Cortical tracking of speakers' spectral changes predicts selective listening

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

[1]Instituto de Investigaciones Biológicas Clemente Estable;

[2]Facultad de Psicología, Universidad de la República

*For correspondence: fcervantes@psico.edu.uy

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

Keywords: frequency modulation | phase locking | selective attention | speech

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 frequency modulations in speech 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 frequency contours in the fourth and fifth formant band, equivalent to the 3.5–5 KHz audible range. Instantaneous frequency 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.

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; for instance, observers may infer the likely size of an unseen

conspecific given the average spacing between the caller's time-varying formants (Fitch 1997; Riede and Fitch 1999; Collins 2000; Reby and McComb 2003; Bruckert et al. 2006; Evans et al. 2006; Puts et al. 2007; Puts et al. 2011). The 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.

What spectral contour dynamics are represented through tracking activity by the auditory cortex or higher associated areas? Cortical tracking is the phenomenon describing the ability of network activity in the cerebral cortex to lock to the structure time-varying stimulus features as in speech, in which case it often refers to following its slow envelope. Such ability has in some cases been shown to correlate with listeners' behavioral intelligibility levels (Vanthornhout et al. 2018). Analysis of cortical tracking while applying transcranial alternating current stimulation has further demonstrated the causal influence of phase-locked responses on intelligible speech perception (Wilsch et al. 2018; Zoefel et al. 2020).

Speech signals that are being attended to amid ambient noise elicit selective tracking responses, whereby the slow dynamics of cortical activity appears more consistent with the target speech envelope than the background (Ding & Simon 2012; Zion Golumbic et al. 2013; O'Sullivan et al. 2015). In scenes, selective tracking may manifest the existence of a cleaned-up representation of the target speech, where the accuracy of such representation possibly determines comprehension (Brodbeck and Simon 2020). There are now indications that cortical tracking, despite showing dominant responses to the speech signal's envelope, also extends to some of the spectral cues of speech needed to establish speaker identity (Tang et al. 2017; Teoh et al. 2019; Van Canneyt et al. 2021; Brodbeck and Simon 2022). These studies have shown that the auditory cortex may track the dynamic contours and changes of voiced pitch, which offers opportunities to evaluate and generalize the association between cortical tracking and speech intelligibility. Much of speech intelligibility research has traditionally required systematic manipulation on spectral uncertainty, such as in noise-vocoded speech. In this regard, it is important to note that when the level of spectral detail is degraded to the point of unintelligibility, cortical tracking may be affected in spite of a preserved temporal envelope (Peelle et al. 2013). This underscores the need the contribution to tracking from the process resolving spectral contours in detail, as it offers a potential to connect with research demonstrate the tracking of higher-level linguistic units (Di Liberto et al. 2015).

Moreover, building on this area may offer an attractive way to probe the relationship of tracking frequency modulations generated by vocal tract resonances with behavior in the face of multiple voiced maskers. It specifically involves the question of what speaker-specific spectral features activate speech-tracking patterns in the brain that are relevant for speaker-selective behavior. Evidence of cortical encoding of frequency modulations (FM) in the voice hence represents a promising step to unravel how the brain targets "who" speaks under such conditions (Brodbeck and Simon 2022). A natural limitation in studies of spectral contour tracking is of course that cortical auditory population responses weakly resolve frequency-following responses even at pitch range (85-200 Hz). Yet, it is clear that representing spectral content at the higher frequency formant bands matters for cortical systems that establish speaker identity and phonemic analyses (López et al. 2013; Mathias and von Kriegstein 2014; Lee et al. 2019). 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 additional slow changes of instantaneous FM in the formant range. To evaluate this possibility, we employed a

stimulus reconstruction approach (Mesgarani 2014; Van Canneyt et al. 2021) designed to estimate the accuracy of recovering the variations of a given feature, such as instantaneous frequency, from a linear mapping of a listener's electroencephalography (EEG) dataset. As with forward-model approaches (Teoh et al. 2019; Van Canneyt et al. 2021; Brodbeck and Simon 2022), decoding methods are evaluated in terms of their generalization performance. Due to scalp-wide, fine-grained multivariate information integration, backward modeling is a reliable means to test generalization performance. Performance from decoding models of envelope-tracking responses has been systematically used to predict behavioral responses in speech processing as objective measures of speech intelligibility (Vanthornhout et al., 2018). Importantly, when decoders show that specific stimulus information contributes to successful behavior, they effectively model the computational process that generates the observer's expected response from coded stimulus features (Kriegeskorte and Douglas 2019).

In the present study, we addressed cortical tracking of spectral features in speech via reconstruction methods, at different bands across the formant range. Because of its potential basis for grouping features by speaker, we further approached the spectral feature which describes the instantaneous spacing pattern between formants arising as a result of their dynamic resonant structure. The average inter-formant spread has sometimes been used as proxy estimator for vocal tract length (VTL), a relevant parameter in automatic speaker recognition and vowel normalization among multiple talkers (Johnson 2020; Johnson and Sjerps 2021). Consequently, accurate decoding of a family of spectral signatures from the EEG of listeners, quantified through a single decodability score, may reasonably indicate their individual ability to follow dynamics stemming from changes in the apparent VTL of an unfamiliar talker as they speak. If reliable spectral contour reconstruction is quantified, such parameter may be furthermore expected to predict voice-specific selective listening ability of different listeners. We evaluated both hypotheses by, first, measuring the ability to faithfully represent spectral changes of different talkers in a single-speaker task through decodability scores. These were then compared with the behavioral accuracy of the same subjects probed on a comprehension report on aspects of target speakers' identities, words and topics, as part of a selective listening task in a 'cocktail-party' scenario.

Materials and Methods

Participants

Seventy-three subjects (51 female; mean age 25.6 ± 4.9 SD) participated in the studies. Participants were native Rioplatense Spanish speakers with no hearing impairment or neurological disorder, and correct vision. 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 original databases containing short phrase audios (mean duration 8.6 s, ±0.8 s) individually narrated by 607 speakers including 264 different male voices. The stimuli were sourced from Uruguayan news broadcasts, audiobooks, interviews, and recordings produced from Uruguayan volunteers. Recorded volunteer audios were read after phrases translated from a public database (Dolan and Brockett, 2005).

Task

As part of the EEG study, participants were instructed to listen to each sound presentation with their eyes closed. In each trial, participants were first presented with a single-speaker speech stimulus ("single speaker listening", SSL phase), followed by a cocktail party mixture stimulus in the second part (Supplementary Figure 1). Listeners were tasked with selectively attending to one of the two speech streams in the mixture, i.e., to select the leading or the interjecting talker. The indicated target stream was instructed randomly with equal probability. At the end of the mixture, participants were presented with a nine-option questionnaire. They were required to respond by exclusively selecting three options or more corresponding to the indicated target. Subsequent to response validation, a feedback screen was displayed with the trial and the accumulated scores (Supplementary Materials and Methods).

Speech spectral feature analyses

Single-speaker speech stimuli from the SSL trial phase were submitted to an AM-FM analysis (Potamianos and Maragos 1996; Sadjadi and Hansen 2015; Mowlaee et al. 2016) in the original sampling rate. To optimize signal-to-noise ratio, the analysis was limited to male voice presentations as higher formants contain substantially less energy in female voices (Baumann and Belin 2010). Moreover, evidence of f_0 tracking in the EEG to female voices is weak as decoder performance may be inversely proportional to frequency (Van Canneyt et al. 2021). A fixed 6-band gammatone filterbank was hence applied to each audio sample, assessing spectral variations corresponding to male f_0 and formants' F_1-F_5 frequency bands. Filterbank coefficients were based on six centre frequencies fc_i set at 146, 671, 1870, 2870, 3946, and 4869 Hz that were defined from formant analysis of a local public voice database (López et al. 2013) conducted with the Praat software. Coefficients were estimated with the Large Time/Frequency Analysis toolbox (Søndergaard et al. 2012; Průša et al. 2014). Channel bandwidths equalled thrice the critical bandwidth of the auditory filter at the corresponding fc_i, and were defined in equivalent rectangular bandwidths (ERBs) (Glasberg and Moore 1990). The real part of the filter coefficients was applied to each audio signal. Each filter output was analyzed to obtain its instantaneous frequency $f_i(t)$, which was estimated by differentiating the instantaneous unwrapped phase $\phi_i(t)$ of the corresponding analytic signal, computed by the Hilbert method. Expressed in polar coordinates, the analytic signal conveniently separates from the effects of instantaneous amplitude content $a_i(t)$, which is extracted in envelope tracking for instance. A multiband demodulation analysis was performed by mean-amplitude weighted short-time estimation of the instantaneous frequency per sub-band (Grimaldi and Cummins 2008). At time instant t_0, the short-time estimate for the spectral modulations in the frequency F_i was obtained by integration of the weighing instantaneous amplitude function over a τ=25 ms window, according to the formula:

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

In effect, spectral modulations are separated from amplitude modulations within each frequency band by rendering the band's estimated instantaneous frequency first, and smoothing it over short windows, weighing each frequency by the relative power of the filtered signal at that specific instant. The resulting set of discrete short-time weighted instantaneous frequency estimates $F_i(t)$ is sometimes referred to as the pyknogram of the speech signal (Grimaldi and Cummins 2008). It may be used as a basis for speech formant estimation, alternatively to linear predictive coding methods for instance (Potamianos and Maragos 1996). Here, the pyknogram was used to represent the dynamic waveform timeseries for instantaneous frequencies within the respective ranges of f_0 and formants' F_1-F_5.

The dynamic waveform for the instantaneous frequency spacing 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 time instant t_k represented by the k-th sample, the model of the spectral observation F_i is:

$$F_i\left[t=t_k\right]=\beta_{0,k}+\beta_{1,k}n_i+\varepsilon_k \tag{2}$$

with β_1 representing the linear slope of the regression model of F_i values against their associated formant band numbers n_i={1,2,3,4,5} (the model predictors), and β_0 the intercept; ε represents the linear model error. 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 spectral spacing (Johnson 2020):

$$\Delta F\left[t=t_k\right]=\beta_{1,k} \tag{3}$$

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 by a channel variance-based criterion (Junghöfer et al. 2000) blind procedure, implemented with confidence coefficient $\lambda_P=4$. The 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 a second time on the full set. EEG data were decomposed into independent components using FastICA (Hyvarinen 1999) and the two independent components with maximal broadband power were automatically selected and projected out of the data. A time-shifted principal component analysis (de Cheveigné and Simon 2007) 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 (de Cheveigné and Simon 2008).

EEG data time samples with a value greater than 7 standard deviations were set to zero after which single trial neural data were aligned with stimuli data, and trials were concatenated in time. The set of EEG and stimulus signals was then band-pass filtered between 1 and 8 Hz

with a second-order Butterworth filter in the forward and reverse direction, downsampled to 128 Hz, and converted to standardized units.

Stimulus-response alignment and decoder estimation

Using EEG responses from single speaker listening phases, we assessed performance in reconstructing the slow modulations of the instantaneous frequency waveform of speech at (*i*) the f_0 band, (*ii-vi*) the first to fifth formant bands ($F1$ to $F5$), (*vii*) the spectral spacing parameter ΔF; and (*viii*) the instantaneous waveform of speech broadband envelope (see *Supplementary material*); resulting in eight separate decoders based on the empirical EEG activity, per listener. Eight additional decoders were separately estimated, based on trial-reshuffled EEG activity, so as to compare each individual empirical performance measure to an estimation noise reference.

The operation to obtain the linear reconstruction $\tilde{S}_i(t)$ of a univariate stimulus timeseries $S_i(t)$ from a set of N EEG channel responses $R_n(t)$ can be formulated as:

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

where $D_{i,n}(t)$ is the i-th linear decoder that operates on the response timeseries $R_n(t)$ from channel n in order to reconstruct the i-th acoustic feature. Each individual decoder D_i is a $N \times T'$ matrix with $N = 64$ channels and $T' = 400$ ms denoting the size of the temporal integration window. Hence, at time point $t=t_k$, the stimulus timeseries is reconstructed by linear combination of all channel responses spanning from t_k until t_k+0.4 s. Decoders were obtained using the boosting method with 10-fold cross-validation (David et al. 2007; Puvvada and Simon 2017). The resulting stimulus reconstruction approximates the true stimulus by a residual response waveform $\epsilon_i(t)$ not explained by the linear decoder model:

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

Empirical decodability of the i-th feature, $\rho_{i,e}$, is measured by the Spearman rank correlation between $S_i(t)$ and $\tilde{S}_i(t)$. For each feature we additionally obtained its reference decodability $\rho_{i,r}$, defined as the Spearman rank correlation between $S_i(t)$ and the linear reconstruction that results from training a decoder D_i* based on trial-reshuffled EEG data R_n*(t). Decoder reliability (also referred to as neural decoding performance here), was expressed as $\left(\rho_{i,e}-\rho_{i,r}\right)/\rho_{i,r}$ *100%, i.e., the percent relative change of empirical decodability with respect to the performance of the corresponding reference decoder.

Decoder performance as predictor of listening performance at the cocktail party

Upon identification that the reliably decodable spectral features were the modulation from bands F_4, F_5, and the ΔF measure (see Results), their common empirical decodabilities were indexed by the Euclidean norm, per participant:

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

A participant's high score in this more general spectral feature reconstruction parameter represents systematically higher decodability of all three significant spectral features. $\|\rho\|$

was used to evaluate the hypothesis that cortical tracking of vocal tract features from speech predicts listener performance upon facing cocktail-party scenarios. The proportion of correct trials in the cocktail party task was taken as a measure of participant behavioral performance. The association between subjects' $||\rho||$ and cocktail-party performance was evaluated with Spearman's rank correlation r_s with interval testing (Pernet et al. 2013), and confirmatory tests were subsequently performed via Shepherd's correlation (Schwarzkopf et al. 2012) and robust regression with the Huber weighting function.

Results

The ability to decode seven acoustic features that relate frequency modulations in voiced speech, in addition to the speech envelope, was investigated from EEG recordings using stimulus reconstruction methods. The spectral features are delta-theta band modulations of instantaneous frequency modulations in (*i*) the fundamental frequency band, (*ii-vi*) the F_1, F_2, F_3, F_4, F_5 formant bands, and (*vii*) their interband spacing (Figure 1). Decoder analyses were based on single-trial, single-speaker male voice listening epochs, as in the EEG these offer relatively higher precision frequency tracking data than female speaker listening [9]. Figure 2A depicts an example of stimulus reconstruction applied to delta-theta band modulations of the speech envelope from one sentence stimulus. Here, the original envelope waveform is reconstructed from a decoder based on empirical EEG data to a degree that, in relative terms, exceeds what a second, independent decoder may achieve using reshuffled EEG data as reference. In a similar way, the same EEG data additionally contain tracking responses to the dynamics of instantaneous interband frequency spacing ΔF, which denotes the time-varying dispersion of spectral resonances in a speaker's voice.

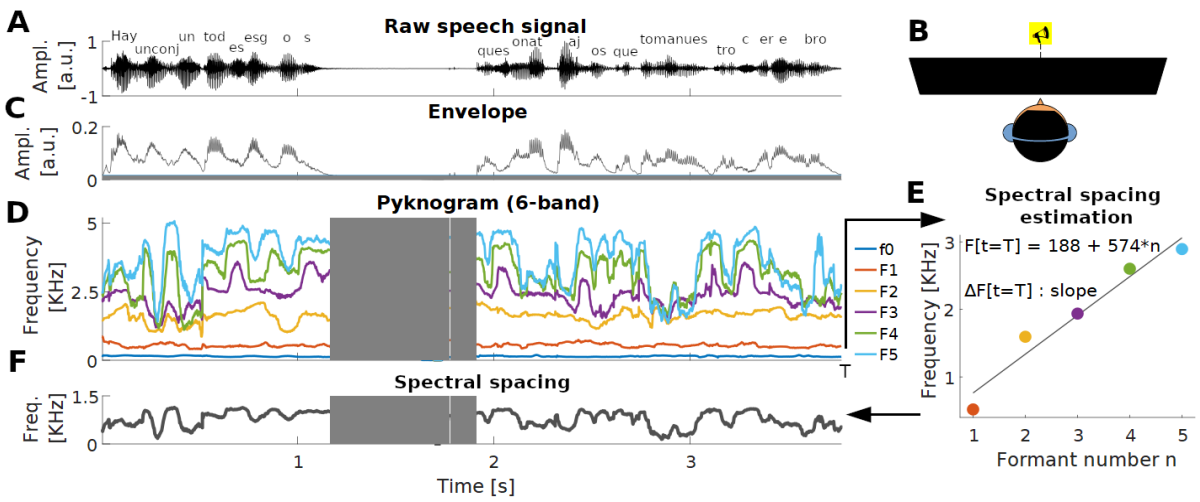


Figure 1. ***Setup for extraction of spectral features from voiced speech*. (A)** Trial of a single-speaker speech sentence. **(B)** Participants listen diotically to centrally presented speech trials while undergoing EEG. **(C)** Broadband envelope of the speech sentence. **(D)** Instantaneous frequency estimates from the sentence. Each estimate corresponds to a spectral band centred on the fundamental frequency f_0 or F_1-F_5 formant bands. Silent regions (grey areas) are not analyzed. **(E)** Method of ΔF estimation, based on linear regression of instantaneous frequency values indexed by the number of the associated formant band. For this example, ΔF at time sample T is 574 Hz.

Robust decodability of vocal spectral dynamics from single-trial EEG recordings

Neural decoding performance, quantified by the empirical-to-reference reconstruction performance ratio, referred to as empirical ρ change, was estimated per acoustic feature (Figure 2B). The slow (i.e., delta-theta) modulations of the broadband envelope were, as expected, robustly decoded from the single-trial EEG (mean empirical ρ change 19.9%, SD = 29.9%) $t(72) = 5.7$, adjusted (Bonferroni-Holm) $p = 10^{-5}$. In addition, the analyses showed that slow modulations of the instantaneous frequency waveforms from higher formant ranges were also reliably decodable: 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. $p = 0.016$. Furthermore, robust decodability was also found in EEG responses to the dynamics of instantaneous interband frequency spacing ΔF (mean 10.8%, SD = 26.5%) $t(72) = 3.5$, adj. $p = 0.003$. Supplementary Figure 2 presents the raw (empirical) reconstruction performance in terms of Spearman rank correlations.

At the lower frequency ranges, decodability of slow modulations of the instantaneous frequency waveform from the fundamental f_0 band was not found to be robust (mean 1.6%, SD = 28.4%) $t(72) = 0.5$, adj. $p > 0.31$. Null results were similarly observed for decodabilities of spectral features from: 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. $p > 0.15$; or F_3 bands (mean 6.6%, SD=30.1%) $t(72) = 1.9$, adj. $p = 0.13$. Hence, slow modulations in the spectral variations across the lower frequency bands of voiced speech were not reliably decoded from single-trial EEG data. The results indicate that the dynamics of spectral contours associated with the higher formant ranges, and spacing information used to index vocal identity, may be however tracked by cortical populations during natural speech listening.

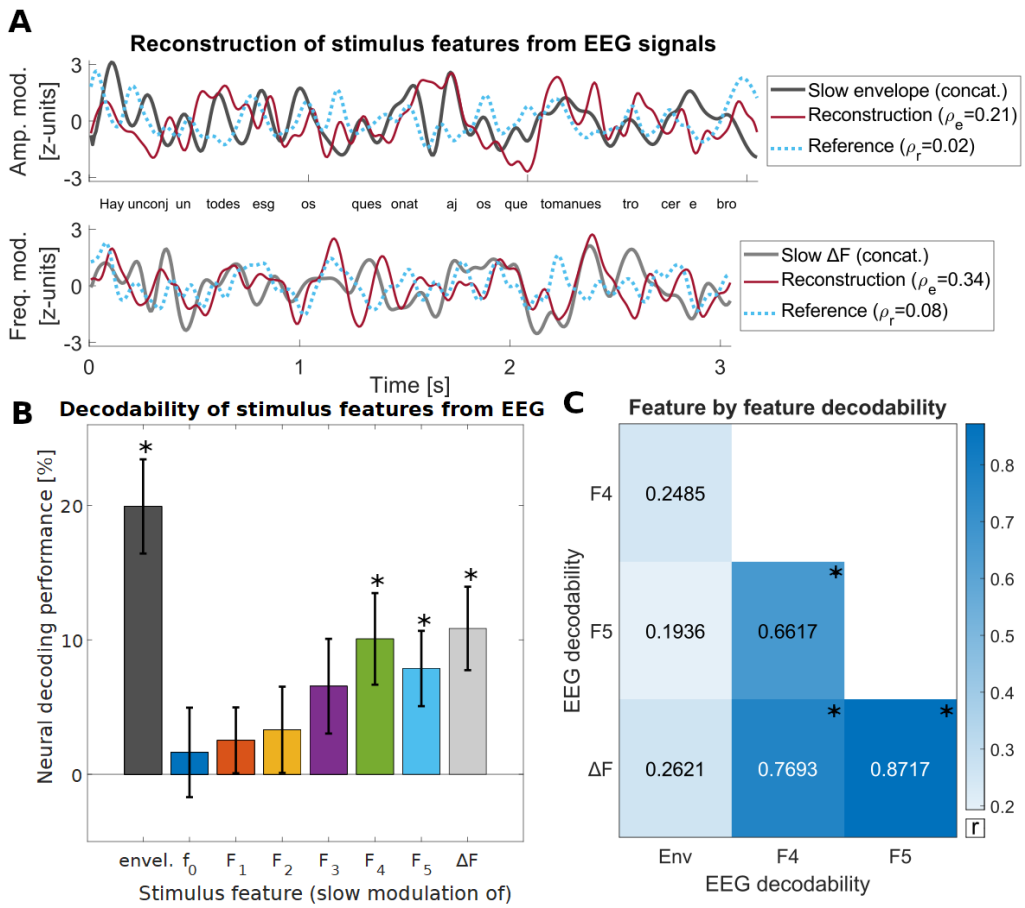


Figure 2. ***Spectral feature waveforms from voiced speech are reliably decoded from single-trial EEG***. **(A)** Sample reconstruction of speech feature waveforms after the sentence example from Figure 1. Two estimated decoders per listener (N = 19 for this stimulus) were obtained, corresponding to empirical versus reshuffled EEG datasets. *Top*: Slow modulations of the broadband envelope, plus subject-averaged envelope reconstructions by decoder type. Reconstruction performance is here represented by the rank correlation between the stimulus and mean reconstruction; the empirical performance exceeds that of reference decoders estimated after reshuffled EEG data. *Bottom*: Application of the method to a spectral vocal feature, instantaneous frequency interband spacing (ΔF). **(B)** The slow amplitude modulations of the envelope, frequency modulations of spectral features from the higher formant range, and ΔF, are reliably decoded from single-trial EEG across listeners. **(C)** Pairwise correlations between acoustic feature decodabilities, across listeners. The three robustly decodable spectral features significantly correlate with each other, indicating that neural tracking across these dimensions is mutually predictable. These decodabilities however do not predict decodability of the amplitude dynamics of the broadband envelope.

We next addressed whether listeners that achieve better envelope decodability similarly achieve higher decodabilities across the relevant spectral domains F_4, F_5, and ΔF. Using a Pearson correlation matrix on subject decodability data by feature (Figure 2C), significant linear relationships among feature pairs were found, although these were limited within all spectral feature pairs (all adj. p-values < $1.6\text{x}10^{-4}$). Subject decodability for any of these features did not predict subject decodability for the envelope (all adj. p-values > 0.6). These results indicate that listeners may encode the spectral features of speech by relying on temporally matching modulations across spectral variations and their spacing patterns in the formant range. In addition, the results suggest a distinct acoustic processing pathway from that which encodes sound level variations signaled by the speech broadband envelope.

Spectral tracking abilities predict performance gains at the cocktail party

Facing new, unfamiliar speakers, selective listening depends critically on the accurate delimiting of spectral sets that target a speaker's content. Our second aim was to test whether a listener's spectral tracking ability may help her select and resolve speech from a cocktail party of unfamiliar speakers. Spectral tracking ability was quantified by ||ρ||, the L2-norm of the vector of empirical decoder performances ($\rho_{e,F4}$, $\rho_{e,F5}$, $\rho_{e,\Delta F}$) which increases as a subject consistently encodes all three relevant spectral features. We hypothesized that higher ||ρ|| may boost a listener's chances to reliably track a masked speaker of choice, as participants also performed a 'cocktail-party' comprehension task. Their level of intelligibility during this stage of the experiment was evaluated through a challenging comprehension questionnaire (see *Supplementary Materials and Methods*). The proportion of correct cocktail party trials that were achieved by the listener was used as index of performance. A robust correlation analysis with interval testing (Pernet et al. 2013) showed that the summary neural measure ||ρ|| and behavioral cocktail party task performance were found to be positively correlated, r_s = 0.39, CI = 0.18 – 0.57, p = 8.05x10^{-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, ||ρ|| similarly predicted behavioral data when expressed in terms of the amount of points earned per trial, r_s = 0.36, CI = 0.13 – 0.55, p = 0.002 (Figure 3B), confirmed by robust regression $t(70)$ = 3.032; p = 0.003.

The cocktail party questionnaire includes items probing three different aspects of target speech, namely, the speaker's vocal identity, the keywords in speech, and the classification of its subject matter or topic (see Supplementary Methods). The ability of ||ρ|| to predict responses in each of these domains separately was further analyzed. First, ||ρ|| was found to predict the likelihood of listeners' correct choices about target speakers' identities, r_s = 0.31, CI = 0.09 – 0.51, p = 0.007 (Supplementary Figure 3A), confirmed by robust regression $t(70)$ = 3.225; p = 0.002. Second, ||ρ|| was similarly found to predict the likelihood of listeners' correct choices about target speech keywords, r_s = 0.32, CI = 0.09 – 0.54, p = 0.006 (Supplementary Figure 3B), confirmed by robust regression $t(70)$ = 2.382; p = 0.020. Third, ||ρ|| did not significantly predict the likelihood of listeners' correct choices about target speech topic, however, r_s = 0.21, CI = -0.03 – 0.41, p = 0.078 (Supplementary Figure 3C). Moreover, cocktail-party data involved voices of either sex, and we verified by robust regression that ||ρ|| measures of spectral tracking from male listening EEG data also predicted listener's cocktail party behavioral perfomance during the targeting of a male speaker, $t(70)$ = 2.913; p = 0.004. The positive correlation similarly held for the subset of cocktail party trials exclusively involving male+male voice combinations, $t(70)$ = 3.192; p = 0.002.

Finally, we examined whether the neural tracking of spectral features ($\rho_{e,F4}$, $\rho_{e,F5}$, $\rho_{e,\Delta F}$) may individually predict cocktail party performance (Figure 3C). In each case, the individual neural tracking measure and behavioral cocktail party task performance were found to be positively correlated, 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. Each of the significant correlations was confirmed by robust regression analysis, 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 findings are consistent with the notion that successful selection in multispeaker conditions is promoted when cortical responses follow through major changes of spectral attributes that serve to distinguish between vocal tracts.

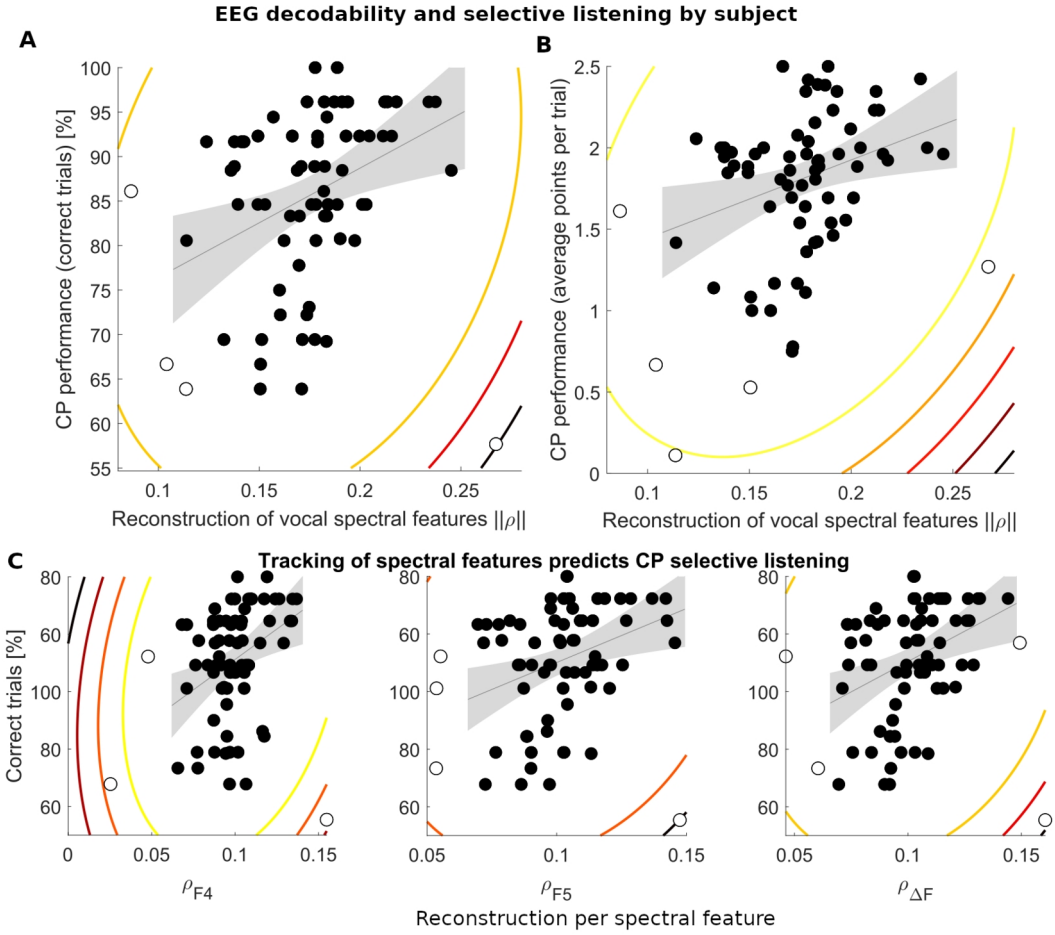


Figure 3. ***Spectral features of voiced speech assist performance in multispeaker conditions***. **(A)** A summary measure of neural decodability across the three relevant spectral features significantly predicts the behavioral performance of listeners at the 'cocktail party'. **(B)** Joint decodability 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 formant range instantaneous frequencies (F_4, F_5) and instantaneous frequency spacing (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. Contour lines indicate distance D_s from the bivariate distribution mean, expressed in z-units (squared) and shown from D_s=12 in steps of 6 (Schwarzkopf et al., 2012). Points with $D_s \geq 6$ are defined as outliers.

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 instantaneous spectral variations

from higher formant bands 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. Decoding findings were also presented here for a key feature in voice recognition studies, the spacing of instantaneous frequencies between formant bands (ΔF), which may be relevant because inter-formant dispersion is often used to infer anatomical dimensions of the source's vocal tract (Lavan, Knight, et al. 2019), (Lammert and Narayanan 2015). All three spectral features considered, their joint decodabilities predicted listeners' ability to selectively attend in a two-speaker scenario.

Cortical tracking of spectral contours in the higher formant band

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 (Friedman 1985). Extracted pitch and spectral cues, including formants, serve to build representations of a speaker's voice as listeners face the problem of determining speaker invariance (Lavan, Knight, et al. 2019). This is a crucial aspect of listeners' organization of multispeaker scenes, which needs estimating the variability that corresponds to one voice with sufficient precision in order to set it apart from maskers' (Darwin et al. 2003) (Lavan, Burton, et al. 2019). Studies suggest that the first 400 ms of auditory processing perform the necessary spectral vocal cue processing analyses (Levy et al. 2001; Charest et al. 2009; Capilla et al. 2013). Continuous sequential analyses of the unfamiliar voice signal are traditionally assumed to trigger the computational process of perceptual voice-identity processing where identity-features are extracted and merged, creating a coherent voice percept (Maguinness et al. 2018). Here, we present sequential filterbank analyses of a wide collection of voice stimuli, based on the phase of the time-varying analytic signal at each band. The results demonstrate cortical tracking coding of frequency (rather than amplitude) modulations within the formant range. It is important to consider the fact that the stimulus reconstructions are based on time-varying frequency contours of content within formant bands, without this necessarily be equal to the formants themselves. The present findings therefore build on previous studies of pitch contours tracked at the auditory cortex (Van Canneyt et al. 2021; Teoh et al. 2019; Brodbeck and Simon 2022). With regards to inconsistent findings on f_0 decoding between these prior reports and the current study, we note three important classes of differences. The first relates to differences in the stimulus representation, where broadband rather than low-frequency temporal modulations of pitch waveforms were previously addressed, assessing spectral variations at correspondingly finer temporal resolutions. In the case of Van Canneyt and colleagues, evidence of f_0 tracking was most robust at the brainstem stage (Etard et al. 2019) but cortical responses up to 40 ms were also inspected. They found cortical tracking to be optimal for male speaker stimuli voiced with low, steady f_0 values. In our case, we found spectral cue decoding across the range of male speakers comprising middle and late stage activity from a 400 ms period, but this did not include the f_0 band. The inconsistency would tentatively suggest that phase-locked responses to the coarser, slow modulations of f_0 that we sample may vanish from population activity following a temporal code, over cortical stages across multiple speakers.

Second, both Teoh et al., 2019 and Brodbeck and Simon, 2022 focussed on cortical responses up to 300 ms (or more), finding cortical tracking of f_0 contours in the low frequency bands of M/EEG responses. Unlike Van Canneyt et al., 2021, the stimuli representations they investigated convey critical information about all silent periods in presented speech. For this reason, their findings of f_0 tracking responses may be potentially

accompanied by on/off response coding that is also signaled via envelope tracking. We consider that such types of contribution to forward models of any continuous stimulus feature may result in small but non-negligible performance advantages over distributions of trial-reshuffled data. For this reason we recommend that, given the dominance of envelope tracking in auditory cortical responses, silent periods be removed prior to modeling whether in f_0 tracking or other standalone continuous features (an alternative is to introduce a voicing binary variable, cf. Tang et al. 2017). When envelope tracking is compared against tracking of other features as is the case here, it is further adequate to excise unvoiced periods from the envelope signal itself as well. Despite remaining segments missing all elements of natural prosody (of which silences may play an important role), removal provides for comparisons of resolving spectral contours unbiased by systematic temporal edge responses.

Third, prior studies (Teoh et al. 2019; Van Canneyt et al. 2021; Brodbeck and Simon 2022) addressed coding models of a speaker's normalized relative pitch. By contrast, our backward decoding models incorporate joint data across multiple speakers on a common spectral frame. The present approach hence rather models the temporal modulations of absolute changes in f_0 waveforms across speakers because they are proportional to their differences. Using electrocorticography, Tang et al. 2017 directly addressed the distinction between absolute and relative pitch encoding in cortical networks. Their findings indicated the two types of encoding may be subserved by separate networks in the auditory cortex, with a possible dominance of relative pitch encoding. Our negative finding might reflect that network activity from absolute pitch encoding does not generally predict neural discriminability of temporal responses to different intonation contours in the auditory cortex (Tang et al. 2017). Although the previous M/EEG studies employed one or few different speakers, their f_0 tracking findings appear consistent with relative pitch encoding, which does predict neural discriminability of intonation contours. Nevertheless, from the perspective of a 'cocktail-party' setting, tracking models of absolute spectral features, including the formant range, remain crucial because they are the basis for speaker voice segregation. Our lack of significant findings for spectral cues in the f_0, F_1–F_3 bands may suggest a lower bandwidth limit for slow absolute FM changes encoded in delta/theta activity. It is worth noting that, while phonemically important fast rates of change occur in movements of the lower formants that encode short vowels (Baumann and Belin 2010), higher formant band changes appear less systematically related to vowel quality. Moreover, vowel encoding is considered to be subject to subsequent speaker-invariant representations via rescaling processes (Johnson and Sjerps 2021). As such, determination of spectral cues from higher formant bands, in absolute terms, could provide a reliable channel for sustained discrimination of unfamiliar voices over the first few seconds of multi-speaker listening.

Spectral tracking as predictor of listeners' speech segregation performance

Spectral contour reconstruction analyses further demonstrated their ability to relate the likelihood of listeners' correct choices on relevant aspects of comprehension behavior in a 'cocktail-party' task. One fundamental aspect of effective speaker selectivity is the successful discrimination of target speaker identities from the background, such as on the basis of their perceived sex or age. Neural decodability measures predicted such ability, which is consistent with the encoding of absolute spectral features from the dynamic voice signal as a basis for correct speaker identification. Importantly, reconstruction accuracy analyses demonstrated the ability to predict the likelihood of listeners' identification of correct word utterances from the target speaker. This finding is consistent with the involvement of spectral contour tracking in the processes that lead to extraction of lexical content during selective

speech listening. At the level of topic information, we did not find evidence of similar correlations of listener's correct choices with reconstruction accuracy however. This null result may reflect the limitations of relatively lower-level spectral cue coding in the inference of higher-level thematic content.

Neural measures obtained under optimal listening conditions remain a key element of models of speech intelligibility for sub-optimal listening conditions with irrelevant sources, such as different noise levels (e.g. neurometric lapse rate, Vanthornhout et al. 2018). Here, neural measures were obtained from the single-speaker task, and used to predict behavioral performance of the two-speaker cocktail-party task. Hence, we tested whether the quality of neural representations from noise-free scenarios is useful to predict the degree of listeners' intelligibility in multiple-speaker listening conditions – without recourse to its likely contingency on noise adaptation mechanisms (Khalighinejad et al. 2019). This means that, in our behavioral testing, we did not specifically address whether neural representations from ideal conditions remain noise-invariant. Nevertheless, the observation that such representations impact behavioral performance may suggest that, across listeners, clean representations may be commensurate with those potentially achieved under noise adaptation, although this remains an open question.

For the reason above, our current approach may be insufficient when one aims to predict within-subject intelligibility across interference levels, for instance. Yet, it may be advantageous regarding the auditory scene analysis of elements continually subject to causal inference during multi-speaker masking. This refers to the process by which the observer adjudicates common or separate causes (e.g. speakers) to the auditory signal. Its computation may involve the different resolved spectral contours and their combinations, where clean speech occurrences from occurring masker-free periods become particularly useful. Prior research has shown that processing of such overt periods is expedited when part of the attended target, as opposed to its masked elements, whose processing is delayed (Brodbeck et al. 2020). In practical terms, responses based on single-speaker epochs convey listeners' tracking ability of a target's key spectral features for periods that are exempt from causal inference. Such opportunities represent a means for listeners to constrain and establish an unfamiliar vocal identity from the varying speech signal on summary statistical patterns (Lavan, Knight, et al. 2019). In computational voice coding models, static pitch and formant dispersion parameters have traditionally represented the coordinates of hypothesized vocal space (Baumann and Belin 2010; Latinus et al. 2013; Lavan, Knight, et al. 2019). 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. Spectral cues that are based on formant spacing may show higher reliability for speaker identification than pitch (Gaudrain et al. 2009). Inter-formant spacing is extracted by the auditory system (Ives et al. 2005; Smith et al. 2005; Pisanski et al. 2014) as early as at the medial geniculate body (von Kriegstein et al. 2006), and is cortically represented in posterior superior temporal lobe networks that include temporal voice areas (von Kriegstein et al. 2007; von Kriegstein et al. 2010; Mathias and von Kriegstein 2014). These regions may extend across the posterior aspect of the right superior temporal sulcus to the superior temporal pole (Belin et al. 2000; Belin and Zatorre 2003; Belin et al. 2004; Kriegstein and Giraud 2004; Latinus et al. 2011; Latinus et al. 2013; Pernet et al. 2015; Belin 2017; Belin et al. 2018; Mathias and von Kriegstein 2019). Responses in this region may be considered to be involved in low-level voice processing when they remain specific to voice-like stimuli regardless of intelligibility (Rosen et al. 2011). For this circuitry, multispeaker conditions is naturally impose short time spans to attempt reliable voice discrimination.

Spectral cues within these reduced windows then provide variations of an unfamiliar voice that may be harnessed and reliably summarized as in clean conditions. Albeit at a slower rate, the process could sequentially buildup to form within-person vocal prototype representation, e.g. (Lavan, Knight, et al. 2019; Lavan, Burton, et al. 2019). Future studies may determine whether tracking abilities, at masked periods, possibly sustain the likelihood that an inferred spectral grouping can be maintained in consideration that alternative formant complexes are part of causal inference.

Finally, many studies have consistently shown the neural tracking of more abstract and higher-level properties of speech than shown here, ranging from detailed narrowband amplitude dynamics to phonemic changes and lexical probabilistic structure (Di Liberto et al. 2015). The present work extends these prior findings by introducing a basis for effective targeting of a *caller* of choice, with potential selective communication across wider animal models. Lexical processing of speech is constrained by the precise encoding of vocal tract modulations (Perrachione and Wong 2007; Chandrasekaran et al. 2011; Perrachione et al. 2011; McGettigan and Scott 2012; Zäske et al. 2017). Relative temporal displacements between spectral components affect intelligibility performance (Stilp et al. 2010), for example from surface reverberation effects that lead to weaker segregation in multispeaker conditions (Abouchacra et al. 2011; Sayles et al. 2016). Bandwidth tuning loss at higher frequencies under presbycusis also severely limits the ability to target speech in multispeaker conditions. The process by which these levels of uncertainty affect the monitoring or tracking of spectral groupings and, more generally, the causal arbitration to integrate or segregate them, remain unclear. Future studies may also 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.

Conclusion

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.

Acknowledgments and funding sources

Conflict of interest: None to declare. 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

Abouchacra KS, Besing J, Koehnke J, Letowski T. 2011. 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. 50(7):468–476. doi:10.3109/14992027.2011.565424.

Baumann O, Belin P. 2010. Perceptual scaling of voice identity: common dimensions for different vowels and speakers. Psychol Res-Psychol Forsch. 74(1):110–120. doi:10.1007/s00426-008-0185-z.

Belin P. 2017. Similarities in face and voice cerebral processing. Vis Cogn. 25(4–6):658–665. doi:10.1080/13506285.2017.1339156.

Belin P, Bodin C, Aglieri V. 2018. A "voice patch" system in the primate brain for processing vocal information? Hear Res. 366:65–74. doi:10.1016/j.heares.2018.04.010.

Belin P, Fecteau S, Bédard C. 2004. Thinking the voice: neural correlates of voice perception. Trends Cogn Sci. 8(3):129–135. doi:10.1016/j.tics.2004.01.008.

Belin P, Zatorre RJ. 2003. Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport. 14(16):2105–2109. doi:10.1097/00001756-200311140-00019.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. 2000. Voice-selective areas in human auditory cortex. Nature. 403(6767):309–312. doi:10.1038/35002078.

Brodbeck C, Jiao A, Hong LE, Simon JZ. 2020. Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers. PLOS Biol. 18(10):e3000883. doi:10.1371/journal.pbio.3000883.

Brodbeck C, Simon JZ. 2020. Continuous speech processing. Current Opinion in Physiology. 18:25–31. doi:10.1016/j.cophys.2020.07.014.

Brodbeck C, Simon JZ. 2022. Cortical tracking of voice pitch in the presence of multiple speakers depends on selective attention. Front Neurosci. 16. doi:10.3389/fnins.2022.828546.

Bruckert L, Liénard JS, Lacroix A, Kreutzer M, Leboucher G. 2006. Women use voice parameters to assess men's characteristics. Proc R Soc B-Biol Sci. 273(1582):83–89. doi:10.1098/rspb.2005.3265.

Capilla A, Belin P, Gross J. 2013. The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. Cereb Cortex N Y N. 23(6):1388–1395. doi:10.1093/cercor/bhs119.

Chandrasekaran B, Chan AHD, Wong PCM. 2011. Neural Processing of What and Who Information in Speech. J Cogn Neurosci. 23(10):2690–2700. doi:10.1162/jocn.2011.21631.

Charest I, Pernet CR, Rousselet GA, Quinones I, Latinus M, Fillion-Bilodeau S, Chartrand J-P, Belin P. 2009. Electrophysiological evidence for an early processing of human voices. BMC Neurosci. 10:127. doi:10.1186/1471-2202-10-127.

de Cheveigné A, Simon JZ. 2007. Denoising based on time-shift PCA. J Neurosci Methods. 165(2):297–305. doi:10.1016/j.jneumeth.2007.06.003.

de Cheveigné A, Simon JZ. 2008. Sensor noise suppression. J Neurosci Methods. 168(1):195–202. doi:10.1016/j.jneumeth.2007.09.012.

Collins SA. 2000. Men's voices and women's choices. Anim Behav. 60:773–780. doi:10.1006/anbe.2000.1523.

Darwin CJ, Brungart DS, Simpson BD. 2003. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. J Acoust Soc Am. 114(5):2913–2922. doi:10.1121/1.1616924.

David SV, Mesgarani N, Shamma SA. 2007. Estimating sparse spectro-temporal receptive fields with natural stimuli. Netw Comput Neural Syst. 18(3):191–212. doi:10.1080/09548980701609235.

Di Liberto GM, O'Sullivan JA, Lalor EC. 2015. Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. Current Biology. 25(19):2457–2465. doi:10.1016/j.cub.2015.08.030.

Ding N, Simon JZ. 2012. Emergence of neural encoding of auditory objects while listening to competing speakers. PNAS. 109(29):11854–11859. doi:10.1073/pnas.1205381109.

Dolan B, Brockett C. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. [accessed 2024 Apr 4]. https://www.microsoft.com/en-us/research/publication/automatically-constructing-a-corpus-of-sentential-paraphrases/.

Etard O, Kegler M, Braiman C, Forte AE, Reichenbach T. 2019. Decoding of selective attention to continuous speech from the human auditory brainstem response. Neuroimage. 200:1–11. doi:10.1016/j.neuroimage.2019.06.029.

Evans S, Neave N, Wakelin D. 2006. Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. Biol Psychol. 72(2):160–163. doi:10.1016/j.biopsycho.2005.09.003.

Fitch WT. 1997. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. J Acoust Soc Am. 102(2):1213–1222. doi:10.1121/1.421048.

Friedman D. 1985. 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. Vol. 10. p. 1121–1124.

Gaudrain E, Li S, Ban VS, Patterson RD. 2009. 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. p. 152–155.

Glasberg BR, Moore BCJ. 1990. Derivation of auditory filter shapes from notched-noise data. Hear Res. 47(1–2):103–138. doi:10.1016/0378-5955(90)90170-T.

Grimaldi M, Cummins F. 2008. Speaker Identification Using Instantaneous Frequencies. IEEE Trans Audio Speech Lang Process. 16(6):1097–1111. doi:10.1109/TASL.2008.2001109.

Hyvarinen A. 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw. 10(3):626–634. doi:10.1109/72.761722.

Ives DT, Smith DRR, Patterson RD. 2005. Discrimination of speaker size from syllable phrases. J Acoust Soc Am. 118(6):3816–3822. doi:10.1121/1.2118427.

Johnson K. 2020. The ΔF method of vocal tract length normalization for vowels. Lab Phonol J Assoc Lab Phonol. 11(1):10. doi:10.5334/labphon.196.

Johnson K, Sjerps MJ. 2021. Speaker Normalization in Speech Perception. In: The Handbook of Speech Perception. John Wiley & Sons, Ltd. p. 145–176. [accessed 2021 Aug 17]. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119184096.ch6.

Junghöfer M, Elbert T, Tucker DM, Rockstroh B. 2000. Statistical control of artifacts in dense array EEG/MEG studies. Psychophysiology. 37(04):523–532. doi:null.

Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N. 2019. Adaptation of the human auditory cortex to changing background noise. Nat Commun. 10(1):2509. doi:10.1038/s41467-019-10611-4.

Kriegeskorte N, Douglas PK. 2019. Interpreting encoding and decoding models. Curr Opin Neurobiol. 55:167–179. doi:10.1016/j.conb.2019.04.002.

Lammert AC, Narayanan SS. 2015. On Short-Time Estimation of Vocal Tract Length from Formant Frequencies. PLoS ONE. 10(7):e0132193. doi:10.1371/journal.pone.0132193.

Latinus M, Crabbe F, Belin P. 2011. Learning-Induced Changes in the Cerebral Processing of Voice Identity. Cereb Cortex. 21(12):2820–2828. doi:10.1093/cercor/bhr077.

Latinus M, McAleer P, Bestelmeyer PEG, Belin P. 2013. Norm-Based Coding of Voice Identity in Human Auditory Cortex. Curr Biol. 23(12):1075–1080. doi:10.1016/j.cub.2013.04.055.

Lavan N, Burton AM, Scott SK, McGettigan C. 2019. Flexible voices: Identity perception from variable vocal signals. Psychon Bull Rev. 26(1):90–102. doi:10.3758/s13423-018-1497-7.

Lavan N, Knight S, McGettigan C. 2019. Listeners form average-based representations of individual voice identities. Nat Commun. 10:2404. doi:10.1038/s41467-019-10295-w.

Lee Y, Keating P, Kreiman J. 2019. Acoustic voice variation within and between speakers. J Acoust Soc Am. 146(3):1568–1579. doi:10.1121/1.5125134.

Levy DA, Granot R, Bentin S. 2001. Processing specificity for human voice stimuli: electrophysiological evidence. Neuroreport. 12(12):2653–2657. doi:10.1097/00001756-200108280-00013.

López S, Riera P, Assaneo MF, Eguía M, Sigman M, Trevisan MA. 2013. Vocal caricatures reveal signatures of speaker identity. Sci Rep. 3(1):3407. doi:10.1038/srep03407.

Maguinness C, Roswandowitz C, von Kriegstein K. 2018. Understanding the mechanisms of familiar voice-identity recognition in the human brain. Neuropsychologia. 116:179–193. doi:10.1016/j.neuropsychologia.2018.03.039.

Mathias SR, von Kriegstein K. 2014. How do we recognise who is speaking? Front Biosci Sch Ed. 6(1):92–109. doi:10.2741/s417.

Mathias SR, von Kriegstein K. 2019. Voice Processing and Voice-Identity Recognition. In: Siedenburg K, Saitis C, McAdams S, Popper AN, Fay RR, editors. Timbre: Acoustics, Perception, and Cognition. Cham: Springer International Publishing. (Springer Handbook of Auditory Research). p. 175–209. doi:10.1007/978-3-030-14832-4_7.

McGettigan C, Scott SK. 2012. Cortical asymmetries in speech perception: what's wrong, what's right and what's left? Trends Cogn Sci. 16(5):269–276. doi:10.1016/j.tics.2012.04.006.

Mesgarani N. 2014. Stimulus Reconstruction from Cortical Responses. In: Jaeger D, Jung R, editors. Encyclopedia of Computational Neuroscience. Springer New York. p. 1–3. doi:10.1007/978-1-4614-7320-6_108-1.

Mowlaee P, Saeidi R, Stylianou Y. 2016. Advances in phase-aware signal processing in speech communication. Speech Commun. 81:1–29. doi:10.1016/j.specom.2016.04.002.

O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC. 2015. Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. Cereb Cortex. 25(7):1697–1706. doi:10.1093/cercor/bht355.

Peelle JE, Gross J, Davis MH. 2013. Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. Cerebral Cortex. 23(6):1378–1387. doi:10.1093/cercor/bhs118.

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

Pernet CR, McAleer P, Latinus M, Gorgolewski KJ, Charest I, Bestelmeyer PEG, Watson RH, Fleming D, Crabbe F, Valdes-Sosa M, et al. 2015. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. NEUROIMAGE. 119:164–174. doi:10.1016/j.neuroimage.2015.06.050.

Perrachione TK, Del Tufo SN, Gabrieli JDE. 2011. Human Voice Recognition Depends on Language Ability. Science. 333(6042):595–595. doi:10.1126/science.1207327.

Perrachione TK, Wong PCM. 2007. Learning to recognize speakers of a non-native language: implications for the functional organization of human auditory cortex. Neuropsychologia. 45(8):1899–1910. doi:10.1016/j.neuropsychologia.2006.11.015.

Pisanski K, Fraccaro PJ, Tigue CC, O'Connor JJM, Roeder S, Andrews PW, Fink B, DeBruine LM, Jones BC, Feinberg DR. 2014. Vocal indicators of body size in men and women: a meta-analysis. Anim Behav. 95:89–99. doi:10.1016/j.anbehav.2014.06.011.

Potamianos A, Maragos P. 1996. Speech formant frequency and bandwidth tracking using multiband energy demodulation. J Acoust Soc Am. 99(6):3795–3806. doi:10.1121/1.414997.

Průša Z, Søndergaard PL, Holighaus N, Wiesmeyr C, Balazs P. 2014. 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. p. 419–442. doi:10.1007/978-3-319-12976-1_25.

Puts DA, Barndt JL, Welling LLM, Dawood K, Burriss RP. 2011. Intrasexual competition among women Vocal femininity affects perceptions of attractiveness and flirtatiousness. Personal Individ Differ. 50(1):111–115. doi:10.1016/j.paid.2010.09.011.

Puts DA, Hodges CR, Cardenas RA, Gaulin SJC. 2007. Men's voices as dominance signals:: vocal fundamental and formant frequencies influence dominance attributions among men. Evol Hum Behav. 28(5):340–344. doi:10.1016/j.evolhumbehav.2007.05.002.

Puvvada KC, Simon JZ. 2017. Cortical Representations of Speech in a Multitalker Auditory Scene. J Neurosci. 37(38):9189–9196. doi:10.1523/JNEUROSCI.0938-17.2017.

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

Riede T, Fitch T. 1999. Vocal tract length and acoustics of vocalization in the domestic dog (*Canis familiaris*). J Exp Biol. 202(20):2859–2867.

Rosen S, Wise RJS, Chadha S, Conway E-J, Scott SK. 2011. Hemispheric Asymmetries in Speech Perception: Sense, Nonsense and Modulations. PLOS ONE. 6(9):e24672. doi:10.1371/journal.pone.0024672.

Sadjadi SO, Hansen JHL. 2015. Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification. SPEECH Commun. 72:138–148. doi:10.1016/j.specom.2015.04.005.

Sayles M, Stasiak A, Winter IM. 2016. Neural Segregation of Concurrent Speech: Effects of Background Noise and Reverberation on Auditory Scene Analysis in the Ventral Cochlear Nucleus. In: VanDijk P, Baskent D, Gaudrain E, DeKleine E, Wagner A, Lanting C, editors. Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing. Vol. 894. Berlin: Springer-Verlag Berlin. p. 389–397.

Schwarzkopf D, de Haas B, Rees G. 2012. Better Ways to Improve Standards in Brain-Behavior Correlation Analysis. Front Hum Neurosci. 6. doi:10.3389/fnhum.2012.00200.

Smith DRR, Patterson RD, Turner R, Kawahara H, Irino T. 2005. The processing and perception of size information in speech sounds. J Acoust Soc Am. 117(1):305–318. doi:10.1121/1.1828637.

Søndergaard PL, Torrésani B, Balazs P. 2012. The linear time frequency analysis toolbox. Int J Wavelets Multiresolution Inf Process. 10(04):1250032. doi:10.1142/S0219691312500324.

Stilp CE, Kiefte M, Alexander JM, Kluender KR. 2010. Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences. J Acoust Soc Am. 128(4):2112–2126. doi:10.1121/1.3483719.

Tang C, Hamilton LS, Chang EF. 2017. Intonational speech prosody encoding in the human auditory cortex. Science. 357(6353):797–801. doi:10.1126/science.aam8577.

Teoh ES, Cappelloni MS, Lalor EC. 2019. 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. 50(11):3831–3842. doi:10.1111/ejn.14510.

Van Canneyt J, Wouters J, Francart T. 2021. Neural tracking of the fundamental frequency of the voice: The effect of voice characteristics. Eur J Neurosci. 53(11):3640–3653. doi:10.1111/ejn.15229.

von Kriegstein K, Smith DRR, Patterson RD, Ives DT, Griffiths TD. 2007. Neural representation of auditory size in the human voice and in sounds from other resonant sources. Curr Biol. 17(13):1123–1128. doi:10.1016/j.cub.2007.05.061.

von Kriegstein K, Smith DRR, Patterson RD, Kiebel SJ, Griffiths TD. 2010. How the Human Brain Recognizes Speech in the Context of Changing Speakers. J Neurosci. 30(2):629–638. doi:10.1523/JNEUROSCI.2742-09.2010.

von Kriegstein K, Warren JD, Ives DT, Patterson RD, Griffiths TD. 2006. Processing the acoustic effect of size in speech sounds. NeuroImage. 32(1):368–375. doi:10.1016/j.neuroimage.2006.02.045.

von Kriegstein KV, Giraud AL. 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. NeuroImage. 22(2):948–955. doi:10.1016/j.neuroimage.2004.02.020.

Wilsch A, Neuling T, Obleser J, Herrmann CS. 2018. Transcranial alternating current stimulation with speech envelopes modulates speech comprehension. NeuroImage. 172:766–774. doi:10.1016/j.neuroimage.2018.01.038.

Zäske R, Awwad Shiekh Hasan B, Belin P. 2017. It doesn’t matter what you say: FMRI correlates of voice learning and recognition independent of speech content. Cortex J Devoted Study Nerv Syst Behav. 94:100–112. doi:10.1016/j.cortex.2017.06.005.

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, et al. 2013. Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” Neuron. 77(5):980–991. doi:10.1016/j.neuron.2012.12.037.

Zoefel B, Allard I, Anil M, Davis MH. 2020. Perception of Rhythmic Speech Is Modulated by Focal Bilateral Transcranial Alternating Current Stimulation. Journal of Cognitive Neuroscience. 32(2):226–240. doi:10.1162/jocn_a_01490.

Supplemental material for

Cortical tracking of speakers' spectral changes predicts selective listening

Francisco Cervantes Constantino and Ángel A. Caputi

Supplementary Materials and Methods

General overview

The data upon which this article is based on is part of a larger dataset that was obtained in the course of a series of studies on the cocktail party effect (Sánchez-Costa et al., results to be published elsewhere). These original EEG studies aim to investigate how prior speech listening experience modulates speech segregation, based on forward modeling during multi-speaker listening. In the present study, we addressed the stimulus reconstruction (i.e., backward modeling) of acoustic features in single-speaker speech. Neural estimates obtained from these data were compared with behavioral performance data at multi-speaker listening (cocktail party) by the same subjects. See *Task* in the main text and Supplementary Figure 1 below for 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 and task

Volunteer recordings were obtained with an Olympus recorder (VN-541OC, OM Digital Solutions Corporation, Japan). 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. All speech stimuli were marked for apparent sex and age (below or above 35 years-old). Between 2 and 4 keywords were selected per stimulus, which were limited to nouns, adverbs, or verbs in the sentence. Each sentence was also associated to at least one topic from a designated list of 13 themes.

Stimulus mixtures were created for the cocktail party task by adding two different stimuli from the database without replacement. A unique set of cocktail party stimuli was generated per participant, with equal probability of creating same- and different-sex combinations. Each cocktail party stimulus was built with a 750 ms onset asynchrony between speaker streams, and the resulting overlap time epoch marked the end of the stimulus. Stimuli were equalized in root mean squared (RMS) intensity for the overlapping epoch. All stimuli (single speaker or cocktail mixtures) were individually normalized.

In the present study, cocktail party behavioral data involves all control trials from the original studies (number of trials: 26 to 36). These trials consist of SSL phases that are unrelated to cocktail party listening phases, as they always involve different speakers, sentences, and content.

Questionnaire and scoring

Trial questionnaires consisted of a 9-option multiple choice set with available options that were based on aspects specific to the target or background speech. Participants were asked to select three options that best matched the indicated target. Questionnaires always contained one option that was related to speaker identity, plus eight options that were related to sentence content. With equal probability, the single identity option prompted participants on vocal qualities, either of the 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 lists of keywords corresponding to the two sentences. Finally, four more content options prompted participants to report on the topic of the attended speaker. The options were randomly selected from the databases' 12-theme list. 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: a) for identity options, selecting the opposite gender or opposite age of the attended speaker resulted in a point deduction; b) for keyword options, selecting a keyword related to the masker speech resulted in a point deduction; c) 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 that were entered and 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 broadband 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 broadband 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 discarded in corresponding periods of the envelope and in all the spectral features.

EEG preprocessing

After preprocessing, if a head channel resulted in activity with 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, after which the variance-based automated rejection procedure was applied across single channels and epochs' timeseries per participant. This resulted in < 1% rejected channel-epoch timeseries on average (subject range 0.09%–1.52%).

Stimulus-response alignment and decoder estimation

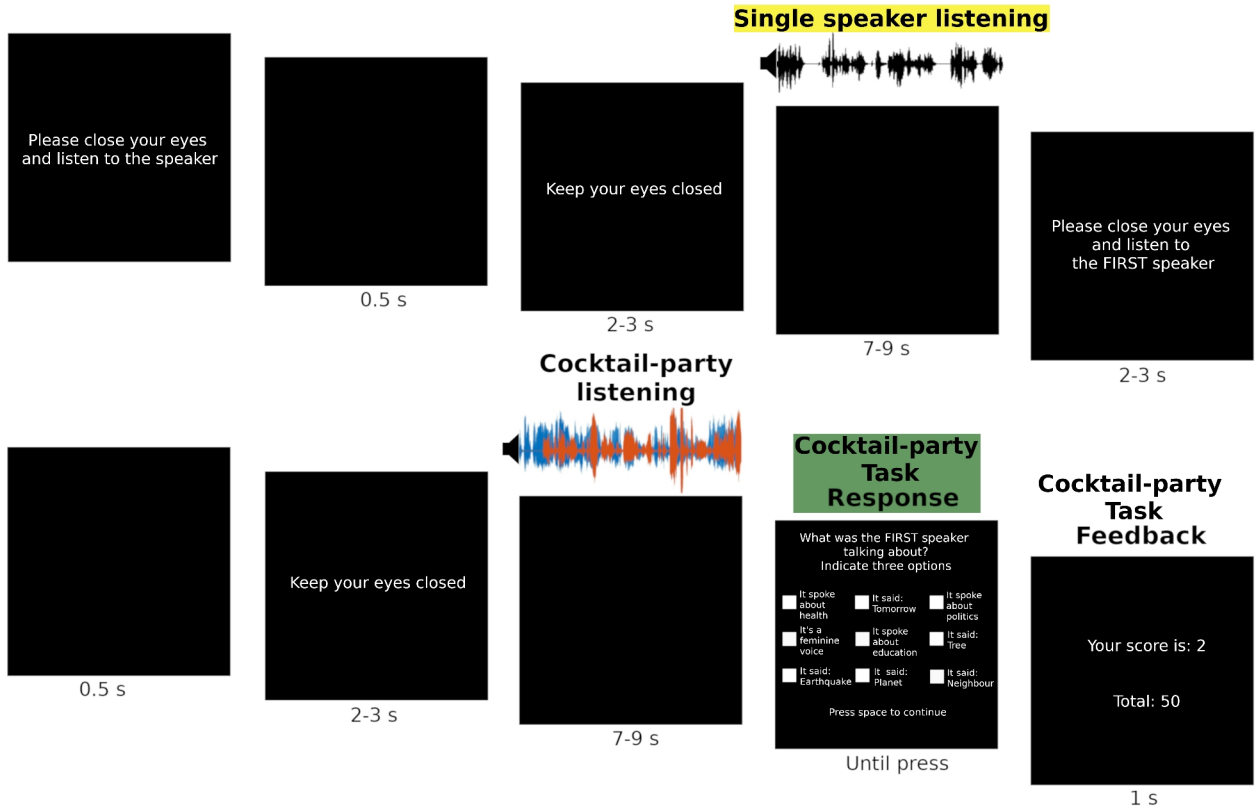
Male voice presentations were selected from each participant's EEG SSL phase data, resulting in 46 trials per participant on average (range: 31 – 62 trials). For any given time sample, channel activity levels exceeding 7 standard deviations of all values in the rest of the data were set as an empty value. EEG data were temporally aligned with the acoustic feature timeseries sets (including the envelope, f_0, F_1-F_5, and ΔF) by cross-correlation of the audio signal sampled by the EEG system with the corresponding raw audio original file downsampled to 1024 Hz. EEG and acoustic feature waveforms were aligned on the basis of the estimated cross-correlation delay. All data epochs when the broadband envelope signal was below a level threshold (see *Speech spectral feature analyses*, above) were removed at this point. 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 in the forward and reversed directions and resampled to 128 Hz. Prior to decoder estimation, all timeseries were transformed to z-score units.

Statistical analysis of decoder data

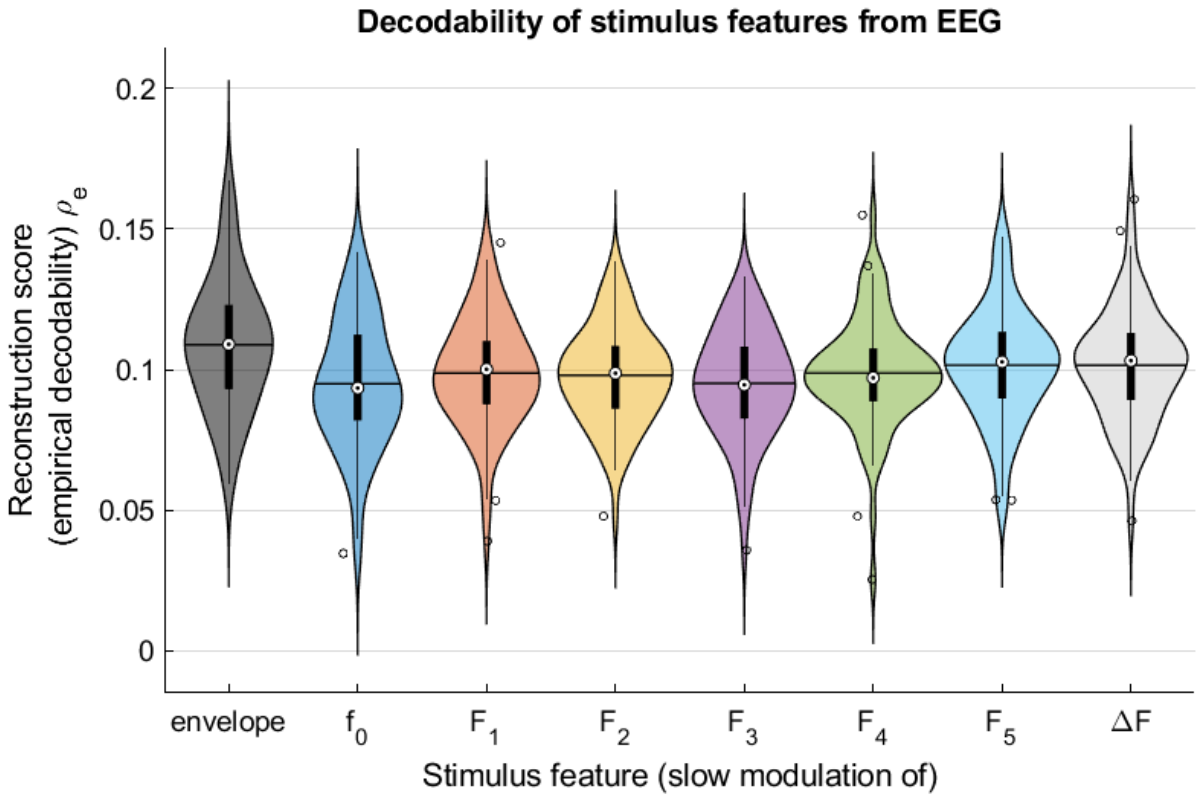
As null hypothesis, slow modulations in the *i*-th auditory feature were considered to not be reliably decoded from the EEG data, which may correspond to non-positive decoder reliability (neural decoding performance) estimates across listeners. To test the alternative hypothesis, one-sided Student's *t*-tests were run per auditory feature evaluating whether each feature modulation is reliably decodable; *p*-values were adjusted by the Holm-Bonferroni method to correct for multiple testing.

It is worthwhile noting that, in decoding studies, validation metrics may be affected by how decodability is assessed. Inspection of raw scores (empirical ρ values, Supplementary Fig. 2) suggests that correlations appear generally greater than a fixed reference level of $\rho=0$. To address generalization we have included cross-validation at the estimation stage, plus training of decoders de novo from reshuffled data. This latter step follows equal performance optimization rules as do decoders estimated from empirical data; generalization is then inferred upon greater performance for empirical- relative to reference-optimized results. Because this multi-stage approach is modeled after the statistical structure of both sampled stimuli and responses, it may impose relatively more stringent conditions than reference approaches based on models exclusively limited to the stimulus or the response.

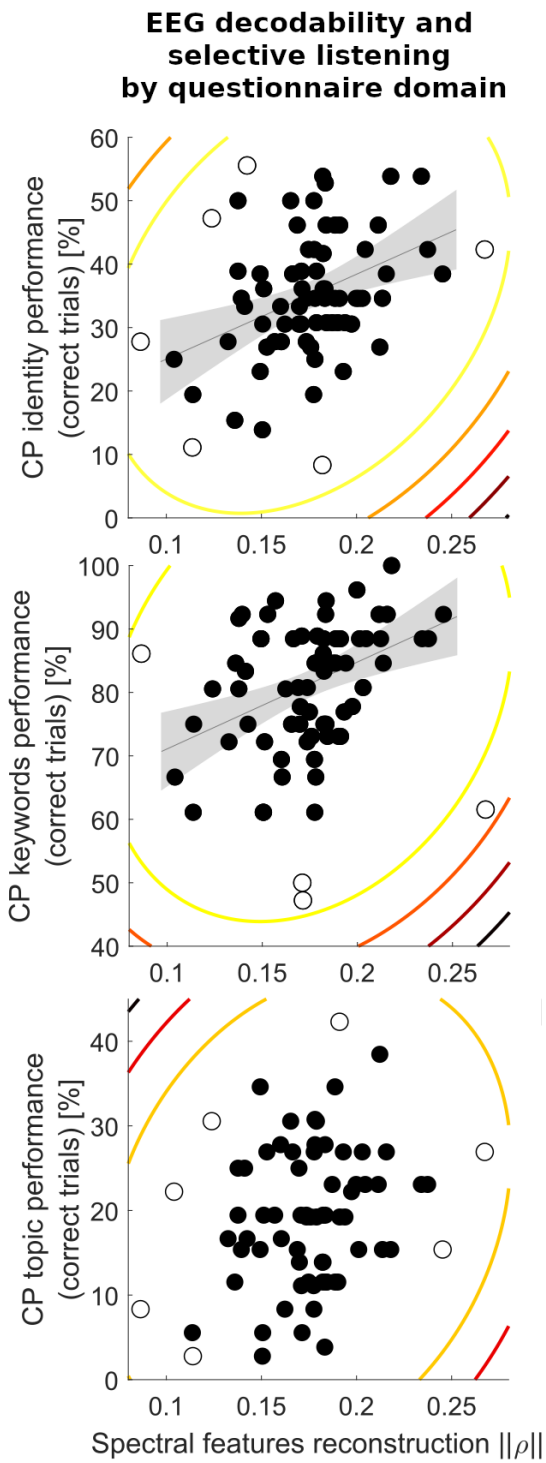
Supplementary Figures



Supplementary Figure 1. Trial design. In the first part of the trial, participants are instructed to listen to a single speaker presentation (yellow highlight) with their eyes closed. In the second part of the trial, they are given an instruction to attend to either the leading ("first") or interjecting ("second") talker at the upcoming cocktail party speech mixture with their eyes closed as well. After listening, a questionnaire prompts for different aspects of the indicated target speech, including speaker sex/age, topical content and specific keywords (green highlight). Trial and accumulated feedback are shown upon validation. Modified from Sánchez-Costa, et al.



Supplementary Figure 2. Empirical performance of EEG decoders by stimulus feature. Each plot represents the distribution across listeners (N=73) of the Spearman's rank correlation value evaluating the ability to reconstruct empirical data, with the mean indicated by horizontal lines. Boxplots indicate distribution median, IQR as well as outlier data. Across most features, the similar reconstruction scores taken from raw decoders may illustrate the need to estimate baseline performance from reference decoders.



Supplementary Figure 3. Spectral features' reconstruction and subject performance addressed by separate aspects of the cocktail party task. (A) Listeners' ||ρ|| predicts their ability to correctly respond to the part of the questionnaire that relates to target speaker's identity (the apparent age or sex of the voice) during selective listening. There is a significant robust correlation relationship observed between both neural and behavioral variables, Shepherd's Pi = 0.35, p = 0.007. (B) Similarly, ||ρ|| significantly predicts listeners' ability to discriminate and report the words expressed by the target speaker during the 'cocktail-party', Shepherd's Pi = 0.39, p = 0.002. (C) By contrast, listener correct responses about topic aspects of the questionnaire were not significantly predicted by ||ρ||, Shepherd's Pi = 0.18, p = 0.319.