activity in the cortex, part of which is known to track slow oscillations of the acoustic envelope of speech, may also provide a channel to represent the slow changes of formants' instantaneous FM. To evaluate this possibility, we employed a stimulus reconstruction approach [15] designed to estimate the accuracy of recovering the variations of a given feature, such as a formant, from a linear mapping of a listener's electroencephalography (EEG) dataset. Of particular interest among spectral features was the contour of the average inter-formant spacing. This measure is sometimes used as proxy estimator for vocal tract length (VTL), a relevant parameter in automatic speaker recognition and vowel normalization among multiple talkers [16], [17]. Voice processing systems that faithfully and continually represent changes in the formants and formant spacing of an unfamiliar talker, may be more likely to scaffold the variable utterances belonging to a phonetically invariant source. If these dynamics are temporally encoded during speech processing, we then addressed their ability to predict success in the problem of selective listening during a multispeaker 'cocktail-party' scenario.

Results

Using stimulus reconstruction, we investigated the ability to decode seven acoustic features describing important frequency modulations of voiced speech, in addition to its envelope. These relate to delta/theta band modulations of FM changes in (*i*) the instantaneous fundamental frequency, (*ii*-*vi*) instantaneous F_1, F_2, F_3, F_4, F_5 formant frequencies, and (*vii*) instantaneous inter-formant frequency spacing (Figure 1). To avoid the confound of encoding differences arising from the significant sexual dimorphism of vocal traits, and that higher-frequency tracking may not be clearly observed in the EEG from female speaker listening [9], decoder analyses were limited to masculine voice neural data. Figure 2A depicts a schematic example of slow envelope reconstruction from a single speech stimulus. Here, the average reconstruction waveform is built after EEG recordings from listeners that were presented with the same example sentence. By contrast, decoders trained from reshuffled EEG data yield a lower baseline performance reference level, an indication of robustness in empirical neural performance. In addition, the same EEG data suggest parallel evidence of tracking responses to the instantaneous spectral dynamics of inter-formant frequency spacing ΔF.

Decodable slow dynamics of vocal spectral features of speech from single-trial recordings

To assess whether such decoding is significant, the relative change in performance between empirical and reference reconstruction performance was measured for each of the eight analyzed acoustic features in natural speech (Figure 2B). As expected, the decodability of slow modulations of the acoustic envelope from the single-trial EEG during single-speaker listening was significantly greater than zero (mean 19.9%, SD = 29.9%) $t(72) = 5.7$, adjusted (Bonferroni-Holm) $p = 10^{-5}$. In the case of slow FMs of the instantaneous higher formants, decodability was also found to be reliable, i.e., significantly greater than zero: F_4 (mean 10.1%, SD=29.1%) $t(72) = 3.0$, adj. $p = 0.013$; F_5 (mean 7.9%, SD=23.8%) $t(71) = 2.8$, adj. $p = 0.016$. Moreover, decodability of slow modulations in formant dispersion dynamics ΔF was also significantly greater than zero (mean 10.8%, SD = 26.5%) $t(72) = 3.5$, adj. $p = 0.003$. At the lower frequency ranges, decodability of slow FMs in the instantaneous fundamental frequency f_0 did not show significantly greater values than zero (mean 1.6%, SD = 28.4%) $t(72) = 0.5$, adj. $p > 0.31$. Null results were similarly observed for decodabilities of slow FMs of instantaneous: F_1 (mean 2.5%, SD = 20.9%) $t(72) = 1.0$, adj. $p > 0.15$; F_2 (mean 3.3%, SD = 27.4%) $t(72) = 1.0$, adj. $p > 0.15$; and F_3 (mean 6.6%, SD=30.1%) $t(72) = 1.9$, adj. $p = 0.13$. The results thus indicate that slow modulations in the spectral variations across the lower frequency bands of voiced speech were not reliably decoded from single-trial EEG data. Nevertheless, the dynamics of higher formant and spacing information used to index vocal identity may be tracked by cortical populations as a result of listening to a single speaker.

We next addressed the correspondence between neural tracking systems dedicated to AM versus FM. Specifically, we investigated whether listeners that achieve better envelope decodability are those that also achieve higher decodabilities across the relevant spectral domains F_4, F_5, and ΔF. Using a Pearson correlation matrix (Figure 2C), significant linear relationships were found yet only within the set of decodable spectral features (all adj. p-values $< 1.6\text{x}10^{-4}$). Subject decodability at any of these features did not predict the decodability of the envelope (all adj. p-values > 0.6). These results suggest that listeners that successfully encode the spectral features rely on synergistic co-modulations between the dynamics of higher formants and formant dispersion. Also, that spectral processing may provide distinctive information from the neural encoding of general sound level variations that are signaled by the speech envelope.

Successful formant reconstructions predict performance at the cocktail party

Unfamiliar multispeaker conditions involve alternative hypotheses on a priori undelimited sets of formant modulations. Our second aim was to test whether spectral cue tracking may help listeners resolve a cocktail party of unfamiliar speakers. The ability to decode spectral vocal information across speakers was summarized through the L2-norm of the vector of empirical decoder performances ($\rho_{e,F4}$, $\rho_{e,F5}$, $\rho_{e,\Delta F}$). Such a measure, $||\rho||$, is higher if subjects consistently encode slow modulations across all three significant spectral domains. We hypothesized that this ability may predict a listener's chances to reliably track one target speaker of their choice. Advantage was made of the

fact that participants also engaged in a 'cocktail-party' task: after listening to an individual speaker, they were presented with an independent two-speaker 'cocktail-party' mix and instructed to follow one. Their level of intelligibility during this stage was evaluated through a challenging comprehension questionnaire (see *Supplementary Materials and Methods*). The proportion of correct trials achieved by the listener was used as index of performance. A robust correlation analysis with interval testing [18] showed that the summary measure of spectral decodability $\|\rho\|$ significantly predicted selective performance at multispeaker conditions, r_s = 0.39, CI = 0.18 – 0.57, p = 8.05×10^{-4} (Figure 3A). The significant correlation was robust to potential outliers (Shepherd's Pi = 0.41; p = 0.001), and confirmed by a robust regression analysis $t(70)$ = 2.994; p = 0.004. Furthermore, the predictive ability of $\|\rho\|$ also extended to finer-grained subject performance measured by the average amount of questionnaire points earned per trial, r_s = 0.36, CI = 0.13 – 0.55, p = 0.002 (Figure 3B). This significant association was confirmed by robust regression $t(70)$ = 3.032; p = 0.003.

Finally, we examined whether the decodabilities of each of the three spectral features may separately predict cocktail party performance (Figure 3C). In each case, spectral decodability significantly predicted intelligibility, i.e., F_4: r_s = 0.35, CI = 0.11 – 0.55, p = 0.002; F_5: r_s = 0.32, CI = 0.09 – 0.53, p = 0.006; ΔF: r_s = 0.32, CI = 0.09 – 0.51, p = 0.006. As with the summary neurometric $\|\rho\|$, each of the correlations with individual decodabilities ($\rho_{e,F4}$, $\rho_{e,F5}$, $\rho_{e,\Delta F}$) were confirmed by robust regression analyses, F_4: $t(70)$=3.259, p=0.002; F_5: $t(70)$=2.587, p=0.012; ΔF: $t(70)$=2.563, p=0.013. The overall findings provide strong evidence in favor of the hypothesis that, to engage selective attention successfully in challenging multispeaker conditions, listeners may rely on the ability to tune to spectro-temporal properties of speech signals typically related to speaker identity.

Discussion

Cortical systems that process natural speech avail themselves of higher-range frequency modulations, as demonstrated here with noninvasive single-trial EEG methods. First, stimulus reconstruction techniques retrieved, as they do in the case of the speech envelope, delta-theta band modulations contained in various spectral indices of speaker identity. Subsequently, listeners' ability to reconstruct spectral features predicted success at following a talker's speech in a 'cocktail-party'. Slow modulations to spectral variations in higher formants F_4 and F_5 were reliably decoded from EEG, yet this ability was not explained by reconstruction of amplitude modulations contained in the envelope, suggesting alternate processing mechanisms. We additionally present decoding findings for a key feature in voice recognition studies, the spacing between neighboring formants (ΔF), often used to infer anatomical dimensions of the source's vocal tract [19], [20]. All three spectral features considered, their joint decodabilities predicted listeners' ability to selectively attend in a two-speaker scenario.

The peripheral auditory system isolates frequency-localized components of speech and subsequent central processing for auditory scene analysis of these components occurs mainly in the time domain [21]. The extracted pitch and formant cues serve to build representations of a speaker's voice as listeners face the problem of determining speaker invariance [19]. This is a crucial aspect of listeners' organization of multispeaker scenes that needs estimating the variability that corresponds to one voice with sufficient precision to set it apart from maskers' [22], [23]. Studies suggest

that the first 400 ms of auditory processing perform the necessary spectral vocal cue processing analyses [24], [25], [26]. Continuous sequential analyses of the incoming voice signal, such as in neural tracking, further guide sustained perception of vocal identity and speech recognition [27]. The present findings add to reports of pitch contours tracked at the early auditory cortex [9], [10]. Here, pitch waveforms (75–125 Hz) were not addressed in the EEG signals at corresponding rates [9]; we did not observe the reconstruction of its slow modulations either. Our lack of significant findings for f_0, along with lower formants F_1–F_3, may suggest a lower bandwidth limit for slow FM changes in delta/theta activity. Faster rates of change occur in movements of the lower formants that encode short vowels in natural speech [28], with higher formants remaining relatively invariant. As such, higher formants may represent more stable cues for a sustained discrimination of unfamiliar speakers over just a few seconds.

Extraction of summary statistics is a means for listeners to constrain and establish an unfamiliar vocal identity from the varying speech signal [19]. In computational voice coding models, static f_0 and ΔF parameters represent coordinates of hypothesized vocal space [19], [28], [29]. These follow from distinguishable origins in source and filter models of production, namely, the individual contributions made by the larynx to pitch and by the supra-laryngeal vocal tract to formant structure. Cues based on formant dispersion like ΔF may show higher reliability for speaker identification than pitch [30]. This important feature is directly extracted by the auditory system [31], [32], [33] as early as at the medial geniculate body [34], and is cortically represented in posterior superior temporal lobe networks that include temporal voice areas [14], [35], [36]. These regions may extend across the posterior aspect of the right superior temporal sulcus to the superior temporal pole [29], [37], [38], [39], [40], [41], [42], [43], [44], [45]. Responses may be specific to voice-like stimuli regardless of intelligibility [46], indicating the ability to sustain low-level voice processing. Multispeaker conditions impose this circuitry with shorter time spans for reliable voice discrimination, yet cortical processing operates faster during the intermittent periods of masker-devoid listening [47]. Spectral cues within these reduced windows represent variations of an unfamiliar voice that can be reliably summarized at higher encoding precision. By contributing to sequential buildup of a within-person vocal prototype representation, e.g. [19], [23], tracking may increase the likelihood that a formant grouping is maintained in the presence of alternative complexes at masked periods.

Lexical processing of speech depends too on the precise encoding of vocal spectral modulations [48], [49], [50], [51], [52]. Relative temporal displacements between spectral components of speech may affect intelligibility performance [53]; this may naturally arise from surface reverberation effects that lead to weaker segregation in multispeaker conditions [54], [55]. In another example of dependence on spectral encoding, bandwidth tuning loss at higher frequencies under presbycusis severely limits the ability to target speech in multispeaker conditions. The extents to which spectrotemporal uncertainty affects the monitoring of formant groupings, and the arbitration between their integration versus segregation, remain unclear. Future studies may address what impact do limitations to tracking of spectral features of speech based on multiple formants at any given time, imply for normal hearing and clinical populations.

Materials and Methods

Participants

Seventy-three subjects (51 female; mean age 25.6 ± 4.9 SD) that were native Rioplatense Spanish speakers with no hearing impairment or neurological disorder, and correct vision, participated in the studies. Behavioral data from one participant was not stored in error. The Ethics in Research Committee at the Facultad de Psicología, Universidad de la República Uruguay, approved the study. All participants gave written consent in accordance with Declaration of Helsinki guidelines. Participants received a chocolate bar in appreciation for their time after they completed the task.

Stimuli

Stimuli were obtained from our lab original databases containing a set of short phrase audios (mean duration 8.6 s, ±0.8 s) individually narrated by 607 speakers (including 264 different male speakers). These single-speaker stimuli were sourced from Uruguayan news broadcasts, audiobooks, interviews, and recordings produced from volunteers that read aloud translated paraphrase sentences [56].

Task

Participants were instructed to listen to every sound with their eyes closed and minimizing body movement. In the first part of each trial, participants were presented with a single-speaker speech stimulus; in the second part, they were presented with a cocktail party mixture. Participants were tasked with selectively attending to one of the two speech streams in the mixture, i.e. the leading or the interjecting talker (Supplementary Figure 1), randomly with equal probability. At the end of the mix they were presented with a nine-option questionnaire, and required to respond by selecting three options corresponding to the indicated target exclusively. A feedback screen was subsequently displayed with the trial's and the accumulated scores.

Speech spectral feature analyses

Audios from presented male voice speakers at the solo speech listening period were submitted to an AM-FM analysis [57], [58], [59] in the original sampling rate. Sex selection was made in this initial study for two reasons; the first is that higher formants contain typically much less energy in female voices compared to male voices [28], likely resulting in a poor signal-to-noise ratio in EEG decoding analyses. The second reason is that evidence of f_0 tracking in the EEG to female voices appears weak as tracking strength may be inversely proportional to frequency [9]. A fixed 6-band gammatone filterbank was thus applied to each sample to assess spectral variations corresponding to male f_0 and formants' F_1-F_5 frequencies. Filterbank coefficients were based on six centre frequencies fc_i set at 146, 671, 1870, 2870, 3946, and 4869 Hz that we defined from our analysis of a local public voice database [12]. Coefficients were estimated with the Large Time/Frequency Analysis toolbox [60], [61]. Sub-bandwidths equalled thrice the critical bandwidth of the auditory filter at the corresponding fc_i, and were defined in equivalent rectangular bandwidths (ERBs) [62]. The real part of the filter coefficients was applied to each audio signal. Each filter output was analyzed to obtain instantaneous frequency $f_i(t)$ and, separately, instantaneous amplitude $a_i(t)$ content using the Hilbert method. To perform a multiband demodulation analysis, a mean-amplitude weighted short-time estimate $F_i(t)$ of the instantaneous frequency $f_i(t)$ was computed per sub-band [63]. Integration of the weighing function was

approximated by using a centred moving average with a τ=25 ms window, according to the formula:

$$F_i[t=t_0]=\frac{\sum_{t=t_0}^{t=t_0+\tau}[f_i(t)\cdot a_i^2(t)]}{\sum_{t=t_0}^{t=t_0+\tau}[a_i^2(t)]} \quad (1)$$

The resulting set of discrete short-time instantaneous frequency estimates, sometimes referred to as the pyknogram of the speech signal [63], was used to represent the dynamic waveform timeseries for f_0 and formants F_1-F_5.

The dynamic waveform for the formant dispersion ΔF measure was constructed from k=[1,…,K] successive independent linear regression models, with K the total of time samples in the stimulus. At the k-th sample, the formant model is:

$$F_i[t=t_k]=\beta_{0,k}+\beta_{1,k}n_i+\varepsilon_k \quad (2)$$

with formant frequencies F_i taken as observations at sample k, and scaled by their formant numbers n={1,2,3,4,5} which are taken as predictors. Model coefficient estimation was performed via an iteratively reweighted least squares algorithm with the bisquare weighting function (MATLAB `robustfit`). The regression slope served as the measure of formant dispersion [16]:

$$\Delta F[t=t_k]=\beta_{1,k} \quad (3)$$

The seven resulting spectral contour timeseries waveforms (f_0, F_1–F_5, ΔF) were downsampled to 1024 Hz.

EEG preprocessing

EEG data analyses were implemented in MATLAB 2018b. EEG signals were bandpass-filtered between 1 and 28 Hz with a 10-order elliptic IIR filter of 60 dB attenuation and 0.5 dB passband ripple, with group delay correction. Single channel data were rejected in a blind manner according to a channel variance-based criterion [65] with confidence coefficient $\lambda_P=4$. The same rejection procedure was performed separately for the external reference channels, with a more restrictive confidence coefficient $\lambda_P=2$. Head (and reference) sensor datasets were common median-referenced, after which the automated rejection procedure was performed for the full set. EEG data were decomposed into independent components using FastICA [66] and the two independent components with maximal broadband power were automatically selected and projected out of the data. A time-shifted principal component analysis [67] was then applied to discard environmental signals recorded on the external reference sensors, using ±4 ms time shifts. A sensor noise suppression algorithm was used to attenuate artifacts specific to any single channel, using 63 neighbors [68].

Stimulus-response alignment and decoder estimation

We measured the linear reconstruction from EEG responses of the slow modulations of (*i*) the speech stimulus envelope; (*ii*) the fundamental frequency f_0 waveform; (*iii*-*vii*) the waveforms of formants $F1$ to $F5$; and (*viii*) variations in formant dispersion ΔF dynamics. Eight decoders were separately estimated based on the empirical EEG activity. Eight additional decoders were estimated based on trial-reshuffled EEG activity, totaling a pair of decoders per stimulus feature. This served to obtain both an

empirical measure of decodability of stimulus features, and an estimation noise reference.

The operation of reconstruction of a stimulus timeseries $S_i(t)$ from c EEG channel responses R_c(t) can be formulated as:

$$\widetilde{S}_i(t)=\sum_{c=1}^{N}\sum_{\tau=0}^{T'}[R_c(t+\tau)\cdot D_{i,c}(\tau)] \tag{4}$$

where $D_{i,c}(t)$ is the linear decoder for the i-th acoustic feature at channel c. Here, N = 64 channels, and T' = 400 ms denotes the end time of the integration window. Hence, a stimulus feature timeseries at time point $t=t_k$ is reconstructed using neural responses spanning from t_k until t_k+0.4 s. Each full decoder $D_i(t)$ was independently obtained by boosting with 10-fold cross-validation [69], [70]. The stimulus reconstruction is equivalent to the true stimulus plus a the residual response waveform not explained by the linear decoder model.

$$S_i(t)=\widetilde{S}_i(t)+\epsilon_i(t) \tag{5}$$

Decoder performance as predictor of listening performance at the cocktail party

Upon identification of reliably decodable spectral features (F_4, F_5, and ΔF, see *Results*) and analysis of their joint ability to be reliably decoded from the EEG, a summary statistic was created by indexing for each participant the Euclidean norm of the vector of full empirical decodabilities:

$$\|\rho\|=\sqrt{\rho^2_{F_4,e}+\rho^2_{F_5,e}+\rho^2_{\Delta F,e}} \tag{6}$$

High scores in this parameter represent systematically high empirical decodability across all three significant spectral domain features of a participant's EEG data. The norm was used as a unified neurometric to evaluate the general hypothesis that finer reconstruction of solo vocal features' variations from the EEG predicts the listener's improved performance in the more demanding cocktail-party scenario. The proportion of correct trials in the cocktail party task was taken as a measure of participants' performance. We evaluated the quality r_s of the predictive neural-behavioral association between subjects' ||ρ|| and correct rate via Spearman's correlation with interval testing [18]. To confirm the results, additional tests were performed, including Shepherd's correlation [71] and robust regression with the Huber weighting function.

Acknowledgments and funding sources

Original EEG and behavioral data were from Thaiz Sánchez-Costa's master's thesis project, and we are thankful for her excellent data collection. We also thank Rodrigo Caramés Harcevnicow for his assistance in data analysis. The research presented in this manuscript received funding from the Agencia Nacional de Investigación e Innovación, Uruguay, under the research grant FCE_1_2019_1_155889 to FCC.

References

[1] D. A. Puts, C. R. Hodges, R. A. Cardenas, and S. J. C. Gaulin, 'Men's voices as dominance signals:: vocal fundamental and formant frequencies influence dominance attributions among men', *Evol. Hum. Behav.*, vol. 28, no. 5, pp. 340–344, Sep. 2007, doi: 10.1016/j.evolhumbehav.2007.05.002.

[2] D. A. Puts, J. L. Barndt, L. L. M. Welling, K. Dawood, and R. P. Burriss, 'Intrasexual competition among women Vocal femininity affects perceptions of attractiveness and flirtatiousness', *Personal. Individ. Differ.*, vol. 50, no. 1, pp. 111–115, Jan. 2011, doi: 10.1016/j.paid.2010.09.011.

[3] T. Riede and T. Fitch, 'Vocal tract length and acoustics of vocalization in the domestic dog (*Canis familiaris*)', *J. Exp. Biol.*, vol. 202, no. 20, pp. 2859–2867, Oct. 1999, doi: 10.1242/jeb.202.20.2859.

[4] L. Bruckert, J. S. Liénard, A. Lacroix, M. Kreutzer, and G. Leboucher, 'Women use voice parameters to assess men's characteristics', *Proc. R. Soc. B-Biol. Sci.*, vol. 273, no. 1582, pp. 83–89, Jan. 2006, doi: 10.1098/rspb.2005.3265.

[5] W. T. Fitch, 'Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques', *J. Acoust. Soc. Am.*, vol. 102, no. 2, pp. 1213–1222, Aug. 1997, doi: 10.1121/1.421048.

[6] S. Evans, N. Neave, and D. Wakelin, 'Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice', *Biol. Psychol.*, vol. 72, no. 2, pp. 160–163, May 2006, doi: 10.1016/j.biopsycho.2005.09.003.

[7] S. A. Collins, 'Men's voices and women's choices', *Anim. Behav.*, vol. 60, pp. 773–780, Dec. 2000, doi: 10.1006/anbe.2000.1523.

[8] D. Reby and K. McComb, 'Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags', *Anim. Behav.*, vol. 65, pp. 519–530, Mar. 2003, doi: 10.1006/anbe.2003.2078.

[9] J. Van Canneyt, J. Wouters, and T. Francart, 'Neural tracking of the fundamental frequency of the voice: The effect of voice characteristics', *Eur. J. Neurosci.*, vol. 53, no. 11, pp. 3640–3653, 2021, doi: 10.1111/ejn.15229.

[10] E. S. Teoh, M. S. Cappelloni, and E. C. Lalor, 'Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features', *Eur. J. Neurosci.*, vol. 50, no. 11, pp. 3831–3842, Dec. 2019, doi: 10.1111/ejn.14510.

[11] C. Brodbeck and J. Z. Simon, 'Cortical tracking of voice pitch in the presence of multiple speakers depends on selective attention', *Front. Neurosci.*, vol. 16, Aug. 2022, doi: 10.3389/fnins.2022.828546.

[12] S. López, P. Riera, M. F. Assaneo, M. Eguía, M. Sigman, and M. A. Trevisan, 'Vocal caricatures reveal signatures of speaker identity', *Sci. Rep.*, vol. 3, no. 1, Art. no. 1, Dec. 2013, doi: 10.1038/srep03407.

[13] Y. Lee, P. Keating, and J. Kreiman, 'Acoustic voice variation within and between speakers', *J. Acoust. Soc. Am.*, vol. 146, no. 3, pp. 1568–1579, Sep. 2019, doi: 10.1121/1.5125134.

[14] S. R. Mathias and K. von Kriegstein, 'How do we recognise who is speaking?', *Front. Biosci. Sch. Ed.*, vol. 6, no. 1, pp. 92–109, Jan. 2014, doi: 10.2741/s417.

[15] N. Mesgarani, 'Stimulus Reconstruction from Cortical Responses', in *Encyclopedia of Computational Neuroscience*, D. Jaeger and R. Jung, Eds., Springer New York, 2014, pp. 1–3. doi: 10.1007/978-1-4614-7320-6_108-1.

[16] K. Johnson, 'The ΔF method of vocal tract length normalization for vowels', *Lab. Phonol. J. Assoc. Lab. Phonol.*, vol. 11, no. 1, Art. no. 1, Jul. 2020, doi: 10.5334/labphon.196.

[17] K. Johnson and M. J. Sjerps, 'Speaker Normalization in Speech Perception', in *The Handbook of Speech Perception*, John Wiley & Sons, Ltd, 2021, pp. 145–176. doi: 10.1002/9781119184096.ch6.

[18] C. Pernet, R. Wilcox, and G. Rousselet, 'Robust Correlation Analyses: False Positive and Power Validation Using a New Open Source Matlab Toolbox', *Front. Psychol.*, vol. 3, 2013, Accessed: Nov. 24, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00606

[19] N. Lavan, S. Knight, and C. McGettigan, ‘Listeners form average-based representations of individual voice identities’, *Nat. Commun.*, vol. 10, p. 2404, Jun. 2019, doi: 10.1038/s41467-019-10295-w.

[20] A. C. Lammert and S. S. Narayanan, ‘On Short-Time Estimation of Vocal Tract Length from Formant Frequencies’, *PLoS ONE*, vol. 10, no. 7, p. e0132193, Jul. 2015, doi: 10.1371/journal.pone.0132193.

[21] D. Friedman, ‘Instantaneous-frequency distribution vs. time: An interpretation of the phase structure of speech’, in *ICASSP ’85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr. 1985, pp. 1121–1124. doi: 10.1109/ICASSP.1985.1168461.

[22] C. J. Darwin, D. S. Brungart, and B. D. Simpson, ‘Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers’, *J. Acoust. Soc. Am.*, vol. 114, no. 5, pp. 2913–2922, Oct. 2003, doi: 10.1121/1.1616924.

[23] N. Lavan, A. M. Burton, S. K. Scott, and C. McGettigan, ‘Flexible voices: Identity perception from variable vocal signals’, *Psychon. Bull. Rev.*, vol. 26, no. 1, pp. 90–102, Feb. 2019, doi: 10.3758/s13423-018-1497-7.

[24] D. A. Levy, R. Granot, and S. Bentin, ‘Processing specificity for human voice stimuli: electrophysiological evidence’, *NEUROREPORT*, vol. 12, no. 12, pp. 2653–2657, Aug. 2001, doi: 10.1097/00001756-200108280-00013.

[25] I. Charest *et al.*, ‘Electrophysiological evidence for an early processing of human voices’, *BMC Neurosci.*, vol. 10, p. 127, Oct. 2009, doi: 10.1186/1471-2202-10-127.

[26] A. Capilla, P. Belin, and J. Gross, ‘The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG’, *Cereb. Cortex N. Y. N*, vol. 23, no. 6, pp. 1388–1395, Jun. 2013, doi: 10.1093/cercor/bhs119.

[27] C. Maguinness, C. Roswandowitz, and K. von Kriegstein, ‘Understanding the mechanisms of familiar voice-identity recognition in the human brain’, *Neuropsychologia*, vol. 116, pp. 179–193, Jul. 2018, doi: 10.1016/j.neuropsychologia.2018.03.039.

[28] O. Baumann and P. Belin, ‘Perceptual scaling of voice identity: common dimensions for different vowels and speakers’, *Psychol. Res.-Psychol. Forsch.*, vol. 74, no. 1, pp. 110–120, Jan. 2010, doi: 10.1007/s00426-008-0185-z.

[29] M. Latinus, P. McAleer, P. E. G. Bestelmeyer, and P. Belin, ‘Norm-Based Coding of Voice Identity in Human Auditory Cortex’, *Curr. Biol.*, vol. 23, no. 12, pp. 1075–1080, Jun. 2013, doi: 10.1016/j.cub.2013.04.055.

[30] E. Gaudrain, S. Li, V. S. Ban, and R. D. Patterson, ‘The role of glottal pulse rate and vocal tract length in the perception of speaker identity’, in *INTERSPEECH 2009: 10TH ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION 2009, VOLS 1-5*, Baixas: Isca-Int Speech Communication Assoc, 2009, pp. 152–155. Accessed: Mar. 29, 2024. [Online]. Available: https://www.webofscience.com/wos/woscc/full-record/WOS:000276842800036

[31] D. T. Ives, D. R. R. Smith, and R. D. Patterson, ‘Discrimination of speaker size from syllable phrases’, *J. Acoust. Soc. Am.*, vol. 118, no. 6, pp. 3816–3822, Dec. 2005, doi: 10.1121/1.2118427.

[32] D. R. R. Smith, R. D. Patterson, R. Turner, H. Kawahara, and T. Irino, ‘The processing and perception of size information in speech sounds’, *J. Acoust. Soc. Am.*, vol. 117, no. 1, pp. 305–318, Jan. 2005, doi: 10.1121/1.1828637.

[33] K. Pisanski *et al.*, ‘Vocal indicators of body size in men and women: a meta-analysis’, *Anim. Behav.*, vol. 95, pp. 89–99, Sep. 2014, doi: 10.1016/j.anbehav.2014.06.011.

[34] K. von Kriegstein, J. D. Warren, D. T. Ives, R. D. Patterson, and T. D. Griffiths, ‘Processing the acoustic effect of size in speech sounds’, *NEUROIMAGE*, vol. 32, no. 1, pp. 368–375, Aug. 2006, doi: 10.1016/j.neuroimage.2006.02.045.

[35] K. von Kriegstein, D. R. R. Smith, R. D. Patterson, S. J. Kiebel, and T. D. Griffiths, ‘How the Human Brain Recognizes Speech in the Context of Changing Speakers’, *J. Neurosci.*, vol. 30, no. 2, pp. 629–638, Jan. 2010, doi: 10.1523/JNEUROSCI.2742-09.2010.

[36] K. von Kriegstein, D. R. R. Smith, R. D. Patterson, D. T. Ives, and T. D. Griffiths, ‘Neural representation of auditory size in the human voice and in sounds from other resonant sources’, *Curr. Biol.*, vol. 17, no. 13, pp. 1123–1128, Jul. 2007, doi: 10.1016/j.cub.2007.05.061.

[37] P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, and B. Pike, ‘Voice-selective areas in human auditory cortex’, *Nature*, vol. 403, no. 6767, pp. 309–312, Jan. 2000, doi: 10.1038/35002078.

[38] P. Belin, ‘Similarities in face and voice cerebral processing’, *Vis. Cogn.*, vol. 25, no. 4–6, pp. 658–665, Jul. 2017, doi: 10.1080/13506285.2017.1339156.
[39] P. Belin, C. Bodin, and V. Aglieri, ‘A “voice patch” system in the primate brain for processing vocal information?’, *Hear. Res.*, vol. 366, pp. 65–74, Sep. 2018, doi: 10.1016/j.heares.2018.04.010.
[40] P. Belin, S. Fecteau, and C. Bédard, ‘Thinking the voice: neural correlates of voice perception’, *Trends Cogn. Sci.*, vol. 8, no. 3, pp. 129–135, Mar. 2004, doi: 10.1016/j.tics.2004.01.008.
[41] S. R. Mathias and K. von Kriegstein, ‘Voice Processing and Voice-Identity Recognition’, in *Timbre: Acoustics, Perception, and Cognition*, K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay, Eds., in Springer Handbook of Auditory Research. , Cham: Springer International Publishing, 2019, pp. 175–209. doi: 10.1007/978-3-030-14832-4_7.
[42] C. R. Pernet *et al.*, ‘The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices’, *NEUROIMAGE*, vol. 119, pp. 164–174, Oct. 2015, doi: 10.1016/j.neuroimage.2015.06.050.
[43] M. Latinus, F. Crabbe, and P. Belin, ‘Learning-Induced Changes in the Cerebral Processing of Voice Identity’, *Cereb. CORTEX*, vol. 21, no. 12, pp. 2820–2828, Dec. 2011, doi: 10.1093/cercor/bhr077.
[44] P. Belin and R. J. Zatorre, ‘Adaptation to speaker’s voice in right anterior temporal lobe’, *NEUROREPORT*, vol. 14, no. 16, pp. 2105–2109, Nov. 2003, doi: 10.1097/00001756-200311140-00019.
[45] K. V. Kriegstein and A. L. Giraud, ‘Distinct functional substrates along the right superior temporal sulcus for the processing of voices’, *NEUROIMAGE*, vol. 22, no. 2, pp. 948–955, Jun. 2004, doi: 10.1016/j.neuroimage.2004.02.020.
[46] S. Rosen, R. J. S. Wise, S. Chadha, E.-J. Conway, and S. K. Scott, ‘Hemispheric Asymmetries in Speech Perception: Sense, Nonsense and Modulations’, *PLOS ONE*, vol. 6, no. 9, p. e24672, Sep. 2011, doi: 10.1371/journal.pone.0024672.
[47] C. Brodbeck, A. Jiao, L. E. Hong, and J. Z. Simon, ‘Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers’, *PLOS Biol.*, vol. 18, no. 10, p. e3000883, Oct. 2020, doi: 10.1371/journal.pbio.3000883.
[48] C. McGettigan and S. K. Scott, ‘Cortical asymmetries in speech perception: what’s wrong, what’s right and what’s left?’, *Trends Cogn. Sci.*, vol. 16, no. 5, pp. 269–276, May 2012, doi: 10.1016/j.tics.2012.04.006.
[49] B. Chandrasekaran, A. H. D. Chan, and P. C. M. Wong, ‘Neural Processing of What and Who Information in Speech’, *J. Cogn. Neurosci.*, vol. 23, no. 10, pp. 2690–2700, Oct. 2011, doi: 10.1162/jocn.2011.21631.
[50] T. K. Perrachione, S. N. Del Tufo, and J. D. E. Gabrieli, ‘Human Voice Recognition Depends on Language Ability’, *Science*, vol. 333, no. 6042, pp. 595–595, Jul. 2011, doi: 10.1126/science.1207327.
[51] T. K. Perrachione and P. C. M. Wong, ‘Learning to recognize speakers of a non-native language: implications for the functional organization of human auditory cortex’, *Neuropsychologia*, vol. 45, no. 8, pp. 1899–1910, Apr. 2007, doi: 10.1016/j.neuropsychologia.2006.11.015.
[52] R. Zäske, B. Awwad Shiekh Hasan, and P. Belin, ‘It doesn’t matter what you say: FMRI correlates of voice learning and recognition independent of speech content’, *Cortex J. Devoted Study Nerv. Syst. Behav.*, vol. 94, pp. 100–112, Sep. 2017, doi: 10.1016/j.cortex.2017.06.005.
[53] C. E. Stilp, M. Kiefte, J. M. Alexander, and K. R. Kluender, ‘Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences’, *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2112–2126, Oct. 2010, doi: 10.1121/1.3483719.
[54] K. S. Abouchacra, J. Besing, J. Koehnke, and T. Letowski, ‘The effects of reverberation on a listener’s ability to recognize target sentences in the presence of up to three synchronized masking sentences’, *Int. J. Audiol.*, vol. 50, no. 7, pp. 468–476, Jul. 2011, doi: 10.3109/14992027.2011.565424.
[55] M. Sayles, A. Stasiak, and I. M. Winter, ‘Neural Segregation of Concurrent Speech: Effects of Background Noise and Reverberation on Auditory Scene Analysis in the Ventral Cochlear Nucleus’, in *PHYSIOLOGY, PSYCHOACOUSTICS AND COGNITION IN NORMAL AND*

IMPAIRED HEARING, vol. 894, P. VanDijk, D. Baskent, E. Gaudrain, E. DeKleine, A. Wagner, and C. Lanting, Eds., Berlin: Springer-Verlag Berlin, 2016, pp. 389–397. doi: 10.1007/978-3-319-25474-6_41.

[56] B. Dolan and C. Brockett, 'Automatically Constructing a Corpus of Sentential Paraphrases', presented at the Third International Workshop on Paraphrasing (IWP2005), Jan. 2005. Accessed: Apr. 04, 2024. [Online]. Available: https://www.microsoft.com/en-us/research/publication/automatically-constructing-a-corpus-of-sentential-paraphrases/

[57] P. Mowlaee, R. Saeidi, and Y. Stylianou, 'Advances in phase-aware signal processing in speech communication', *SPEECH Commun.*, vol. 81, pp. 1–29, Jul. 2016, doi: 10.1016/j.specom.2016.04.002.

[58] S. O. Sadjadi and J. H. L. Hansen, 'Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification', *SPEECH Commun.*, vol. 72, pp. 138–148, Sep. 2015, doi: 10.1016/j.specom.2015.04.005.

[59] A. Potamianos and P. Maragos, 'Speech formant frequency and bandwidth tracking using multiband energy demodulation', *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3795–3806, Jun. 1996, doi: 10.1121/1.414997.

[60] Z. Průša, P. L. Søndergaard, N. Holighaus, C. Wiesmeyr, and P. Balazs, 'The Large Time-Frequency Analysis Toolbox 2.0', in *Sound, Music, and Motion: 10th International Symposium, CMMR 2013, Marseille, France, October 15-18, 2013. Revised Selected Papers*, Berlin, Heidelberg: Springer-Verlag, Dec. 2014, pp. 419–442. doi: 10.1007/978-3-319-12976-1_25.

[61] P. L. Søndergaard, B. Torrésani, and P. Balazs, 'The linear time frequency analysis toolbox', *Int. J. Wavelets Multiresolution Inf. Process.*, vol. 10, no. 04, p. 1250032, May 2012, doi: 10.1142/S0219691312500324.

[62] B. R. Glasberg and B. C. J. Moore, 'Derivation of auditory filter shapes from notched-noise data', *Hear. Res.*, vol. 47, no. 1–2, pp. 103–138, Aug. 1990, doi: 10.1016/0378-5955(90)90170-T.

[63] M. Grimaldi and F. Cummins, 'Speaker Identification Using Instantaneous Frequencies', *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 6, pp. 1097–1111, Aug. 2008, doi: 10.1109/TASL.2008.2001109.

[64] J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart, 'Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope', *J. Assoc. Res. Otolaryngol.*, vol. 19, no. 2, pp. 181–191, Apr. 2018, doi: 10.1007/s10162-018-0654-z.

[65] M. Junghöfer, T. Elbert, D. M. Tucker, and B. Rockstroh, 'Statistical control of artifacts in dense array EEG/MEG studies', *Psychophysiology*, vol. 37, no. 04, pp. 523–532, Jul. 2000, doi: 10.1111/1469-8986.3740523.

[66] A. Hyvarinen, 'Fast and robust fixed-point algorithms for independent component analysis', *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999, doi: 10.1109/72.761722.

[67] A. de Cheveigné and J. Z. Simon, 'Denoising based on time-shift PCA', *J. Neurosci. Methods*, vol. 165, no. 2, pp. 297–305, Sep. 2007, doi: 10.1016/j.jneumeth.2007.06.003.

[68] A. de Cheveigné and J. Z. Simon, 'Sensor noise suppression', *J. Neurosci. Methods*, vol. 168, no. 1, pp. 195–202, Feb. 2008, doi: 10.1016/j.jneumeth.2007.09.012.

[69] S. V. David, N. Mesgarani, and S. A. Shamma, 'Estimating sparse spectro-temporal receptive fields with natural stimuli', *Netw. Comput. Neural Syst.*, vol. 18, no. 3, pp. 191–212, Jan. 2007, doi: 10.1080/09548980701609235.

[70] K. C. Puvvada and J. Z. Simon, 'Cortical Representations of Speech in a Multitalker Auditory Scene', *J. Neurosci.*, vol. 37, no. 38, pp. 9189–9196, Sep. 2017, doi: 10.1523/JNEUROSCI.0938-17.2017.

[71] D. Schwarzkopf, B. de Haas, and G. Rees, 'Better Ways to Improve Standards in Brain-Behavior Correlation Analysis', *Front. Hum. Neurosci.*, vol. 6, 2012, Accessed: Nov. 24, 2023, doi: 10.3389/fnhum.2012.00200.

Figures

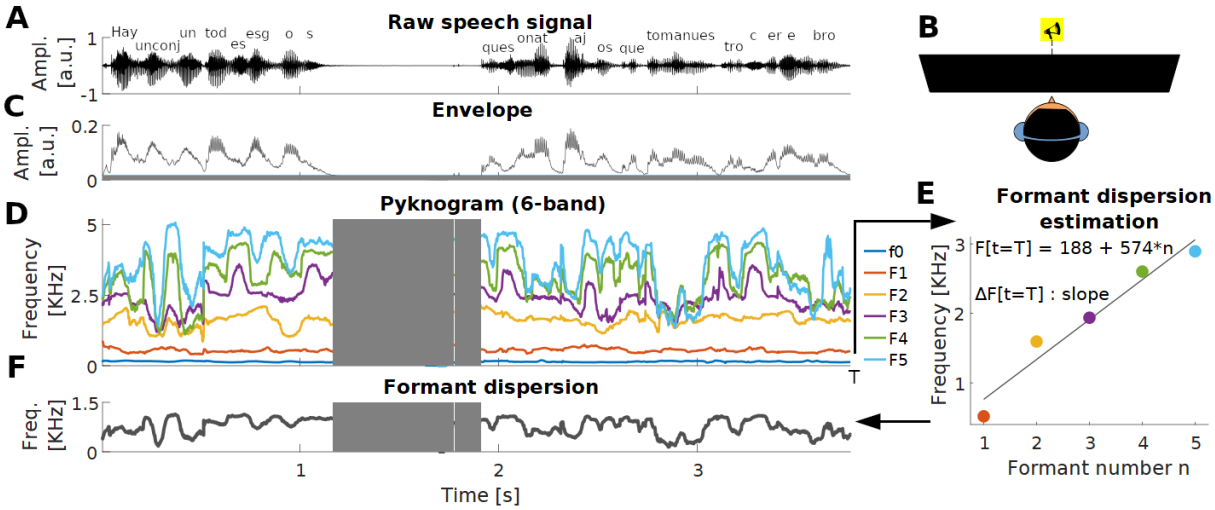


Figure 1. ***Setup for extraction of spectral features of voiced speech***. **(A)** Example of a single-speaker speech phrase as presented in a trial. **(B)** Participants listen diotically to centrally presented speech while undergoing EEG recordings. **(C)** Envelope of the speech example at the original sampling rate. **(D)** Short-time estimates (mean amplitude-weighted) of instantaneous frequency from the speech example correspond to six spectral bands representing the fundamental frequency f_0 and formants F_1-F_5. Silent regions (grey areas) are not further analyzed. **(E)** Method of formant dispersion dynamics ΔF estimation, based on sample-by-sample linear regression of formant frequencies. For this example, ΔF at time T is 574 Hz.

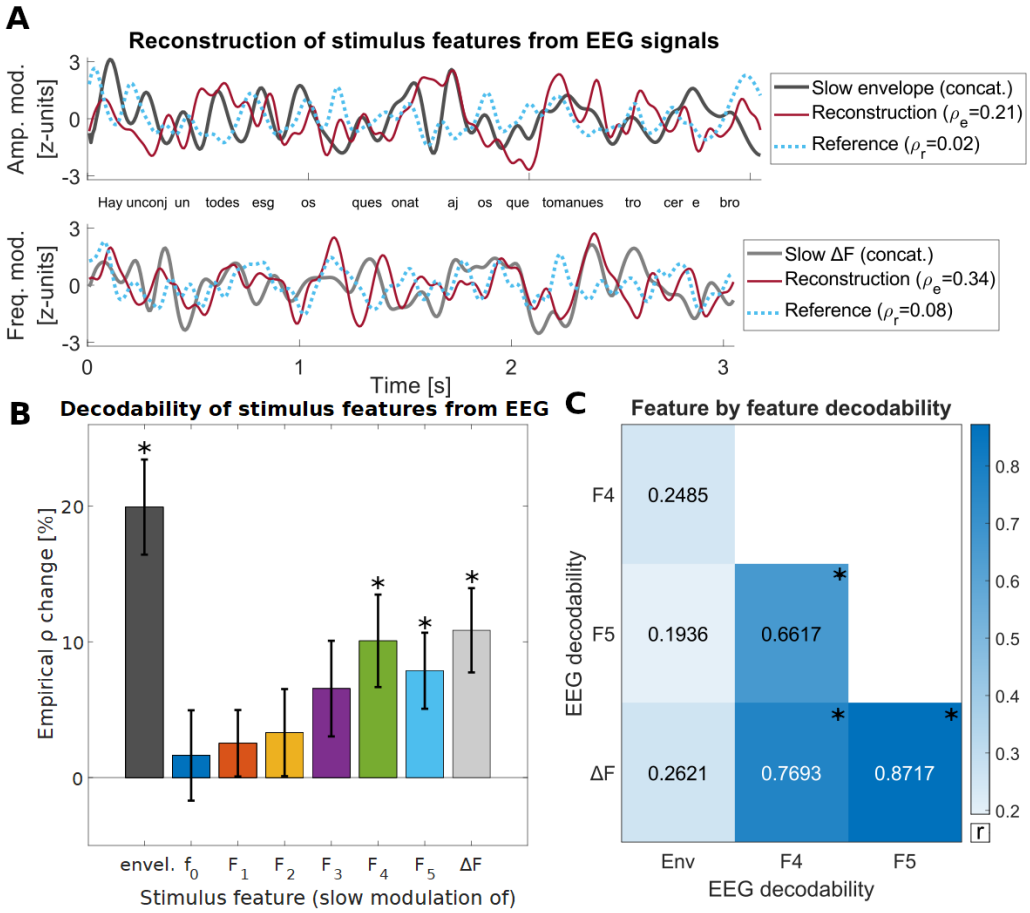


Figure 2. ***Spectral features of voiced speech are reliably decoded from single-trial EEG dynamics***. **(A)** Sample reconstruction of speech feature waveforms from the example in Figure 1. Decoders estimated per listener (N = 19 for this stimulus) are applied to corresponding subject datasets. *Top*: average of resulting envelope reconstructions. Reconstruction performance is indicated by the rank correlation between the original and mean reconstructed stimuli. Here, performance exceeds that of decoders estimated after reshuffled EEG data as reference. *Bottom*: The method can be successfully applied to spectral vocal features such as formant dispersion dynamics (ΔF). **(B)** Modulations of envelope, higher formants, and ΔF are reliably decoded from single-trial EEG. **(C)** Unlike the dynamics of the envelope, spectral features are co-modulated to the extent that subject decodability across these dimensions is mutually predictable.

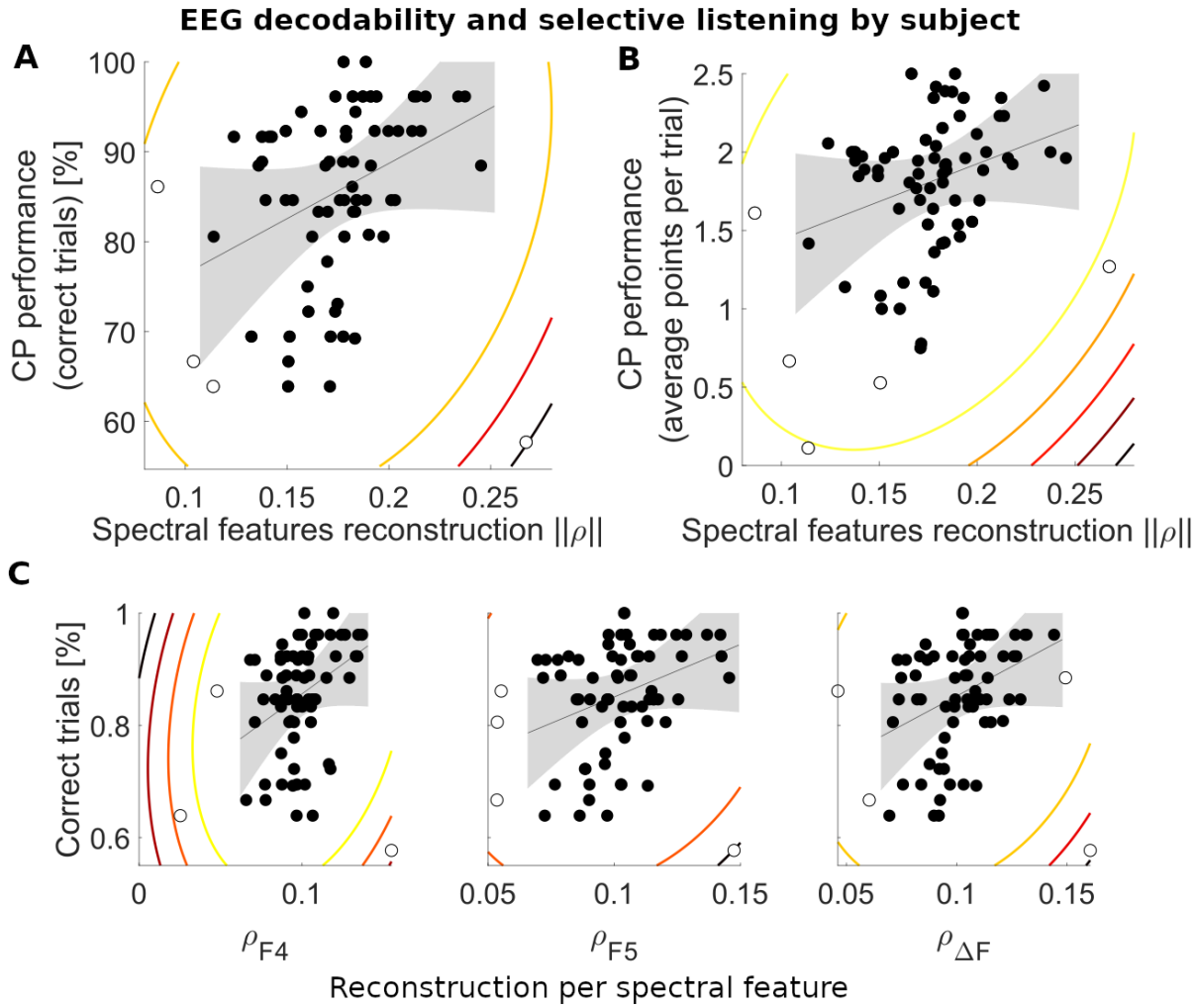


Figure 3. ***Spectral features of voiced speech assist performance in multispeaker conditions***. **(A)** A summary measure of neural decodability across the three relevant spectral features significantly predicts the behavioral performance of listeners at the 'cocktail party'. **(B)** Joint decodability of spectral contours also predicts subject performance measured by mean points per trial at the cocktail party task, with a significant robust correlation relationship observed, Shepherd's Pi = 0.32, p = 0.015. **(C)** Decodabilities of higher formants (F_4, F_5) and formant dispersion (ΔF) dynamics individually predict performance at the cocktail party task. Robust correlation relationships are observed for F_4 (left), Shepherd's Pi = 0.39, p = 0.002; F_5 (middle), Shepherd's Pi = 0.35, p = 0.007; and ΔF (right), Shepherd's Pi = 0.37, p = 0.004.

Supplemental material for

Cortical tracking of speakers' formant changes predicts selective listening

Francisco Cervantes Constantino and Ángel A. Caputi

Supplementary Materials and Methods

General overview

This article is part of a more exhaustive series of studies on the cocktail party effect (Sánchez-Costa et al., results to be published elsewhere) that aims to investigate how previous listening experiences modulate multi-speaker segregation. Here, we applied a stimulus reconstruction technique from the EEG of subjects recorded while listening to single-speaker speech presentations (yellow highlight, Supplementary Figure 1). Neural estimates obtained from these data were compared with behavioral data addressing performance at the cocktail party by the same subjects (green highlight, Supplementary Figure 1). See *Task* in the main text and below for full details.

Setup

Subjects were seated 50 cm in front of a 40 cm monitor (E. Systems, Inc., CA). Audio was presented diotically with a Sound Blaster Z sound card (Creative Labs, Singapore), a Scarlett 414 sound interface (Focusrite Plc, UK) and high-quality Sennheiser HD 25 headphones (Sennheiser, Germany). EEG recordings were made with a BioSemi ActiveTwo 64-channel system (BioSemi, The Netherlands) sampled at 1024 Hz; audio was re-registered with the EEG system with a ActiveTwo ERGO Opticlink (Biosemi) at the same rate. The participants adjusted sound volume to their comfortable listening level before starting the task.

Stimuli

Volunteer recordings were made with an Olympus recorder (VN-541OC Olympus, Vietnam) in open spaces. All stimuli were marked for speaker/voice apparent sex and age (i.e., below or above 35 years-old). Sentence topic was designated from a list of 12 topical themes, with a sentence possibly corresponding to more than one topic. Audio recordings were processed with MATLAB® (MathWorks, United States) at a sampling rate of 44100 Hz. Cosine ramps (5 ms duration) were applied at the beginning or end of the stimulus, and surrounding manual excisions made in cases of long silences and repeated filler words.

For the cocktail party task, stimulus mixtures were created by selecting two stimuli from the database and adding them together. A unique set of cocktail party stimuli was created for each participant, involving same- and different-sex combinations in equal proportions. Each cocktail party stimulus had an initial 750 ms onset asynchrony between speaker streams, so one speaker was presented initially while the other was interjected later. The end of the stimulus equalled to the resulting overlap time. Stimuli were equalized in root mean squared (RMS) intensity for the overlapping epoch. All stimuli were individually normalized.

Task

In the present study we only addressed behavioral data from all control trials in the original study series. Unlike other conditions, the control trials consist of initial solo-listening presentations that are independent from stimuli present in the mixture, i.e., involve different speakers, messages, and sentences. The number of control trials ranged from 26 to 36 in the studies.

Questionnaire and scoring

Trial questionnaires contained one option that was related to the speaker's identity plus eight options that were related to sentence content. Participants were asked to select three best-match options on the indicated target. With equal probability, the single identity option prompted participants on vocal qualities: either speaker's sex (masculine/feminine) or age (younger/older). Four additional content-related options were based on specific keywords pronounced by either or both talkers in the mix; these options were randomly sampled from the sum of two 4-keyword lists corresponding to the sentences. These keywords are sentence nouns, adverbs, or verbs, collected from the database. Finally, four more content options prompted participants to report on the topic of the attended speaker. The options were randomly selected from the database 12-theme list; sentences may or may not correspond to the options shown. All questionnaire options were shown in randomized order.

For scoring, any correctly chosen option always earned one point. Incorrectly chosen options deducted points as follows. For identity options, selecting the opposite gender or opposite age of the attended speaker resulted in a point deduction. For keyword options, selecting a keyword related to the masker speech resulted in a point deduction. For topic options, missed choices or choices related to the masker speech did not result in point deductions. Trial scores equaled the sum of point earnings and deductions across all options validated by the participant. Validating no responses in a trial did not gain any points.A trial was deemed correct if its score was greater than zero.

Speech spectral feature analyses

To discount the spurious spectral contributions of silent periods and weak transients in natural speech, all periods that corresponded to the signal's envelope below a certain threshold were automatically discarded. The threshold was set per speech sample at the bottom quintile of the full envelope histogram for that speech signal. Removal was performed after temporal alignment with the EEG data (see *Stimulus-response alignment and decoder estimation*, below). To obtain the full envelope, the raw speech signal was submitted to a bank of 28 gammatone filters with ERB-spaced centre frequencies corresponding to the 52 – 4956 Hz range. From each sub-band, the envelope was extracted by taking the magnitude value of each sample and raising it to the power of 0.6 [64]. The resulting sub-band envelopes were averaged, and this total envelope was resampled to 1024 Hz. Below-threshold epochs were jointly discarded in corresponding periods of the envelope and the seven spectral features.

EEG preprocessing

After preprocessing, if a head channel resulted in broadband power above 3 standard deviations relative to all other head channels across the experiment session, it was removed. Single-trial presentations of solo speech stimuli were epoched. The variance-based rejection procedure was repeated now across single channels and epochs' timeseries per participant, resulting in < 1% rejected channel-epoch timeseries on average (subject range 0.09%–1.52%).

Stimulus-response alignment and decoder estimation

Male speaker voice presentations were selected from each participant's EEG data, resulting in 46 trials per participant on average (range: 31 – 62 trials). Any time sample, in a given head channel and trial,

whose recorded absolute value exceeded 7 standard deviations of all values in the rest of the data, was set as an empty value.

EEG data were temporally aligned with the acoustic feature timeseries sets (including the envelope, f_0, F_1-F_5, and ΔF) as follows. Delay corrections were performed via cross-correlation between the audio signal sampled by the EEG system versus the corresponding raw audio original file downsampled to 1024 Hz. EEG and acoustic feature waveforms were then aligned, after which all data time periods when the speech envelope was below a level criterion (see *Speech spectral feature analyses*, above) were removed. All co-aligned EEG and stimuli epoch trials were concatenated in time. All signals were 1–8 Hz bandpass filtered with an order 2 Butterworth filter, with group delay correction, and resampled to 128 Hz. All timeseries were transformed to z-score units along the temporal dimension.

The robustness of a subject's neural representation for the *i*-th feature was estimated by reconstruction accuracy, i.e., the distance between $S_i(t)$ and $\widetilde{S}_i(t)$ measured as the strength of the Spearman rank correlation coefficient $\rho_{i,e}$ between both timeseries. If the decoder was estimated using reshuffled EEG trial data instead, the reference correlation $\rho_{i,r}$ was similarly obtained. To compare across auditory features, the decodability of the *i*-th feature from EEG was expressed in relative terms as the empirical percent change in ρ_i, i.e. $(\rho_{i,e} - \rho_{i,r}) / \rho_{i,r}$.

Statistical analysis of decoder data

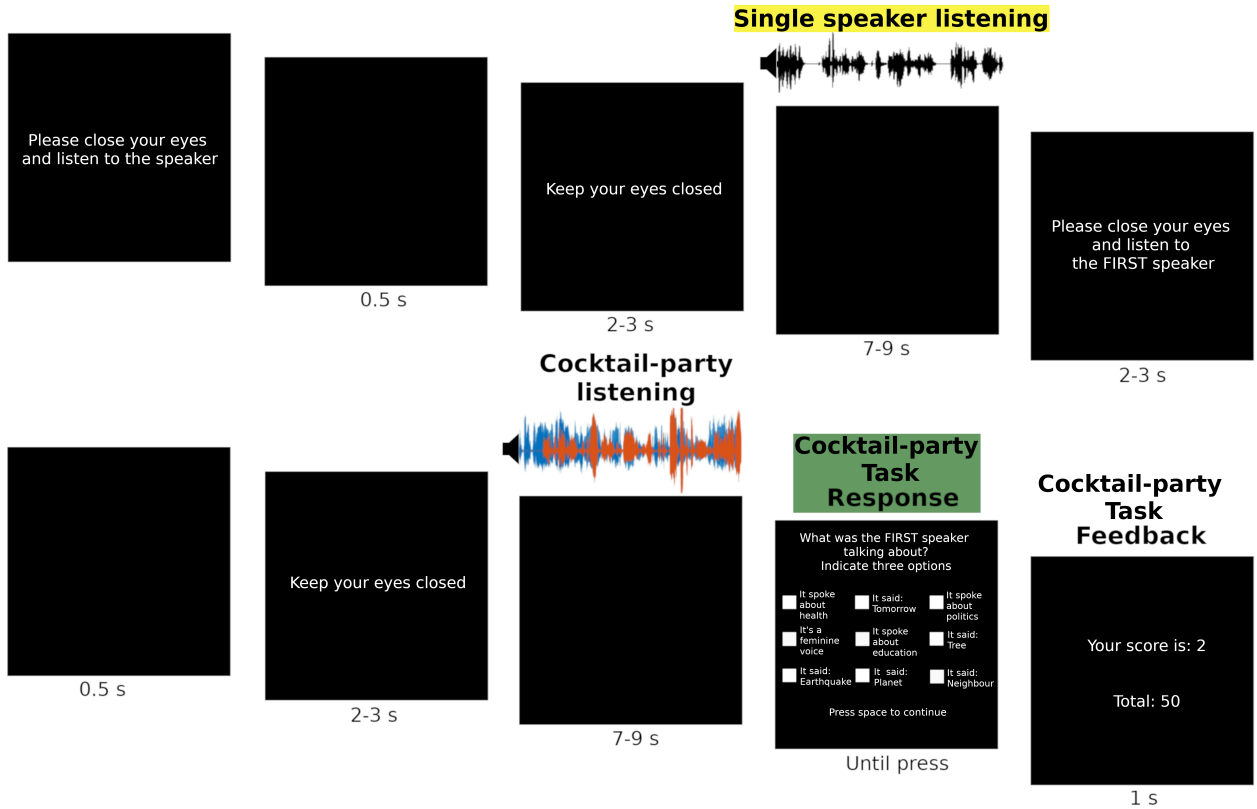
The null hypothesis that the empirical percent change is not positive states that slow modulations in the *i*-th auditory feature are not reliably decoded from slow modulations in the EEG data. To test the alternative hypothesis, a one-sided Student's *t*-tests was run for each auditory feature, evaluating whether each feature modulation is decodable. To correct for multiple testing, *p*-values were adjusted by the Holm-Bonferroni method.

Supplementary Text

Successful formant reconstructions predict performance at the cocktail party

With regards to cocktail-party (but not single-speaker listening) data involving voices of either sex, we verified that neural predictions of subject correct rates indeed remained for data where participants targeted a male speaker specifically, as confirmed by robust regression $t(70) = 2.913$; $p = 0.004$. The relationship similarly held for the subset of behavioral data involving male-male speaker voices' combinations only, $t(70) = 3.192$; $p = 0.002$.

Supplementary Figures



Supplementary Figure 1. Current experimental approach in the context of a wider cocktail party effect study. In each trial, participants hear the solo speech example (yellow highlight) and are presented later with an unrelated two-speaker mixture to select and attend to one target speaker. After listening, a questionnaire prompts for different aspects of the attended speech, including sex/age, topic, and specific keywords (green highlight); feedback is shown upon validation. Trial performance is measured by a point scoring system. Modified from Sánchez-Costa, et al.