Neurocognitive dynamics of near-threshold voice signal detection and affective voice evaluation

Huw Swanborough1,2*, Matthias Staib1,2, Sascha Frühholz1,2,3*

Communication and voice signal detection in noisy environments are universal tasks for many species. The fundamental problem of detecting voice signals in noise (VIN) is underinvestigated especially in its temporal dynamic properties. We investigated VIN as a dynamic signal-to-noise ratio (SNR) problem to determine the neurocognitive dynamics of subthreshold evidence accrual and near-threshold voice signal detection. Experiment 1 showed that dynamic VIN, including a varying SNR and subthreshold sensory evidence accrual, is superior to similar conditions with nondynamic SNRs or with acoustically matched sounds. Furthermore, voice signals with affective meaning have a detection advantage during VIN. Experiment 2 demonstrated that VIN is driven by an effective neural integration in an auditory cortical-limbic network at and beyond the near-threshold detection point, which is preceded by activity in subcortical auditory nuclei. This demonstrates the superior recognition advantage of communication signals in dynamic noise contexts, especially when carrying socio-affective meaning.

INTRODUCTION

Auditory perception requires the human auditory system to accurately detect, recognize, and classify auditory signals and auditory objects from the environment. While auditory processing and auditory object recognition are usually performed with high cognitive (1) and neural accuracy (2, 3) in cases of a clear auditory signal, this task becomes more challenging in real-life conditions containing various sources of background noise. This noise can render auditory signals undetectable and unintelligible and poses critical problems for individuals with hearing impairments (4).

The effects of acoustic noise on voice signal perception has mainly been investigated for the specific but socially very important case of understanding speech of other individuals (5). Understanding speech under challenging and noisy hearing conditions is commonly referred to as speech-in-noise (SIN) perception. SIN has been investigated using different experimental conditions of degrading the speech signal, such as speech resynthesized out of acoustic noise (6), (un) informational noise masking of speech (7, 8), or speech-in-speech perception in multitalker environments (9, 10). Recent neuroscientific studies pointed to a crucial involvement of the human auditory system, especially the auditory cortex, for auditory signal analysis in noise. Successful and unsuccessful SIN performance correlated predominantly with signal in the superior temporal cortex (STC) as part of higher-level auditory cortex (7–9, 11–13) that is sometimes more lateralized to the left hemisphere (14). Auditory cortical activity for SIN appears sometimes functionally connected with other cortical areas (11), based on a co-occurrence of activity in lateral (7, 11) and medial frontal areas (7, 11, 14) as well as in (pre)motor (15) and insula-opercular regions (13).

While SIN perception is certainly important for human interactions, it is only a specific and rather downstream process to a more fundamental problem of signal-in-noise detection. In the literature, the predominant focus of SIN is comprehension and intelligibility of speech, yet prior to SIN reaching sufficient signal-to-noise ratio (SNR) for adequate comprehension, listeners can hear speech and vocalizations at unintelligible levels. These daily life occurrences, where one can hear the voice of another individual within noise but be unable to make sense of it, are common for human communication in various environments (e.g., in noisy urban, factory, or emergency environments). The detection of speech or vocalizations at precomprehension levels as a signal-in-noise target detection problem has little, if any, representation in existing research. It is the behavioral and neural dynamics of detecting this signal that interest us, representative of the most challenging voice-in-noise (VIN) detection tasks that one would have to perform in normal listening. The phenomenon of detecting voices emerging from background noise, yet at insufficient SNR for comprehension, is a typical experience of listeners in loud and noisy acoustic environments. It is worth noting that VIN detection as we discuss it here is distinct from a categorization/discrimination task and is primarily an SNR problem that the auditory system must resolve. While there is extant research on the categorization of precomprehension vocal activity (16–18), these investigations have not addressed the topic as an SNR problem where a continuous signal is embedded within noise. The detection of auditory signals masked by energetic noise requires some form of separation of signal from noise, or formation of the perceptual object before any performance of classification or further speech.

It seems, then, that common SIN studies are thus limited to a certain degree, and this paucity of SIN research on these initial stages limits our broad understanding of the phenomenon in a number of ways. First, SIN studies so far focused overwhelmingly on comprehension of linguistic speech information as the specific information carried by voice signals. Accordingly, the detection and recognition of paralinguistic or nonverbal information simultaneously carried by voice signals presented in noise have been largely neglected, a stark omission given that paralinguistic information adds additional category information to vocal stimuli.

Second, previous research predominantly used paradigms that presented speech signals at a fixed, unchanging SNR (i.e., the relative loudness of the target does not change across the trial) for only short instances of vocal activity (single utterances of words/phonemes), such that the dynamic temporal properties and the influence of varying
SNR levels of this figure-ground separation process on auditory signal detection were neglected. There are some limitations with the ecological validity of such research, particularly regarding the use of single utterances of vocal information. Day-to-day, spoken communication typically relies on continuous vocalizations, and low-frequency information is encoded within continuous speech, such as envelope modulation, that may not be adequately represented in single utterances (19).

Speech in noise, and other signals in noise, processing is not driven by a single linear relationship with SNR and is demonstrated to be informed by temporal properties as well. In sentence length stimuli where global SNR does not increase, perceptual performance can be predicted by the spectrotemporal coherence of “glimpses” of speech signal (20–22); in addition, auditory object formation is well established to stabilize over listening time (23, 24). If VIN detection is in line with most research on perceptual decision-making, then evidence must be accrued until a “voice present” perceptual bound is reached (25). Yet, the predominant approach of using single utterances, defining detection/decision thresholds by participant “hits” and “misses,” does not consider whether a trial was missed due to SNR or lack of temporal exposure. This potentially causes incorrect estimations of perceptual performance of continuous/sentence length speech. Crucially, these comparisons largely limit us to contrasting the neural difference between “detected” and “not detected,” rather than observing a state change as no perceptual or neural state change is experimentally observed, merely two states compared.

In this study, we aimed to address these limitations by investigating the more fundamental process of detecting and recognizing dynamic voice signals hidden in noise, which we refer to as “voice-in-noise perception.” We also aimed to investigate how socially relevant information, such as affective information carried by the paralinguistic channel of speech, is dynamically recognized in VIN. This is an important consideration given the potential recruitment of additional areas in the medial temporal limbic system known for decoding socially relevant information in paralinguistic speech signal (26, 27) and may provide assistance to the auditory system for voice signal detection.

Concerning the first limitations above, we can consider VIN processing itself to be a specific subcategory within auditory scene analysis of separating the individual auditory voice signal from competing sound sources. While the underlying neural networks involved in the formation of more general objects in auditory scene analysis is well investigated (28), the neural process of auditory scene analysis for the more specific auditory object of a human voice is rather unknown. Human voices carry far more sociobiological relevant information than other auditory signals, and they have been demonstrated to recruit specific auditory areas known as the temporal voice areas located in the STC even when speech is unintelligible (29). Recruitment of higher-cognitive areas and processes then potentially further differentiates SNR from general auditory processing if necessary (30).

Next to the largely neglected issues of VIN in auditory scene analysis and the recognition of social information carried by the paralinguistic channel of speech, the second neglected issue as described above is the temporal dynamics of auditory signal-in-noise detection. The dynamic and temporal properties of speech and voice activity hold a large amount of ecological relevance. This study will use continuous stimuli that increase in SNR over the course of the presentation, allowing investigation of detection of dynamic VIN that is not possible during studies using more common staircase procedures with single utterances. A key focus of this paper is to characterize both the perceptual state change and the neural state change. As speech is dynamic and variant, it is scientifically relevant to dynamically measure responses to it. In our first experiment, we will investigate how perceptual thresholds may differ in a dynamic situation, rather than using traditional staircase threshold methods using single stimuli.

Detection of continuous VIN is underexplored within perceptual/cognitive neuroscience; to the author’s knowledge, there exists no previous research that specifically investigates the detection of VIN as an SNR problem. This presents us with a challenge for describing the phenomenon, given how close detection of a signal can be to classification tasks (16–18). Luckily, however, applications in audio technology have a useful analogy in computation voice activity detectors (VADs) (31). These VAD applications are dynamic systems that continuously sample an auditory environment, performing an “on-off” state change when they detect that a voice signal is present. The system itself does not process the voice signal or make any higher judgment other than its presence in the auditory scene. These technical applications therefore provide us with an excellent analogy for investigating this aspect of voice processing, and we are looking for evidence of a dynamic state change in response to the detection of a voice in a dynamic auditory scene. The analogy of a neural implementation of VAD also aligns with perceptual decision-making, as computational VADs continuously accrue evidence for a “voice detection” decision until reaching a decision bound, in line with the core concepts of many perceptual decision-making models (25, 31).

A primary candidate for a neural implementation of VAD and processing perceptually emergent voice signals is a network consisting of the auditory cortex and amygdala. The auditory cortex has shown robust recruitment for concurrent stream analysis and auditory-object formation (32–34), as well as recruitment during voice signal detection and discrimination tasks (35). Next to the auditory processing, analysis, and sound segregation in the auditory cortex, the sociobiological relevance of voice stimuli makes the amygdala an excellent candidate area, being involved in sociobiological detection tasks (26, 36), in processing of paralinguistic affective information (37) and in influencing auditory cortex activity during the affective analysis of sounds (38). The multifunctional role of the amygdala, and possible network connection with the auditory cortex, highlights the methodological consideration that we must focus on how recruited regions are functionally integrated for VAD. Stimulus conditions will plausibly not display functional segregation, but rather, the pathways and connections involved will differ by experimental perturbation.

Considering these recent studies, we will attempt to characterize VAD in human auditory cognition (experiment 1) as well as in the neural dynamics of the human neural auditory and limbic systems (experiment 2). In both experiments, we account for the necessarily dynamic nature of the task as well as the impact from paralinguistic affect information in emergent VIN.

**RESULTS**

**Decisional dynamics of voice and vocal affect detection in dynamic and fixed SNR conditions**

To investigate the cognitive characteristics of detecting and recognizing dynamic voice signals and affective vocal information in noise, we performed a first experiment (experiment 1, n = 26) in which we
presented continuous vocal utterances based on so-called “pseudo-
speech.” This pseudospeech resembled normal speech and followed
basic linguistic rules but was completely free of any semantic mean-
ing. These stimuli allowed focusing on the paralinguistic voice channel
as the carrier of important nonverbal social information, such as
affective information. We used 12-s recordings of these pseudospeech
utterances from six male and six female speakers, and these utter-
ances were intonated with either a neutral, angry, or joyful affective
prosody. These pseudospeeches were presented together with 14-s
recordings of pink noise with a fixed intensity of 70 dB. The pseudo-
speeches had a jittered onset of 0 to 2 s in reference to the onset
of the pink noise and were presented with log-linear increase of
intensity from −26- to −12.5-dB SNR. Participants had to press a
button with their right index finger once they start hearing a voice
hidden in the pink noise (Fig. 1A).

In this original “dynamic” experiment, we quantified the dB SNR
level from the reaction times as a proxy to the perceptual switching
point from the onset of the increase in voice amplitude. We found
that the affective category of vocal stimuli significantly influenced
the dB SNR at the time of detection \( F_{1,65,41.36} = 65.22, P < 0.001; \)
linear mixed model (LMM); fixed effect of the factor condition, with
random effect of the factor participant; \( n = 26 \). In post hoc compar-
isons (with Tukey correction), angry (\( T_{50} = 11.41, P < 0.001 \)) and
happy voices (\( T_{50} = 6.23, P < 0.001 \)) were detected significantly
with a lower dB SNR than neutral voices, and angry voices with a
lower dB SNR than joyful voices (\( T_{50} = 5.17, P < 0.001 \)). An important
note is that these dB SNR data were individually corrected for each
participant’s general response latency on the basis of an additional
experiment where we presented pure tones at fixed 70 dB and of
500-ms duration and without background noise. Participants had
to respond as quickly as they hear the tone, and the mean reaction
time in this pure tone task was used to correct the mean reaction
time in the main experiment. This helped to estimate the perceptual
switching point for the dynamic VIN appearance, which happens
before participants execute the button press.

To assess whether these reaction time differences for the dynamic
VIN detection task were specific to the perception of emergent voices
and vocal affect in noise, we performed two additional experiments
with the participants. First, the same experiment was performed but
with a “scrambled” version of the pseudospeech. This presents an
analogous, temporally dynamic SNR problem as the original stimulus
task. However, the scrambling preserves the mean SNR and frequency-
that is an important cue to vocal signal
power distributions of the original but destroys speech-specific tem-
poral envelope information that is an important cue to vocal signal
and affective voice detection \( (39) \). This scrambling thus controls for
local pitch or intensity envelope peaks in the original stimuli that
could have driven the detection of voice signals at certain time points.
Once again, stimulus dB SNR levels and thus detection times differed
significantly by category \( (F_{1,38,45.53} = 27.768, P < 0.001; \) LMM; \( n = 26 \)).
Angry \( (T_{50} = 7.43, P < 0.001) \) and joyful voices \( (T_{50} = 4.09, P = 0.005) \)
again were detected significantly with lower dB SNR than neutral,
and anger voices were detected with lower dB SNR than joyful voices
\( (T_{50} = 3.734, P = 0.044) \).

Second, we performed the “fixed” condition experiment with
presentations of single words taken from the original pseudospeech
and presented them with 500 ms of pink noise at a nonchanging
SNR, varying the SNR instead across each trial so that it resembled
the original SNR from the dynamic experiment as the mean of each
1-s bin from the 10-s period of the original experiment with in-
creasing dB SNR levels. By constraining the length of time to single
utterances, without globally increasing SNR of the target voices, we
eliminate the almost certain emergence of vocal stimuli that occurs
in the dynamic and scrambled conditions as a function of increasing
exposure time and SNR. This demonstrates any perceptual difference
between our methodology and more common uses of single-utterance
stimuli. On each trial, participants had to indicate whether they
heard a voice or not. We quantified the 50% point of a psychometric
fitting of the data across the 10-dB SNR conditions for each partici-
pants. The 50% dB SNR level across participants was significantly
different between conditions for the detection thresholds \( (F_{1,65,41.17} =
26.24, P < 0.001; \) LMM; \( n = 25 \) ) driven by significant contrasts be-
 tween angry and both neutral \( (T_{50} = 6.91, P < 0.001) \) and joyful trials
\( (T_{50} = 5.33, P < 0.001) \), while joyful and neutral trials did not signifi-
cantly differ \( (T_{50} = 1.58, P = 0.263) \).

Last, we compared the dB SNR levels across all three experiments.
Since for the fixed experiment we did not analyze mean dB SNR
levels (i.e., mean across all trials) but rather quantified the 50% deci-
sion point after a psychometric fitting (i.e., a single point value),
we entered data from dynamic and scrambled studies as mean re-
response times per participant and as single value for the fixed ex-
periment. Analysis of the data across the three experimental parts showed
significant effects on the dB SNR level due to the experiment factor
\( (F_{1,37,34.24} = 31.81, P < 0.001; \) LMM; \( n = 26 \) ) and to the condition
factor \( (F_{1,83,65.64} = 63.34, P < 0.001) \). Concerning the experiment factor,
significantly lower dB SNR levels were found in the dynamic condi-
tion compared with fixed \( (T_{50} = 5.63, P < 0.001) \) and scrambled

\[ A \text{ Dynamic: Experimental conditions} \]

\[ \text{Pink noise} \rightarrow \text{Angry voice} \rightarrow \text{Decision} \rightarrow \text{Pink noise} \rightarrow \text{Neutral voice} \rightarrow \text{Decision} \]

\[ \text{Fixed: Experimental conditions} \]

\[ \text{Pink noise} \rightarrow \text{Affective voice} \rightarrow \text{Decision} \rightarrow \text{Pink noise} \rightarrow \text{Affective voice} \rightarrow \text{Decision} \]

\[ \text{B Behavioral performance} \]

\[ \text{Experiment 1} \rightarrow \text{Fixed Dynamic} \rightarrow \text{Experiment 2} \rightarrow \text{Scrambled Fixed} \]

\[ \text{SNR} \rightarrow \text{dB SNR} \rightarrow \text{SNR} \rightarrow \text{dB SNR} \rightarrow \text{SNR} \rightarrow \text{dB SNR} \]

\[ \text{Fig. 1. Experimental setup and behavioral data.} \]

(A) Experiment 1 included three different parts. The first part (top) included the original dynamic vocal pseudospeech
utterances (neutral, anger, and joy; blue amplitude curve) that were presented with temporally increasing intensity in fixed-intensity pink noise of 14-s duration (gray).
Vocal utterances were presented with a jitter onset of 0 to 2 s from the onset of the pink noise. Participants were asked to press a button (decision) once they hear a voice
masked by the pink noise. The second part of the experiment presented the stimuli in scrambled versions. The third part was a fixed experiment, where single pseudowords
were presented at fixed SNR in pink noise, varying SNR across trials (bottom). (B) The detection latencies from the voice onset were quantified in dB SNR levels for exper-
iment 1 (left) and for the behavioral data of the fMRI experiment (experiment 2; right). Neu, neutral; Ang, anger; Joy.
Dynamic VIN detection supported by a neural pattern switch

In experiment 1, we established that humans can reliably detect dynamic voice signals in noise with a dB SNR advantage against short-utterance VIN detection at fixed SNR and that affective category is perceptually relevant to the speed of this detection. On the basis of this evidence, we performed a second experiment (experiment 2) wherein we presented the same neutral, angry, and joyful pseudospeech utterances to another sample of participants (n = 22) while recording their brain activity using functional magnetic resonance imaging (fMRI). This experiment was performed to investigate the neural dynamics especially in the human auditory system and the limbic system underlying the perceptual switch at the time point when participants start to hear voices that slowly appear in background noise. In experiment 2, we used a specific spectral profile for the background noise that perceptually appeared as pink noise when combined together with the scanner noise. Participants again had to press a button as soon as they were confident to hear a voice. Reaction times again were corrected by the mean response latencies for each participant as quantified by a separate pure tone experiment. Participants responded differently to the stimulus conditions (F1,70,35,74 = 19.32, P < 0.001; LMM; n = 22), with significantly lower dB SNR for anger voices compared with neutral (T42 = 5.92, P < 0.001) and joyful voices (T42 = 4.60, P = 0.001). Joyful and neutral voices did not differ significantly (T42 = 1.32, P = 0.394) (Fig. 1B), the lack of significant contrasts compared to the behavioral study likely owing to the more challenging environment of the fMRI scanner causing more trial-on-trial response variability, reducing significance between conditions. It should be noted that the overall profile of the results is similar to the behavioral data from the dynamic part of experiment 1, but at a much higher SNR (Table 1).

Table 1. Summary of the behavioral data from experiment 1. Mean response times (RT; SEM in brackets) for each experimental condition (neutral, anger, and joy) and approximate dB SNR values as converted from the response times. Data are shown for (A) the dynamic presentation of the original stimuli, (B) the scrambled condition, and (C) the fixed experimental part.

| Conditions            | RT (SEM) | dB SNR (SEM) |
|-----------------------|----------|--------------|
| (A) Dynamic original stimuli |          |              |
| Neutral               | 7.21 s (0.37 s) | −16.27 (0.50) |
| Anger                 | 6.30 s (0.39 s) | −17.50 (0.47) |
| Joy                   | 6.71 s (0.33 s) | −16.94 (0.44) |
| (B) Scrambled stimuli |          |              |
| Neutral               | 9.34 s (0.39 s) | −13.39 (0.53) |
| Anger                 | 8.39 s (0.36 s) | −14.67 (0.48) |
| Joy                   | 8.81 s (0.37 s) | −14.12 (0.50) |
| (C) Fixed experiment  |          |              |
| Neutral               | −         | −15.00 (0.63) |
| Anger                 | −         | −17.30 (0.50) |
| Joy                   | −         | −15.40 (0.50) |

We first analyzed data from a separate functional voice localizer scan to determine the voice-sensitive regions in the auditory cortex (29). This revealed extensive bilateral clusters of activation in the STC covering primary, secondary, and mainly higher-level areas of the auditory cortex (Fig. 2A). Second, we compared the neural activity for the main dynamic experiment for the predecision against the postdecision phase, and vice versa. This analysis was locked to the perceptual switching point (i.e., corrected for the mean response latency of each participant). In the predecision phase, we found significant cortical activity in bilateral insula (MNIxyz; left [34 20 −8], T = 7.53; right [−30 16 −6], T = 6.97) as well as subcortical activity in auditory structures, such as the left medial geniculate nucleus ([−12 −24 0], T = 8.86) and bilateral inferior colliculi (IC; left [−2 −32 −12], T = 8.26; right [2 −34 −12], T = 8.21) (Fig. 2B). In the postdecision phase, we found significant activity in the bilateral auditory cortex with a peak activity in Te3 as part of the STC (left [−60 −14 2], T = 13.09; right [62 0 −8], T = 12.05) as well as activity in bilateral amygdala (left [−18 −8 −20], T = 7.90; right [26 −2 −14], T = 8.67) (Fig. 2C). This activity was similarly found for the same contrasts performed for each affective voice condition, namely, for neutral (left STC [−60 −14 2], T = 7.34; right STC [62 −8 0], T = 6.33; left amygdala [−18 −6 −20], T = 4.12; right amygdala [24 −2 −16], T = 5.28), angry (left STC [−60 −6 −6], T = 7.49; right STC [62 −2 8], T = 8.68; left amygdala [−18 −8 −20], T = 4.84; right amygdala [24 −4 −16], T = 5.21), and joyful voices (left STC [−60 −14 2], T = 7.96; right STC [60 8 −12], T = 6.57; left amygdala [−18 −10 −18], T = 5.26; right amygdala [20 −2 −18], T = 4.76).

Limbic-cortical network connections drive voice from noise separation

We also conducted a directional neural network analysis between the neural regions that were active during the postdecision phase. Since we found activity in bilateral amygdala and STC in this postdecision phase, we wanted to investigate the neural information flow between these regions. We used dynamic causal modeling (DCM), which estimates the most likely driving input to the neural network (C matrix), the endogenous connectivity between regions unmodulated by the experimental conditions (A matrix), and the modulatory influence of the experimental condition on the effective connections (B matrix). We first determined the most likely input model that was providing driving input to the STC-amygdala network as in a previous report (38).

To determine this driving input (C matrix), we fixed the endogenous connectivity between regions as bilateral connections between any of the regions (A matrix). By permuting the model space through any possible bilateral input of all voice trials (neutral, anger, and joy) or of only the affective voice trials (anger and joy) either to the STC, the amygdala, or both regions, we determined the winning model as having driving input of the affective voices to the amygdala, and of all voice trials to the STC (Fig. 3). We then used this winning input model configuration as fixed parameters to permute through the model space of the modulations of connections (B matrix), with the possibility that connections from and between the STC are modulated by the all-voice condition, whereas connections from and between...
the amygdala are modulated by the affective voice conditions. The winning model turned out to include any modulation of connections and the bilateral amygdala-STC network (Fig. 3).

DISCUSSION

In this study, we aimed to define the cognitive and neural characteristics of the dynamically emergent and near-threshold voice signal detection in a dynamic VIN perception task that we supposed to be fundamental to more specific SIN tasks. We accordingly used multiple complimentary perceptual tasks as well as functional neuroimaging in humans to construct the broadest possible understanding of VIN as a supposedly neurocomputational implementation of a VAD-like process and also to specifically determine how affective voice intonations modulate these processes. We found that, first, dynamic VIN on original neutral and affective voices shows an advantage over rather fixed or acoustically scrambled version of VIN, with dynamic affective voices being the least vulnerable to noise masking. Second, VIN is neurally supported by a network change, as indicated by a
change from subcortical auditory and insular activity before detection, to an auditory cortical and amygdala network after voices have been reliably detected in noise. This neural switch is also indicated by the specific effective connectivity demonstrating an integrated coupling between the amygdala and the auditory cortex for an affective and voice information exchange, respectively. The network analysis seems to imply that this neural switch has nested effects driven by the emotional content of the voice, given that specific modulatory activity is present only during affective trials.

This detection advantage for affective compared to neutral voices has been first confirmed by the decisional dynamics and corresponding SNRs (dB SNR) in experiment 1. The data show that VIN detection can be accomplished by the human cognitive system at relatively low levels of SNR when presented dynamically as emergent voice signals appearing in noise. This detection advantage in the dynamic case is probably significant against previous evidence resulting from experiments using a fixed paradigm (7, 8), but a precise comparison to previous reports is not possible. In addition to this general, dynamic VIN detection, affective voices have an additional advantage for being detected at even lower SNRs compared to neutral voices. Because all stimuli were normalized in their intensity, the detection advantage for affective voices seems mainly driven by their emotional nature that is reflected in certain acoustic features and feature variations (40). These distinct acoustic features might enhance the salience of affective voices even in noisy conditions, and affective communication even in noisy conditions seems of social and evolutionary advantage (41).

This VIN detection advantage for affective voices was confirmed by a comparison to scrambled neutral and affective voices. The significantly higher SNR detection level was probably the result of destroyed temporal modulations (i.e., spectral and amplitude modulation rates and depth), which are a strong cue for voice signals and affective voice information (42). Furthermore, dynamic voice signals were detected at lower SNRs compared with fixed voice signals, at least for neutral and joyful voices. A possible attribution of this detection advantage is that our dynamic experiment provides the optimal conditions for evidence accumulation for target detection.

As stated in Introduction, perceptual performance can be informed by a number of speech characteristics that are driven, at least in part, by length of exposure. Notably, glimpsing studies show that naturally occurring, local SNR variations in speech can accurately predict perceptual performance and that specific spectrotemporal regions of speech are more informative than others (21, 22). In any theoretical VIN trial where the target is missed, it cannot be disentangled whether the miss was due to insufficient time allowing relevant regions to be revealed to the listener, or whether global SNR was insufficient to ever reveal those regions. By increasing the SNR over sentence length stimuli, we have provided advantageous conditions for detection compared to our staircase procedure with single utterances.

To our knowledge, the exact relationship between exposure length and global SNR has not been fully characterized. In addition, when taken within the context of decision-making literature, and the core concept of a percept reaching a decisional bound, this relationship between favorable SNR for detection and relevant information for evidence accumulation has not been accounted for. This makes explicitly applying decision models to our finding somewhat challenging, yet these behavioral differences clearly reveal a perceptual advantage for the auditory system for detecting continuous stimuli, rather than single utterances, supported by current work in auditory stream segregation demonstrating that streaming is a result of evidence accumulation and positively correlated with stimuli length.

The relationship between SNR and presentation time is clearly a complex one. Perceptual performance for angry voices was similar in the dynamic and fixed conditions, likely because angry voices are acoustically very distinct (42) and powerful stimuli to elicit brain responses (39), which might facilitate their detection even in fixed conditions of VIN. The glimpsing framework may offer an explanation as having many more coherent spectrotemporal regions at favorable SNR to the pink noise, thus providing more perceivable, informative signal for detection. Summarizing the data from experiment 1, we show that the commonly accepted method of presenting single instances of SIN returns differing results to more ecologically valid stimuli presentation that reflects the continuous nature of vocal utterances.

Having established that VIN is more effective in dynamic compared to fixed conditions of voice signal processing, we investigated the neural dynamics of dynamic VIN in experiment 2. The perceptual switching point was marked by a neural switch from higher activity in bilateral insula as well as subcortical structures of the ascending auditory pathway [IC and medial geniculate body (MGB)] in the predecision phase, to higher activity in a cortical auditory-amygdala network in the postdecision phase. Neural activity in the IC and MGB might continuously evaluate incoming auditory sensory information for socially relevant information, such as voice signals, given the properties of both regions to perform a complex spectrotemporal analysis of incoming sensory information (43). Research directly concerning the IC and MGB in auditory processing has shown these areas to be responsive to auditory deviance detection (44) and facilitating task-relevant, speech recognition during a dynamic task (45), respectively. As deviance detection is additionally a dynamic task requiring constant evaluation, it is plausible that the detection advantages that we see in the dynamic condition may be predicated on recruitment of these areas; however, our study was not designed to explicitly answer this question in an adequate manner. Further, given the close link of both nodes to the limbic system, they might report emerging voice signals and affective voice information to the limbic system, and especially the amygdala (27), which is then critically relevant for the postdecision phase. The insula is the common brain node for social emotion processing and interceptive awareness of emotion (46) and therefore may track internal states that lead to the conscious presence of voice signals in the neuro-cognitive system.

After the perceptual and neural switching point, we found predominant involvement of bilateral STC extending over several auditory cortical subregions (primary, secondary, and higher-level auditory cortices) as well as of the bilateral amygdala. Activation in STC to conscious voice processing is highly likely once voice signals start to appear in noise, since previous studies have shown STC activity in response to voice signals (29) and vocal affect (26, 37). This confirms that the perceptual switch in dynamic VIN elicits critical activity in STC that is located within the independently defined voice-sensitive cortex. While activation in the amygdala to affective voices is highly expected (37, 47, 48), it cannot be solely attributed to the affective content of stimuli since we also found it for VIN of neutral voices. The amygdala thus seems likely being more generally recruited for the detection of previously undetected, socially relevant stimuli (49), such as voice signal detection hidden in noise.
Although the contrast results show that the amygdala is involved generally in VIN processing and detection, we also show effective connectivity evidence that the amygdala plays a selective role when processing affective rather than nonaffective voices. The permutation through all possible input models of affective/nonaffective stimuli to bilateral amygdala and STC showed that a model with affective trials (anger and joy) being a driving input to the amygdala, while all trials are input to the STC, best explains the driving input to this neural network. Thus, affective rather than neutral voices were likely to provide significant input information to the amygdala (37, 48) based on full connectivity between all amygdala and STC regions. By further permuting through a number of possible modulatory influences from experimental conditions to the connections between brain regions, the model showed bilateral, reciprocal modulation of ipsi- and contralateral amygdala-STC connections, notably that forward connection modulation from the amygdala to the STC is present during affective trials only. The model also showed additional modulations within bilateral amygdala and STC for affective and all postdetection voices, respectively. Assuming that participants did not consciously perceive voice signals in this phase, these modulatory connections may indicate the neural monitoring and updating of auditory sensory information (50) that is responsive to affect even at threshold detection levels. In addition, it is unlikely that the results of our model are influenced by any possible preparation of auditory or affective systems to likely upcoming emotional stimuli, despite most of the stimuli being non-neutral. Such an effect, if present, would affect the predetection phase; however, our model is determined by postdetection neural activity and only compared against prephase to show that this has changed as a result of stimulus detection.

Thus, VIN especially for affective voice modulates the connectivity between bilateral limbic areas and their connection to the STC, which together might facilitate their detection in noise. This might further support our finding for the behavioral data, which show an SNR advantage for affective voices. For the effective connectivity modeling, we further found that the connections between bilateral STC and from the STC to the amygdala were driven by general voice signal processing. The STC thus might share voice signal and voice acoustic information with the amygdala for general VIN, and the amygdala, in turn, shares affective information with the STC (38) for the specific case of VIN for affective voices. This modulation of connections between bilateral STC and from STC to the amygdala was critically also found in the predecision phase. Assuming that participants did not consciously perceive voice signals in the predecision phase, this STC-STC and STC-amygdala activity might represent a neural monitoring and updating activity of auditory sensory information (51). Given that both the amygdala and the STC showed significantly lower activity in the predecision phase, the neural activity centered on the STC may increase after the perceptual switching point, which might be facilitated by additional support by the amygdala to the neural network at and after the perceptual turning point (37, 38, 48).

The neural network for VIN processing and detection, particularly in cases of affective voices, seems thus to have critical inclusion of amygdala with reciprocal connections to the STC as the central neural node for general auditory processing and also specifically for voice signal processing. For the STC in the connectivity analysis, we used a specific peak in higher-level STC located in Te3 as part of the STC. This STC subregion is central to processing voice signals (29) and affective voices (26, 37, 52, 53), but it is only a subpart of the STC. Furthermore, cortical voice processing usually extends across several subregions including primary, secondary, and higher-level STC (29). The central interpretation of these findings is that the auditory system actively influences the STC-limbic network in both pre- and postdetection phases and that the influence of the auditory system is augmented with dynamic, stimuli-dependent, and reciprocal modulations from the limbic areas in response to and after vocal activity detection. While a reciprocal STC-amygdala relationship was previously reported during clear perception of short affective sounds in a non-noisy and nondynamic context (38), here we report a critical neural network switch to an STC-amygdala network at near-threshold levels of voice signal and vocal affect detection indicative of a change from continuous scene evaluation to target detection. This important relationship has not yet been described in previous auditory perceptual research. In particular, our findings are the first to characterize a state change in the auditory system in response to the threshold detection of VIN.

In sum, our data show that the capabilities of the human auditory and neural system to detect communication and voice signals in noise might have been underestimated in previous research using rather fixed SNR protocols to investigate SIN and sound detection in noise. A dynamic SNR context is common to many daily life hearing conditions, and humans show superior voice signal detection in this dynamic noise and SNR contexts, especially for voice signals carrying socio-affective meaning given their relevance for any social interaction. Dynamic SNR contexts allow for subthreshold evidence accrual before deciding about the presence of voice signals hidden in noise. This switch from sensory evidence accrual to voice signal detection was reflected by a switch from subcortical auditory processing to an integrated auditory-limbic neural network dynamic. We hope that these findings provide the groundwork for future work building on VIN processing and demonstrate the utility of investigating auditory processing as a continuous task to fully observe, and make inferences upon, the dynamic change that neural systems undergo.

**METHODS**

As described in Results, we performed two experiments (experiments 1 and 2) with independent sample groups of human participants. All participants self-reported normal hearing and normal or corrected-to-normal vision. No participants presented any atypical neurological or psychiatric history, and all participants gave written informed consent in accordance with the ethical and data security guidelines of the University of Zurich. The experiments were approved by the cantonal ethics committee of the Swiss canton Zurich. All experiments were conducted in accordance with the ethical guidelines of the Swiss canton Zurich.

**Experiment 1: Decisional patterns of VIN detection**

**Participants**

Twenty-six participants took part in this experiment (15 females; mean age, 26.73 years; SD, 4.05). The experiment was divided into three different parts, referred to as dynamic experiment, scrambled experiment, and fixed experiment. One participant was removed from the data after incorrectly performing the experiment.

**Stimuli**

We used pseudospeech samples of 12-s duration consisting of 36 pseudospeech utterances spoken by 12 speakers (6 male, 6 female) in neutral, angry, and joyful intonation. This pseudospeech utterances
were based on speech-like verbal material that did not violate any linguistic rules but which were completely free of any semantic meaning (e.g., “Nikalibam sud molen kudsemina lod belam ...”). A 14-s sample of masking pink noise was pregenerated, and voice and noise stimuli were normalized, pre-experiment, to have a continuous intensity rate of ~70 dB in a sliding window across each stimulus. Stimulus length of 12 s was chosen as we aimed to guarantee detection for all trials, but not at such a rapid rate that we could not reliably measure equivalent SNR at detection time.

All experiments were conducted in an anechoic chamber. Stimuli were presented over Sennheiser HD 280 Pro headphones, using an external RME Fireface UC soundcard.

**Experimental setup**

Across all three experimental parts (dynamic, scrambled, and fixed), participants were tasked with detecting VIN of three affective categories (neutral, angry, and joyful). Participants in all situations were instructed to indicate by button press with the right index finger the first moment they could detect a voice, even if they could not understand the pseudolinguistic carrier of the vocal utterances. At no point were participants asked to identify the affective category nor were they informed of any affective aspect to the study.

In the dynamic condition, target stimuli were continuously spoken utterances of the pseudospeech as described above. The 12-s samples were presented simultaneously with masking pink noise, increasing on a linear dB scale from −26-dB SNR to −12.5-dB SNR (relative to the 70-dB normalized pink noise) over the course of 10 s, with the linear increase starting at either 0, 1, or 2 s into the stimulus. The onset of the pseudospeech was jittered randomly to start at either 0, 1, or 2 s after the onset of the pink noise that obscured the pseudospeech samples. Participants were presented with six runs each containing 21 pseudospeech trials, plus 3 catch trials that did not contain any pseudospeech. Trials were separated by an inter-stimulus-interval (ISI) of 4- to 7-s duration. Each pseudospeech sample was repeated three times throughout the dynamic experiment. Blocks were balanced for presentations of affect and speaker gender.

For the scrambled experimental part, we used the same stimuli as in the dynamic part, but the original stimuli were scrambled. For this scrambling, stimuli were rearranged into 25-ms segments at 12.5-ms steps; thus, all windows have 50% overlap but stimulus length is the same. Hanning windows were applied to each segment to reduce audio artifacts. This process eliminates linguistic-like and envelope-related information conveyed by amplitude modulation while retaining the overall power and frequency distribution of the stimuli. All other experimental and presentation settings were identical to the dynamic experimental part.

The fixed experimental part presented 500-ms bursts of noise containing words cut from the full-length pseudospeech samples. Trial presentations where at 10 different SNRs, linearly spaced between −26- and −12.5-dB SNR as resembling the mean SNR from the 1-s time bins of the original 10-s dB SNR increase. Relative SNR of the VIN did not change across the course of the stimulus presentations. The fixed stimulus was composed of two pseudowords (“belam” and “namil”) extracted from all the neutral, angry, and happy pseudospeech utterances of two of the speakers (1 male and 1 female). The fixed condition presented a single block of 240 pseudoword trials and 48 catch trials containing no pseudoword. Trials were separated by an ISI of 2- to 4-s duration. No instructions were given again regarding the affective content of the target voice. Rather than taking mean response times like the first two experiments, participant thresholds from the fixed condition were derived by fitting psychometric curves per participant, per condition. Using the Curve Fitting Toolbox v3.5.8 (MATLAB 2018a), participant responses were fit to a cumulative Gaussian distribution function using robust linear least-squares fitting with bisquare weighting. The returned parameter indicated the dB SNR for each participant’s 50% detection for each condition. Statistical analysis for the fixed condition was conducted on these derived thresholds and fed into the model below.

**Pure tone reaction time task**

Participants’ general mean reaction times were estimated separately from the other three experimental parts, and these estimates were used to correct reaction times for the three main experimental parts as to estimate the perceptual switching point that is unaffected by individual reaction times. For this estimation of individual reaction times, we presented pure tones (100, 200, 300, 400, and 500 Hz) of 500-ms duration and asked participants to respond with a button press as soon as they heard the tones. Each pure tone was repeated 20 times, resulting in a total of 100 trials. Tones were separated by an ISI of 2- to 4-s duration. Mean reaction times were scored as the mean across all 100 trials.

**Data analysis**

The data were analyzed using R statistics software (v3.5.3), and data from all behavioral experiments (including behavioral results from the fMRI study) were analyzed using the general LMM equation of

\[
\text{Value} \sim \text{Condition} + (\text{Condition} | \text{Participant})
\]

where “value” is dB SNR of the perceptual switching point and “condition” refers to the three affective stimuli (neutral, angry, and joy). Null responses, or detections at implausibly fast times (<2.5 s into a trial), were not included in the data analysis. Analysis of the fixed condition was conducted in a traditional staircase procedure, and hence, psychometric functions where fitted for each condition per participant. Participant threshold SNRs for each condition where then averaged, and the averages were analyzed in the above model. Pairwise comparisons of individual contrasts were performed using Tukey correction.

**Experiment 2: Neural dynamics of VIN detection**

**Participants**

Functional brain imaging data were recorded from 22 participants (14 females; mean age, 23.63; SD, 3.74).

**Experimental setup**

The experimental setup was identical to the dynamic condition as in experiment 1. Referring back to our analogous investigation of computation VAD, we investigated the dynamic state change of a network. This analogy demonstrates the importance of using continuous stimuli, rather traditional staircase methods that are more common, as we are obliged to continuously sample a network to make claims of its state change. Staircase procedures with single trials that are either hits or misses only allow us to describe the difference between conditions, rather than state change.

The behavioral task was also identical, with the exceptions that the masking pink noise was notch filtered to account for scanner noise amplifying masking in a certain frequency band and that the overall SNR range was raised at presentation to around −7.5 to 6 dB SNR. Because of the varying nature of noise canceling, presentation level had to be adjusted in some instances to ensure that participants routinely answered approximately in the middle of the trial. Stimulus
presentations were conducted using the same computer and audio setup in the behavioral study. Stimuli were presented via Optoacoustics OptoActive II active noise canceling headphones with a noise cancellation of about ~20 dB, allowing comfortable background noise levels without obstructing the ear canal. Each imaging session consisted of the same six blocks of 21 pseudospeech trials and 3 catch trials. Block and stimulus balancing is the same as in the dynamic behavioral experiment. The fixed experiment was not pushed through to the fMRI experiment because of invalidity in appropriately comparing 500-ms, fixed trials where voice was not detected to the activity across ~4 to 6 s of predetection speech. Not only does this present methodological concerns for fMRI analysis, but it is also unknown whether a missed trial is equivalent to “predetection” and is just too short or whether the trial is truly too low in SNR to ever be detected. It should be noted that while we did not see a significant, pairwise difference between all affective conditions in the behavioral results of the fMRI study, the overall profile of the results is similar to the original behavioral data but at a much higher SNR. The lack of significant difference is likely caused by the fMRI environment introducing greater trial-on-trial variability.

**Pure tone reaction time task**

See experiment 1 for the settings. This pure tone reaction was performed after the fMRI acquisition for the main experiment.

**Functional voice localizer scan**

To identify human voice–sensitive regions in the bilateral superior cortex, sound clips of 8-s length from an existing database were used (29). These sound files consisted of 20 vocal sounds and 20 nonvocal sounds. Participants were instructed to listen passively to the stimuli.

**Brain data acquisition**

Images were acquired using a Philips Ingenia 3T scanner using a standard 32-channel head coil. High-resolution structural scans were acquired using a T1-weighted sequence [301 contiguous 1.2-mm slices; repetition time (TR)/echo time (TE) = 1.96 s/3.71 ms; field of view = 250 mm; in-plane resolution, 1 × 1 mm] Experimental and voice localizer functional images were recorded using a multislice band sequence (SENSE factor 3) based on a T2*-weighted echo-planar imaging (EPI) sequence (TR = 330 ms; TE = 30 ms; flip angle = 41°; in-plane resolution, 220 × 220 mm; voxel size, 1.72 × 1.72 × 5 mm; gap, 0.3 mm; 12 slices), covering most parts of the temporal cortex inducing the auditory cortex and the amygdala. This partial volume acquisition was chosen to optimize the TR to an optimal temporal resolution and to mainly cover the two major regions of interest in this study (auditory cortex and amygdala). Two additional whole-brain EPIs were acquired for the purpose of coregistration with the partial volume acquisitions (TR = 1 s; TE = 30 ms; 220 × 159 × 220 mm; matrix 72 × 69; 30 descending slices). Respiration and pulse were also acquired in each participant to be used later to correct for nuisance artifacts.

**Data analysis**

The statistical parametric mapping software (version SPM12, fil.ion.ucl.ac.uk/spm/software/spm12) was used for the preprocessing and analysis of the functional brain data. Functional images were realigned and coregistered to the anatomical image; for the main experiment, the partial volume images were first coregistered to the whole-brain EPI images. The realigned functional images were spatially normalized to the Montreal Neurological Institute (MNI) stereotactic template brain using the segmentation procedure implemented in the Computational Anatomy Toolbox (CAT12; www.neuro.uni-jena.de/cat/). Normalized images were spatially smoothed using an isotropic Gaussian kernel with a full width at half maximum of 8 mm.

A general linear model was used for the first-level statistical analyses, including boxcar functions defined by the onset and duration of the pre- and postdecision phase for each of the three conditions (neutral, anger, and joy). The onset of the postdecision phase was defined by the button press minus the mean reaction time for each participant from the separate pure tone reaction time experiment. These boxcar functions were convolved with a canonical hemodynamic response function. For the functional voice localizer scan, we modeled vocal and nonvocal trials in separate regressors. Each model included six additional regressors of no interest based on motion estimates to account for motion artifacts and additional physiological regressors of modeling breathing and heart rate (TAPAS toolbox; https://www.tnu.ethz.ch/en/software/tapas). The general linear model (GLM) was estimated for each subject, and the contrasts from each condition were assessed with a second-level group analysis. Regressors included the six conditions (pre- and postdetection for neutral, angry, and joyful trials) and a regressor for each participant. Functional contrasts between conditions were thresholded at $P < 0.05$ on the basis of a false discovery rate (FDR) correction and a cluster extent of $k = 20$.

**Dynamic causal modeling**

DCM allows us to make inferences of forward and/or backward connections between various regions of interest active during our experimental conditions. Given the strong activation in the STC and amygdala connected to the detection of stimuli, we aimed to investigate characteristics of effective connectivity between these regions. To do so, we implemented a two-stage model selection routine to investigate possible input and modulation conditions of the connectivity model. It should be noted that a strength of DCM compared to traditional contrast analyses is that it affords us inferences on the functional integration of regions that do not display functional segregation, allowing us, in essence, to claim that our conditions elicit differential “pathway” activation, rather than differential regional activation (54); hence, the DCM network was defined from the activity maps of post- minus predetection of all stimuli. This decision was made a priori, under the expectation that contrasts between affective conditions within regions of interest after detection would (i) fail to show significant differences as the amygdala is often recruited during social relevance/novelty detection as it is during affective processing (36, 49, 53), and (ii) informing or constraining the network characteristics by post hoc effects of stimulus affect (itself a nested condition for network modulation and inputs) may be considered somewhat circular analysis, biasing potential findings. In the interests of completion, post hoc beta-weight comparisons between regions can be seen in Fig. 2; contrasts between emotions after detection did not reveal significant clusters.

We implemented our DCM (version DCM12.5) in SPM12. The seeding for our DCMs was based on the group level peak coordinates in bilateral amygdalae and STC from the GLM analysis above (MN1xyz; amygdala [−18−8−20] and [26−2−14]; STC [−60−14 2] and [62 0−8]), with a 3-mm sphere around each coordinate (19 voxels per sphere) to extract the time course in these region as a volume of interest (i.e., first eigenvariate). All DCMs were deterministic, bilinear, and single-state models. Model comparisons where assessed using a Bayesian model selection approach with fixed-effects inferences (FFX) as similarly done in previous studies (51). The BMS approach determines the most likely DCM model across all participants given the data and determines the likelihood of driving inputs, effective connections, and modulations of connections.
In our two-step method, we first assessed multiple stimuli-input models, followed by possible modulation models. Endogenous connectivity between bilateral amygdala and STC is assumed, building on the findings of Kumar and colleagues (38), who previously used DCM to investigate affective sound processing in these areas. In the first step, we aimed to determine the most likely input model (C matrix). For this purpose, we fixed the endogenous connectivity matrix (A matrix), including bidirectional connections between all regions; no modulation of connection by experimental condition (B matrix) was included. Our assumption for driving inputs was that the experimental conditions would generally or selectively drive activity in the STC and amygdala nodes, and we used inputs either consisting of all voices (neutral, anger, and joy) or only affective voices (anger and joy). It should be explicitly noted that we did not constrain any parameters by stimulus condition in either the first or second step, allowing the model space to contain models that would not show an effective of stimuli emotion, although previous work indicates that modulatory connections from the amygdala in response to stimuli affect were likely (38). We permuted through any possible input configuration, with the only restriction that homolog, left-right regions were identical, resulting in \( n = 8 \) possible models (full details shown in fig. S1A). The winning input model was determined on the basis of the log Bayes factor (i.e., relative log evidence for each model) posterior probability of each model. For the latter, a posterior probability of \( P > 95\% \) is commonly regarded as informative (56).

In the second step of the DCM, we took this winning input model as a constraint while permuting through \( n = 12 \) possible modulation models by changing the parameters of the B matrix that defines the modulation of connections by experimental conditions (fig. S1B); the A matrix settings were the same as reported above. These models were selected on an informed basis according to previous studies (38, 51) and on four principle rationales: First, any of the modulatory influence on connections is symmetric across the hemispheres since we did not expect major hemispheric differences; second, connections between the bilateral amygdala and STC could reflect a forward connection from the STC to the amygdala, a backward connection, or bidirectional connections; third, bilateral connections between the amygdala as well as between the STC were included in any possible combination; and fourth, the modulation of connections was set to be the experimental condition that was the driving input to the neural node as the origin of the connection, such that connections originating from the STC were allowed to be modulated by all voice trials, whereas connections originating from the amygdala were allowed to be modulated only by the affective voice trials.

In a third step, we repeated the same DCM of the modulation of connections as in step 2, but included the functional data from the predecision phase. We permuted through the same \( n = 12 \) model space to assess whether the winning model for the postdecision phase could be or not be detected in the predecision phase.

**SUPPLEMENTARY MATERIALS**

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/eabb3884/DC1

**REFERENCES AND NOTES**

1. S. K. Scott, S. Rosen, C. P. Beamman, J. P. Davis, R. J. S. Wise, The neural processing of masked speech: Evidence for different mechanisms in the left and right temporal lobes. *J. Acoust. Soc. Am.* **125**, 1737–1743 (2009).

2. S. Kumar, K. E. Stephan, J. D. Warren, J. K. Friston, T. D. Griffiths, Hierarchical processing of auditory objects in humans. *PLOS Comput. Biol.* **3**, e100 (2007).

3. T. D. Griffiths, J. D. Warren, What is an auditory object? *Nat. Rev. Neurosci.* **5**, 887–892 (2004).

4. J. H. Hwang, C.-W. Li, C.-W. Wu, J.-H. Chen, T.-C. Liu, Aging effects on the activation of the auditory cortex during binaural speech listening in white noise: An fMRI study. *Audiol. Neurotol.* **12**, 285–294 (2007).

5. S. K. Scott, C. McGettigan, The neural processing of masked speech. *Hear. Res.* **303**, 58–66 (2013).

6. S. Evans, J. S. Kyong, S. Rosen, N. Golestanl, J. E. Warren, C. McGettigan, J. Mourão-Miranda, R. J. S. Wise, S. K. Scott, The pathways for intelligible speech: Multivariate and univariate perspectives. *Cereb. Cortex* **24**, 2350–2361 (2014).

7. S. K. Scott, S. Rosen, L. Wickham, R. J. S. Wise, A position emission tomography study of the neural basis of informational and energetic masking effects in speech perception. *J. Acoust. Soc. Am.* **115**, 813–821 (2004).

8. J.-H. Hwang, C.-W. Wu, J.-H. Chen, T.-C. Liu, The effects of masking on the activation of auditory-associated cortex during speech listening in white noise. *Acta Otolaryngol.* **126**, 916–920 (2006).

9. N. Mesgarani, E. F. Chang, Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485**, 233–236 (2012).

10. A. W. Bronkhorst, The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Atten. Percept. Psychophys.* **77**, 1465–1487 (2015).

11. J. Obleser, R. J. S. Wise, M. A. Drisen, S. K. Scott, Functional integration across brain regions improves speech perception under adverse listening conditions. *J. Neurosci.* **27**, 2283–2289 (2007).

12. A. Vouloumanos, K. A. Kiehl, J. F. Werker, P. F. Liddle, Detection of sounds in the auditory stream: Event-related fMRI evidence for differential activation to speech and nonspeech. *J. Cogn. Neurosci.* **13**, 994–1005 (2001).

13. P. C. M. Wong, A. K. Uppunda, T. B. Parrish, S. Dhar, Cortical mechanisms of speech perception in noise. *J. Speech Lang. Hear. Res.* **51**, 1026–1041 (2008).

14. R. J. Salvi, A. H. Lockwood, R. D. Frisina, M. L. Coad, D. S. Welsh, D. Frisina, PET imaging of the normal human auditory system: Responses to speech in quiet and in background noise. *Hear. Res.* **170**, 96–106 (2002).

15. M. R. Schommer, F. Pulvermüller, Is the somatosensory cortex relevant for speech perception and understanding? An integrative review. *Front. Hum. Neurosci.* **10**, 435 (2016).

16. E. Bigand, C. Delbé, Y. Gérard, B. Tillmann, Categorization of extremely brief auditory stimuli: Domain-specific or domain-general processes? *PLOS ONE* **6**, e27024 (2011).

17. S. Bouton, V. Chambon, R. Tyrand, A. G. Guggisberg, M. Seex, S. Karkar, D. van de Ville, A.-L. Giraud, Focal versus distributed temporal cortical activity for speech sound category assignment. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E1299–E1308 (2018).

18. M. De Lucia, A. Tzovara, F. Bernasconi, L. Spierer, M. M. Murray, Auditory perceptual decision-making based on semantic categorization of environmental sounds. *Neuroimage* **60**, 1704–1715 (2012).

19. J. E. Peele, M. H. Davis, Neural oscillations carry speech rhythm through to comprehension. *Front. Psychol.* **3**, 320 (2012).

20. Y. Tang, M. Cooke, B. M. Fazenda, T. J. Cox, A metric for predicting binaural speech intelligibility in stationary noise and competing speech maskers. *J. Acoust. Soc. Am.* **140**, 1858–1870 (2016).

21. Y. Tang, M. Cooke, in *Proceedings of the Annual Conference of the International Speech Communication Association*, INTERSPEECH (International Speech Communication Association, 2016), vol. 8–12–Na2016, pp. 2488–2492.

22. M. I. Mandel, S. E. Yoho, E. W. Healy, Measuring time-frequency importance functions of speech with bubble noise. *J. Acoust. Soc. Am.* **140**, 2542–2553 (2016).

23. J. K. Bizley, Y. E. Cohen, The what, where and how of auditory-object perception. *Nat. Rev. Neurosci.* **14**, 693–707 (2013).

24. S. A. Shamma, M. Elefli, C. Michely, Temporal coherence and attention in auditory scene analysis. *Trends Neurosci.* **34**, 114–123 (2011).

25. R. G. O’Connell, P. M. Dockree, S. P. Kelly, A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nat. Neurosci.* **15**, 1729–1735 (2012).

26. S. Frühholz, W. Trost, S. A. Kots, The sound of emotions—Towards unifying neural network perspectives of affective sound processing. *Neurosci. Biobehav. Rev.* **68**, 96–110 (2016).

27. S. Frühholz, W. Trost, D. Grandjean, The role of the medial temporal limbic system in processing emotions in voice and music. *Prog. Neurobiol.* **123**, 1–17 (2014).

28. A. Bendixen, Predictability effects in auditory scene analysis: A review. *Front. Neurosci.* **8**, 60 (2014).

29. C. R. Pernet, P. McAleer, M. Latini, K. J. Gorgolewski, I. Charest, P. E. G. Bestelmeyer, R. H. Watson, D. Fleming, F. Crabbe, M. Valdes-Sosa, P. Belin, The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage* **119**, 164–174 (2015).

30. G. Hickok, D. Poeppel, The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
31. J. Sohn, N. S. Kim, W. Sung, A statistical model-based voice activity detection, IEEE Signal Process. Lett. 6, 1–3 (1999).
32. B. J. Dyson, C. Alain, Representation of concurrent acoustic objects in primary auditory cortex. J. Acoust. Soc. Am. 115, 280–288 (2004).
33. Y. I. Fishman, J. C. Arezzo, M. Steinschneider, Auditory stream segregation in monkey auditory cortex: Effects of frequency separation, presentation rate, and tone duration. J. Acoust. Soc. Am. 116, 1656–1670 (2004).
34. D. C. Javitt, M. Steinschneider, C. E. Schroeder, H. G. Vaughan Jr., J. C. Arezzo, Detection of stimulus deviance within primate primary auditory cortex: Intracortical mechanisms of mismatch negativity (MMN) generation. Brain Res. 667, 192–200 (1994).
35. A. Capilla, P. Belin, J. Gross, The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. Cereb. Cortex 23, 1388–1395 (2013).
36. J. U. Blackford, J. W. Buckholtz, S. N. Avery, D. H. Zald, A unique role for the human amygdala in novelty detection. Neuroimage 50, 1188–1193 (2010).
37. A. Pannese, D. Grandjean, S. Frühholz, Amygdala and auditory cortex exhibit distinct sensitivity to relevant acoustic features of auditory emotions. Cortex 85, 116–125 (2016).
38. S. Kumar, K. von Kriegstein, K. Friston, T. D. Griffiths, Features versus feelings: Dissociable representations of the acoustic features and valence of aversive sounds. J. Neurosci. 32, 14184–14192 (2012).
39. S. Frühholz, L. Ceravolo, D. Grandjean, Specific brain networks during explicit and implicit decoding of emotional prosody. Cereb. Cortex 22, 1107–1117 (2012).
40. S. Frühholz, W. van der Zwaag, M. Saenz, P. Belin, A.-K. Schobert, P. Vuilleumier, D. Grandjean, Neural decoding of discriminative auditory object features depends on their socio-affective valence. Soc. Cogn. Affect. Neurosci. 11, 1638–1649 (2016).
41. R. H. Wiley, Signal Detection, Noise, and the Evolution of Communication, in Animal Communication and Noise, H. Brumm, Ed. (Springer-Verlag Berlin Heidelberg, 2013), pp. 7–30.
42. S. Patel, K. R. Scherer, E. Björkner, J. Sundberg, Mapping emotions into acoustic space: The role of voice production. Biol. Psychol. 87, 93–98 (2011).
43. A. Pannese, D. Grandjean, S. Frühholz, Subcortical processing in auditory communication. Hear. Res. 328, 67–77 (2015).
44. P. P. Gao, J. W. Zhang, J. S. Cheng, I. Y. Zhou, E. X. Wu, The inferior colliculus is involved in deviant sound detection as revealed by BOLD fMRI. Neuroimage 91, 220–227 (2014).
45. P. Glad Mihai, M. Moerel, F. De Martino, R. Trampel, S. Kiebel, K. von Kriegstein, Modulation of tonotopic ventral medial geniculate body is behaviorally relevant for speech recognition. eLife 8, e48837 (2019).
46. C. Lamm, T. Singer, The role of anterior insular cortex in social emotions. Brain Struct. Funct. 214, 579–591 (2010).
47. S. Frühholz, C. Hofstetter, C. Cristinio, A. Saj, M. Seeck, P. Vuilleumier, D. Grandjean, Asymmetrical effects of unilateral right or left amygdala damage on auditory cortical processing of vocal emotions. Proc. Natl. Acad. Sci. U.S.A. 112, 1583–1588 (2015).
48. S. Frühholz, D. Grandjean, Amygdala subregions differentially respond and rapidly adapt to threatening voices. Cortex 49, 1394–1403 (2013).
49. D. Sander, J. Grafman, T. Zalla, The human amygdala: An evolved system for relevance detection. Rev. Neurosci. 14, 303–316 (2003).
50. S. Molholm, A. Martinez, W. Ritter, D. C. Javitt, J. J. Foxe, The neural circuitry of pre-attentive auditory change-detection: An fMRI study of pitch and duration mismatch negativity generators. Cereb. Cortex 15, 545–551 (2005).
51. S. Chennu, V. Noreika, D. Gueorgiev, Y. Shytrov, A. T. Bekinschtein, R. Henson, Silent Expectations: Dynamic causal modeling of cortical prediction and attention to sounds that weren’t. J. Neurosci. 36, 8305–8316 (2016).
52. T. Ethofer, J. Bretscher, M. Gschwind, B. Kreifelts, D. Wildgruber, P. Vuilleumier, Emotional voice areas: Anatomic location, functional properties, and structural connections revealed by combined fMRI/DTI. Cereb. Cortex 22, 191–200 (2012).
53. S. Frühholz, M. Staib, Neurocircuitry of impaired affective sound processing: A clinical disorders perspective. Neurosci. Biobehav. Rev. 83, 516–524 (2017).
54. K. J. Friston, L. Harrison, W. Penny, Dynamic causal modelling. Neuroimage 19, 1273–1302 (2003).
55. O. T. Ousdal, J. Jensen, A. Server, A. R. Hariri, P. H. Nakstad, O. A. Andreassen, The human amygdala is involved in general behavioral relevance detection: Evidence from an event-related functional magnetic resonance imaging Go-NoGo task. Neuroscience 156, 450–455 (2008).
56. K. F. Stephan, W. D. Penny, R. J. Moran, H. F. E. M. den Ouden, J. Daunizeau, K. J. Friston, Ten simple rules for dynamic causal modeling. Neuroimage 49, 3099–3109 (2010).

Acknowledgments: We thank M. Bobin for the helpful comments on the study and the manuscript. Funding: The study was supported by the Swiss National Science Foundation (SNSF PP00P1_157409/1 and PP00P1_183711/1 to S.F.). Author contributions: H.S. and S.F. contributed to designing the experiment, acquiring and analyzing the data, and writing the manuscript. M.S. contributed to analyzing the data. Competing interests: The authors declare that they have no competing interests. Data and materials availability: All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors. The code used in the analysis of this project is also available from corresponding authors upon reasonable request.

Submitted 20 February 2020
Accepted 29 October 2020
Published 11 December 2020
10.1126/sciadv.abb3884

Citation: H. Swanborough, M. Staib, S. Frühholz, Neurocognitive dynamics of near-threshold voice signal detection and affective voice evaluation. Sci. Adv. 6, eabb3884 (2020).