

# Making sense of periodicity glimpses in a prediction-update-loop—A computational model of attentive voice tracking

Joanna Luberadzka, Hendrik Kayser and Volker Hohmann

Citation: [The Journal of the Acoustical Society of America](#) **151**, 712 (2022); doi: 10.1121/10.0009337

View online: <https://doi.org/10.1121/10.0009337>

View Table of Contents: <https://asa.scitation.org/toc/jas/151/2>

Published by the [Acoustical Society of America](#)

---

## ARTICLES YOU MAY BE INTERESTED IN

[Children's syntactic parsing and sentence comprehension with a degraded auditory signal](#)

[The Journal of the Acoustical Society of America](#) **151**, 699 (2022); <https://doi.org/10.1121/10.0009271>

## REVIEWS OF ACOUSTICAL PATENTS

[The Journal of the Acoustical Society of America](#) **151**, 663 (2022); <https://doi.org/10.1121/10.0009375>

[Perception difference for approaching and receding sound sources of a listener in motion in architectural sequential spaces](#)

[The Journal of the Acoustical Society of America](#) **151**, 685 (2022); <https://doi.org/10.1121/10.0009231>

[Interferometric processing of hydroacoustic signals for the purpose of source localization](#)

[The Journal of the Acoustical Society of America](#) **151**, 666 (2022); <https://doi.org/10.1121/10.0009381>

[The physics of knocking over LEGO minifigures with time reversal focused vibrations for use in a museum exhibit](#)

[The Journal of the Acoustical Society of America](#) **151**, 738 (2022); <https://doi.org/10.1121/10.0009364>

[Reflection on Collins' split-step Padé solution for the parabolic equation](#)

[The Journal of the Acoustical Society of America](#) **151**, R3 (2022); <https://doi.org/10.1121/10.0009374>

---



**Advance your science and career  
as a member of the**

**ACOUSTICAL SOCIETY OF AMERICA**

LEARN MORE



# Making sense of periodicity glimpses in a prediction-update-loop—A computational model of attentive voice tracking

Joanna Luberadzka,<sup>a)</sup> Hendrik Kayser,<sup>b)</sup> and Volker Hohmann<sup>b)</sup>

*Auditory Signal Processing, Department of Medical Physics and Acoustics, University of Oldenburg, Germany*

## ABSTRACT:

Humans are able to follow a speaker even in challenging acoustic conditions. The perceptual mechanisms underlying this ability remain unclear. A computational model of attentive voice tracking, consisting of four computational blocks: (1) sparse periodicity-based auditory features (sPAF) extraction, (2) foreground-background segregation, (3) state estimation, and (4) top-down knowledge, is presented. The model connects the theories about auditory glimpses, foreground-background segregation, and Bayesian inference. It is implemented with the sPAF, sequential Monte Carlo sampling, and probabilistic voice models. The model is evaluated by comparing it with the human data obtained in the study by Woods and McDermott [Curr. Biol. **25**(17), 2238–2246 (2015)], which measured the ability to track one of two competing voices with time-varying parameters [fundamental frequency ( $F_0$ ) and formants ( $F_1, F_2$ )]. Three model versions were tested, which differ in the type of information used for the segregation: version (a) uses the oracle  $F_0$ , version (b) uses the estimated  $F_0$ , and version (c) uses the spectral shape derived from the estimated  $F_0$  and oracle  $F_1$  and  $F_2$ . Version (a) simulates the optimal human performance in conditions with the largest separation between the voices, version (b) simulates the conditions in which the separation is not sufficient to follow the voices, and version (c) is closest to the human performance for moderate voice separation.

© 2022 Acoustical Society of America. <https://doi.org/10.1121/10.0009337>

(Received 15 June 2021; revised 13 November 2021; accepted 3 January 2022; published online 3 February 2022)

[Editor: Jonas Braasch]

Pages: 712–737

## I. INTRODUCTION

Selective attention is a critical aspect of auditory perception (Shamma and Micheyl, 2010; Snyder *et al.*, 2012). It has been demonstrated that attention can actually change perception: The same vibrations of the eardrums can be interpreted differently, depending which sounds are in the listener's focus (Carlyon *et al.*, 2001; Hafter *et al.*, 2008). Attention is especially important in complex auditory scenes with various simultaneously active sound sources (Koch *et al.*, 2011; McDermott, 2009; Xiang *et al.*, 2010). A classical illustration of its role is the listener's ability to attentively follow a given speaker at a cocktail party (Cherry, 1953).

A simple but powerful stimulus for investigating selective attention in the auditory system was presented by Woods and McDermott (2015). They measured the listener's ability to attentively track one of two simultaneously active synthetic voices, whose parameters—fundamental frequency and the first two formants—varied over time. There were no constant, distinctive features between the voices that could facilitate the stream formation (for

example, direction of arrival or timbre) and the multidimensional parameter trajectories crossed over time. They concluded that the listeners can successfully distinguish the attended voice from the background voice if they focus their attention on one of the voices and the parameter trajectories maintain a sufficient separation in the feature space.

Exactly how the mixture of acoustic signals is decomposed into the attended foreground and residual background remains unclear (Carlyon, 2004). Over the last few years, several researchers have addressed this and other questions related to the auditory attention by developing computational models (Di Fu *et al.*, 2020; Kaya and Elhilali, 2017; Shinn-Cunningham, 2008; Szabó *et al.*, 2016; Wrigley and Brown, 2004) and machine-hearing systems (Cohen-Lhyver *et al.*, 2018). Our aim is to contribute our ideas to this family of models. In particular, we address the aspect of the attentive tracking of auditory objects. We introduce a computational model, illustrating how the top-down attention is maintained on a chosen stream over time. We use the model to predict the results obtained by the human listeners (Woods and McDermott, 2015).

On a conceptual level, the model brings together several theories about auditory perception. To begin with, we adopt the notion that perceptual mechanisms involved in auditory scene analysis can be characterized on a scale between the *bottom-up* and *top-down* processes (see the arrows in the upper left and bottom right corners of Fig. 1).

<sup>a)</sup>Also at: Cluster of Excellence Hearing4all, Department of Medical Physics and Acoustics, University of Oldenburg, Germany. Electronic mail: joanna.luberadzka@uni-oldenburg.de

<sup>b)</sup>Also at: Cluster of Excellence Hearing4all, Department of Medical Physics and Acoustics, University of Oldenburg, Germany.

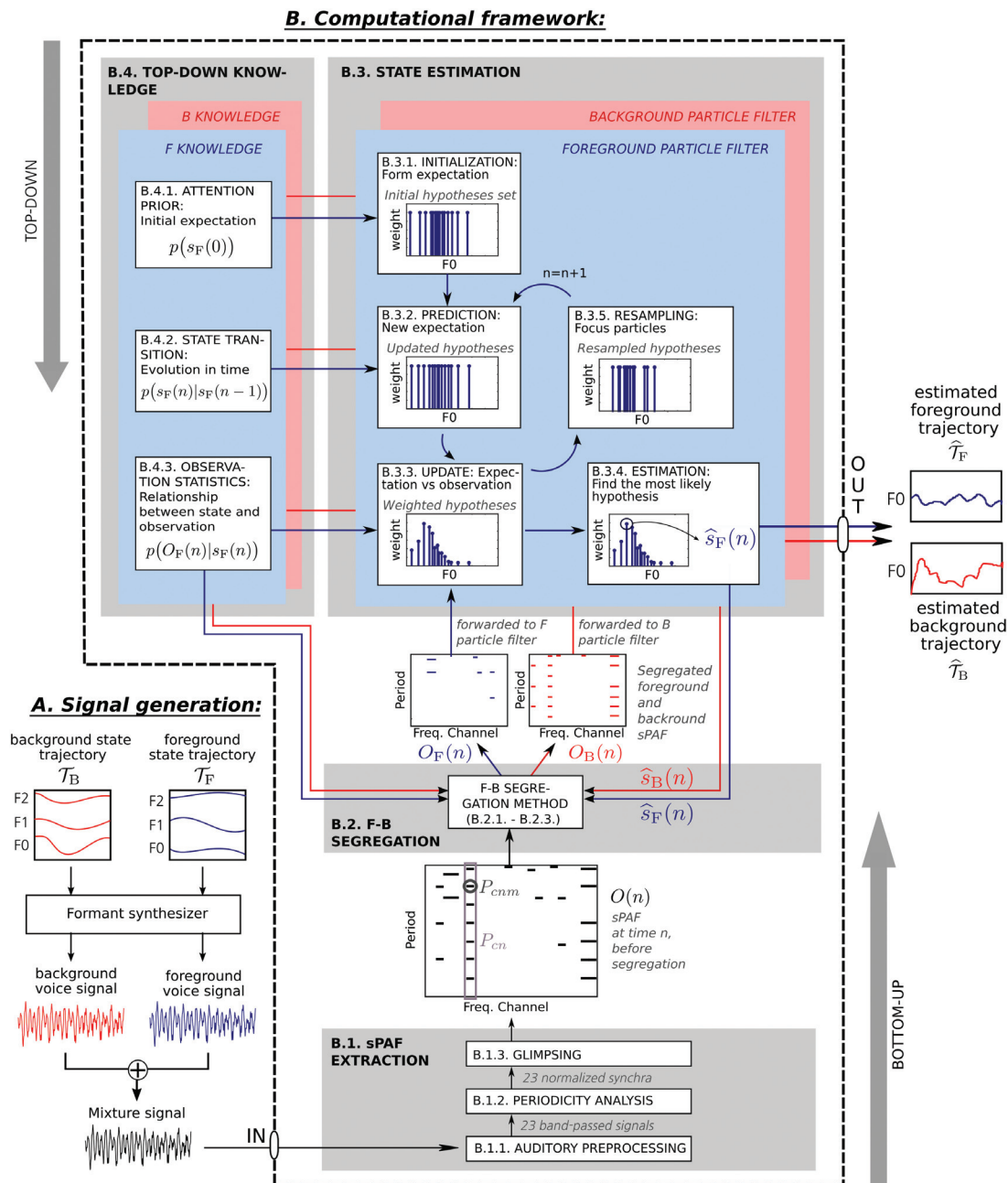


FIG. 1. The modeling framework. (A) The signal generation features two synthetic voiced signals created based on the randomized state trajectories, which comprise the time series of the voice parameters. The mixture of the foreground ( $F$ ) and background ( $B$ ) signals is the input to the model. (B) The computational framework emphasizes the attentive tracking of a foreground sound source in an auditory scene implemented with several computational sub-tasks, placed on a top-down-bottom-up scale (see the arrows in the top left and bottom right corners). The sub-tasks include (B.1.) sparse periodicity-based features extraction, (B.2.) foreground-background segregation, (B.3.) state estimation, and (B.4.) top-down knowledge. The model outputs the estimated state trajectories of the foreground and background voices.

The scale starts with the fully stimulus-driven processing, responsible for extracting the basic auditory features (such as harmonic structure, time onset, and spatial direction; Darwin, 2008). The top-down attention relies on these features. To follow a chosen source, the properties of this source must, to some extent, be reflected in the basic auditory features. Whether this is the case depends on the complexity of the auditory scene. In a complex scenario with simultaneously active sound sources, the energetic and informational masking makes it difficult to extract reliable information about the individual sources.

In such a scenario, the perception can be characterized as *glimpsing* (Darwin, 2008). Reliable auditory features are provided primarily by the time-frequency ( $T$ - $F$ ) units, which are dominated by a single voice: *auditory glimpses*. Several studies suggest that the auditory system navigates in the complex acoustic environment using these non-disrupted pieces of information, coming from one source at time rather than the superposed features from several sound objects (Best et al., 2016; Best et al., 2017; Cooke, 2006; Schoenmaker and van de Par, 2016). The studies mentioned

above usually prove the relevance of the  $T$ - $F$  bins, which are dominated by a single source, by comparing the speech intelligibility for the signals resynthesized from the partial information with the speech intelligibility of the unprocessed mixture. To select the speaker-related information, they use perfect knowledge about the signal and masker energies. Blindly modeling the *auditory glimpses* is, however, not a straightforward task. A series of modeling studies (Josupeit and Hohmann, 2017; Josupeit *et al.*, 2016; Josupeit *et al.*, 2020) developed an approach capable of blindly extracting the speech glimpses from the mixture of signals, which are called the *sparse periodicity-based auditory features* (sPAF). Speech is a highly periodic signal, and the harmonic structure plays a crucial role in voiced speech segregation (Popham *et al.*, 2018). The core idea of the sPAF is to use the salient periodicity as a footprint of speech in a sound mixture.

The auditory model introduced by Josupeit and Hohmann (2017) and further investigated by Josupeit *et al.* (2020) was used to model several aspects of auditory perception in a multi-talker scenario. The results suggested that despite their sparsity, the sPAF contain the bulk of the information needed to decode a complex auditory scene. The *sPAF extraction* stage (see Fig. 1, B.1. and Sec. II B 1] of the attentive tracking model presented here is based on the approach of Josupeit and Hohmann (2017).

Further up the bottom-up-top-down scale, we locate the process of assigning the basic auditory features to streams. Even a complex auditory scene with multiple simultaneously active sound sources usually contains a single object of interest to which the listener attends (Shinn-Cunningham, 2008). Hence, we can identify two main attention-dependent streams: attended foreground, corresponding to the target, and unattended background, comprised of the remaining irrelevant clutter (Elhilali *et al.*, 2009). Segregating a sound mixture into the foreground and background can be relatively straightforward if the sounds in the background are different from the target. An example of this is speech in a stationary noise. In that case, extracting the target-related information is mainly a matter of detecting the  $T$ - $F$  bins, which are qualitatively different from the rest of the mixture. However, the background can also consist of sound sources with similar properties as the signal of interest. Following a speaker in a multi-talker condition is a good example of that scenario. In that case, many of the local  $T$ - $F$  regions are dominated by one of the speakers. The auditory glimpses from the attended target talker must, therefore, be perceptually separated from the glimpses of all of the remaining talkers (Darwin, 2008).

According to the principle of old-plus-new heuristics (Bregman, 1990), the auditory system is prone to interpret the incoming features as a continuation of the already existing streams. The continuous nature of the auditory streams enhances the listeners ability to focus their attention on a desired object (Best *et al.*, 2008; Bressler *et al.*, 2014; Woods and McDermott, 2015). In the *foreground-background segregation* stage (see Fig. 1, B.2. and Sec. II B 2] of

the current model, the sPAF at a given time instance are decomposed into foreground and background features based on the stream estimates from the preceding time step.

The idea of top-down processing is tied to the assumption that the brain is equipped with some preliminary knowledge that helps to interpret the signals received by the sensory organs (Ellis, 1999; Gregory, 1997). This may include the conceptual and contextual schemas and goal-oriented attention. Mesgarani *et al.* (2014) showed that clean speech, reverberant speech, and speech in noise evoke similar neural responses. This suggests a neural denoising mechanism, which provides robust representation of speech, independent of the background sounds. Considering that the brain is able to extract the clean representation of speech in the auditory scene, we assume that the knowledge of clean speech is sufficient to solve the task of attentive tracking. Hence, we represent top-down knowledge in a form of statistical models, describing properties of a single, clean voice in isolation. The uncertainty of the model comes entirely from the variability within the clean voice itself and does not arise from the superposed background signal. The probability models required for the attentive tracking task are discussed in the *top-down knowledge* stage (see Fig. 1, B.4. and Sec. II B 4) of the model. The implementation details of these models are presented in Sec. II C.

The top-down and bottom-up processes are not mutually independent: They are engaged in a machinery of interactions, which are generating perception (Bressler *et al.*, 2014). Many recent studies propose the Bayesian inference as an elegant computational illustration of this synergy. The Bayesian estimation is granted as a plausible model of the optimal inference in a general view on cognition (Aitchison and Lengyel, 2017; Chater *et al.*, 2006; Helmholtz, 1897; Pouget *et al.*, 2013) as well as in the context of auditory perception (Elhilali, 2013; Heilbron and Chait, 2018; Nix and Hohmann, 2007; Schröger *et al.*, 2015).

In a nutshell, the Bayesian models of the perceptual inference postulate that our brains (a) learn and store the statistical models of the environment, (b) use them to constantly generate expectations about what might currently be happening, and (c) confront the expectation with the incoming sensory information to estimate what is really happening. Mismatch negativity in the event-related brain potential, elicited by an unexpected—deviant—sound occurring in a predictable sound sequence is a typical demonstration of the predictive nature of auditory perception (Garrido *et al.*, 2009; Näätänen *et al.*, 1978).

From the mathematical perspective, the sequential Bayesian estimation—finding the posterior state distribution given a series of observations—does not have a single universal solution. There are a variety of methods that can be used, for example, Kalman filtering, hidden Markov models, or the sequential Monte Carlo sampling also known as *particle filtering* (Chen *et al.*, 2003). Although the sampling-based scheme is usually computationally costly, it has the advantage that it does not assume any particular distribution



of the data: A finite set of hypotheses about the state is iteratively evaluated and updated. This embodies the idea of perception as hypotheses testing proposed by Gregory (1980), who speculated that the brain uses “fiction-generators, which may hit upon the truth by producing symbolic structures matching physical reality.” This notion has been revisited in the recent studies comparing the perceptual inference to Bayesian sampling (Friston *et al.*, 2012; Nix and Hohmann, 2007; Sanborn and Chater, 2016; Shi and Griffiths, 2009). We adopt these concepts in our model: The *state estimation* stage (see Fig. 1, B.3. and Sec. II B 3) integrates the components of our model into a sequential Bayesian inference framework. We use the Monte Carlo sampling approach, i.e., particle filtering (Arulampalam *et al.*, 2002), to simultaneously track the state of the foreground and background streams. The particle filters have already been used successfully in the context of speech tracking (Nix and Hohmann, 2007; Spille *et al.*, 2013).

This study presents a computational model of the chosen aspects of human auditory perception, which takes the above-discussed theories into account. In particular, it illustrates the human ability to attentively track a voice in the presence of other sounds. Except for the sPAF extraction stage, the model is novel and was designed and implemented for the first time for the purpose of this study. We demonstrate the feasibility of the model using the auditory scene from the study of Woods and McDermott (2015): two simultaneously active synthetic voices with varying  $F0$ ,  $F1$ , and  $F2$ , whose parameter trajectories cross-in time. This type of stimulus is new among the sPAF-based modeling approaches, which previously worked with multi-talker speech sets (Josuweit and Hohmann, 2017; Josuweit *et al.*, 2020). Instead of the template-matching approach, which evaluated the sPAF on the word scale, the proposed model performs sequential processing. It tracks the fundamental frequency of the competing voices using instantaneous sPAF and probabilistic  $F0$  models (Sec. II C). The sPAF and  $F0$  models are integrated in a particle-filtering-based framework for the first time. Previous modeling approaches used spectral coefficients (Nix and Hohmann, 2007) or output features of a binaural model (Spille *et al.*, 2013) and a codebook-based approach to track the spectral envelope and direction of arrival. Furthermore, the parallel particle filtering as a solution for tracking multiple voices is introduced here for the first time. The main contributions of this work include:

- (1) a theoretical framework unifying the modeling approaches related to the sPAF (Josuweit and Hohmann, 2017; Josuweit *et al.*, 2020) and sequential Bayesian inference (Elhilali, 2013; Elhilali and Shamma, 2008; Nix and Hohmann, 2007);
- (2) an  $F0$  observation model summarizing the statistical relationship between the fundamental frequency and salient periodicity; and

- (3) a comparison of a single-dimensional version of our model ( $F0$  tracking) with human results in the attentive tracking paradigm by Woods and McDermott (2015).

## II. MODEL

In this section, we introduce the computational model used in the current study. The model is depicted in Fig. 1. Section II A describes the generation of the stimuli: synthetic competing voices with time-varying parameters. Section II B guides the reader through the computational blocks of the modeling framework: *sPAF extraction* (Sec. III B 1), where the salient auditory features are extracted from the input mixture; *foreground-background segregation* (Sec. III B 2), where the sPAF are segregated into foreground and background segments; *state estimation* (Sec. III B 3), where the particle filter estimates the voice state based on the segregated sPAF; and *top-down knowledge* (Sec. III B 4), where the probability distributions required for the state estimation are briefly reviewed. Section II C presents the implementation details of the probability distributions related to the voice fundamental frequency, which are used in this study: the  $F0$  transition and  $F0$  observation models.

### A. Signal generation

We assume that an auditory scene consists of the *foreground* ( $F$ ) and the *background* ( $B$ ). The foreground contains the attended stream of information: the target auditory object. All of the remaining components of the auditory scene belong to the background stream.

As a simple example of such an auditory scene, we use two competing voiced signals, whose fundamental frequencies  $F0$  and first two formants  $F1$  and  $F2$  vary over time as in the study by Woods and McDermott (2015). One voice is considered to be the foreground and the other voice is considered to be the background. In each time instance, the signals are defined by the three-dimensional state vectors  $\vec{s}_F$  and  $\vec{s}_B$ , containing the parameter values,

$$\vec{s}_F(n) = \begin{pmatrix} F0_F(n) \\ F1_F(n) \\ F2_F(n) \end{pmatrix}, \quad \vec{s}_B(n) = \begin{pmatrix} F0_B(n) \\ F1_B(n) \\ F2_B(n) \end{pmatrix}, \quad (1)$$

where  $n$  is the time index. The full-length signals are defined by the ground truth state trajectories,

$$\begin{aligned} \mathcal{T}_F &= \{\vec{s}_F(n) | n = 0, \dots, N\}, \\ \mathcal{T}_B &= \{\vec{s}_B(n) | n = 0, \dots, N\}, \end{aligned} \quad (2)$$

where the trajectory sampling rate is  $F_S = 50$  Hz, and  $N$  is the length of the trajectory. The ground truth trajectory of each voice is taken as the input to a formant synthesizer, which generates the synthetic foreground and background acoustic signals. The signals are summed, and the resulting mixture is the input to the model. Figure 1, A, illustrates the signal generation procedure.

## B. Computational framework

### 1. sPAF extraction

The signal containing mixture of the foreground and background is forwarded to the feature extraction stage (see Fig. 1, B.1.). The sPAF extraction stage is based on a series of studies by Josupeit *et al.* (2016), Josupeit and Hohmann (2017), and Josupeit *et al.* (2020). The sPAF represent the robust tonal components of the auditory scene (see Fig. 2). Below, we briefly review the method. For the implementation details, the reader is referred to Appendix A.

The sPAF extraction consists of three main steps (Fig. 1, B.):

- (1) *auditory pre-processing*, which provides the auditory-inspired  $T$ - $F$  representation (Fig. 1, B.1.1.);
- (2) *periodicity analysis*, which analyzes the periodic structure of the sound in each considered frequency band and yields a  $T$ - $F$ -period representation (Fig. 1, B.1.2.);
- (3) *glimpsing*, which removes all the non-salient information from the  $T$ - $F$ -period representation and extracts the salient period values, herein called *period glimpses* (Fig. 1, B.1.3.).

In every time instance  $n$  at the output of the sPAF extraction stage, we obtain an *observation*  $O(n)$  (see Fig. 1, output of the model step B.1.3.) such that

$$O(n) = \{P_{cn} | c = 1, \dots, 23\}. \quad (3)$$

$O(n)$  consists of 23 *channel sets*  $P_{cn}$  for each frequency channel  $c$ ,

$$P_{cn} = \{P_{cnm} | m = 1, \dots, M_{cn}\}. \quad (4)$$

$P_{cn}$  consists of the salient period values—*period glimpses*—denoted as  $P_{cnm}$ . The indices  $c$ ,  $n$ , and  $m$  denote the frequency channel, time instance, and period glimpse

index, respectively.  $M_{cn}$  is the total number of period glimpses in the set  $P_{cn}$ . One channel set  $P_{cn}$  can consist of a single value, multiple values, or no value at all. Hence, the magnitude of the sPAF can change depending on the acoustic scene.

### 2. Foreground-background segregation

The observation  $O(n)$  is segregated into the *foreground observation*  $O_F(n)$  and *background observation*  $O_B(n)$  (see Fig. 1, B.2.). Following the assumption that each channel set represents only one voice, each set  $P_{cn}$  is assigned to either the foreground or the background. This is done by comparing the likelihood that the set  $P_{cn}$  belongs to the foreground with the likelihood that it belongs to the background. The likelihood is derived from the previous foreground and background estimates or based on the ground truth values used in the signal generation. Which values are used depends on the foreground-background segregation method. In this study, we use three alternative methods to evaluate the various aspects of the model. Figure 3 shows the pseudo-code for each of these methods. The purpose of each method is reviewed in more detail in Sec. III B.

### 3. State estimation

The goal of this stage of the model (depicted in Fig. 1, B.3.) is to track the state of the foreground and background voices given the segregated sPAF. For tracking, we use particle filtering, combined with the probability distributions from the *top-down knowledge* stage (Fig. 1, B.4.). The particle filters approximate the Bayesian posterior distribution in an iterative prediction-update procedure (Arulampalam *et al.*, 2002). The key idea is to represent this density function by a set of random samples with associated weights and iteratively compute the state estimate based on these samples and weights.

Although particle filtering allows for the multidimensional tracking of events with arbitrary probability

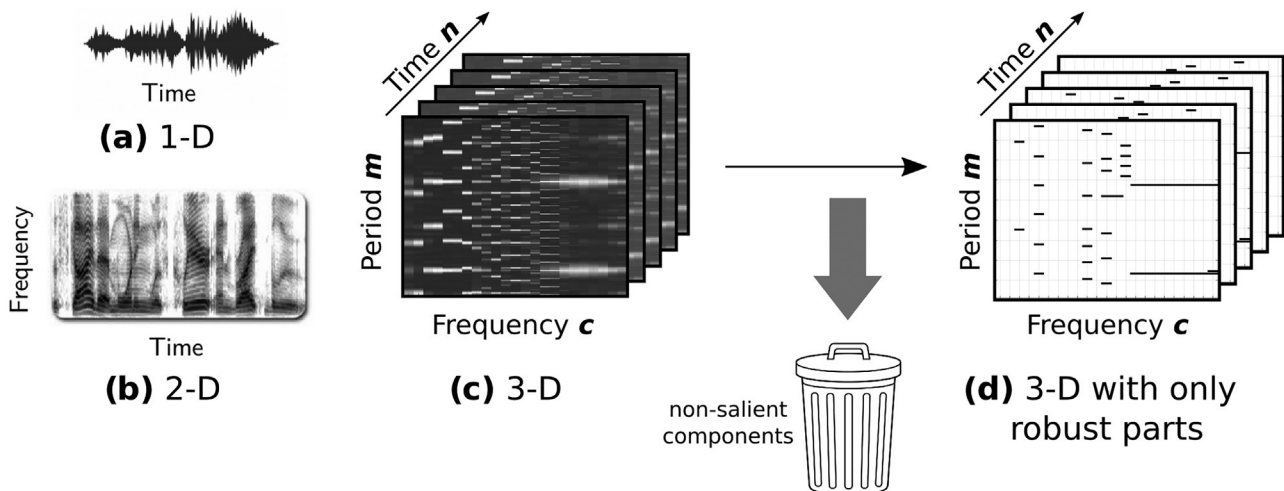


FIG. 2. The sPAF represents the robust tonal components of the auditory scene. In contrast to a one-dimensional time representation (a) or two-dimensional  $T$ - $F$  representation (b), the sPAF analyzes the sound also in a third dimension, period (c). Moreover, the sPAF allows only the salient, robust components originating from a single sound source. Noisy  $T$ - $F$  bins, as well as the bins containing the superposition of many sound sources are eliminated by the glimpsing process (d).

<p><b>B.2.1. F0-guided segregation</b></p> <p><b>FOR</b> all <math>P_{cn}</math> in <math>O(n)</math></p> <p>  # Compute likelihood given   # previous foregr. ground-truth <math>F_0</math></p> <p>  <math>L_F = p(P_{cn} F_0(n-1))</math></p> <p>  # Compute likelihood given   # previous backgr. ground-truth <math>F_0</math></p> <p>  <math>L_B = p(P_{cn} F_0(n-1))</math></p> <p>  <b>IF</b> <math>L_F &gt; L_B</math></p> <p>    assign <math>P_{cn}</math> to <math>O_F(n)</math></p> <p>  <b>ELSE</b></p> <p>    assign <math>P_{cn}</math> to <math>O_B(n)</math></p> <p>  <b>END</b></p> <p><b>END</b></p>	<p><b>B.2.2. Segregation without oracle information</b></p> <p><b>FOR</b> all <math>P_{cn}</math> in <math>O(n)</math></p> <p>  # Compute likelihood given   # previous foregr. <math>F_0</math> estimate</p> <p>  <math>L_F = p(P_{cn} \widehat{F_0}(n-1))</math></p> <p>  # Compute likelihood given   # previous backgr. <math>F_0</math> estimate</p> <p>  <math>L_B = p(P_{cn} \widehat{F_0}(n-1))</math></p> <p>  <b>IF</b> <math>L_F &gt; L_B</math></p> <p>    assign <math>P_{cn}</math> to <math>O_F(n)</math></p> <p>  <b>ELSE</b></p> <p>    assign <math>P_{cn}</math> to <math>O_B(n)</math></p> <p>  <b>END</b></p> <p><b>END</b></p>	<p><b>B.2.3. Formant-guided segregation</b></p> <p><b>FOR</b> all <math>P_{cn}</math> in <math>O(n)</math></p> <p>  # Compute spectral power (SP) in channel <math>c</math> given   # previous foregr. <math>F_0</math> estimate and ground-truth <math>F_1</math> &amp; <math>F_2</math></p> <p>  <math>E_F = \text{SP}(c, \widehat{F_0}(n-1), F_{1F}(n-1), F_{2F}(n-1))</math></p> <p>  # Compute spectral power (SP) in channel <math>c</math> given   # previous backgr. <math>F_0</math> estimate and ground-truth <math>F_1</math> &amp; <math>F_2</math></p> <p>  <math>E_B = \text{SP}(c, \widehat{F_0}(n-1), F_{1B}(n-1), F_{2B}(n-1))</math></p> <p>  <b>IF</b> <math>E_F &gt; E_B</math></p> <p>    assign <math>P_{cn}</math> to <math>O_F(n)</math></p> <p>  <b>ELSE</b></p> <p>    assign <math>P_{cn}</math> to <math>O_B(n)</math></p> <p>  <b>END</b></p> <p><b>END</b></p>
---	--	--

FIG. 3. The foreground-background segregation methods. Method B.2.1. uses the ground truth  $F_0$  values from the preceding time step, which was used to generate the signals before mixing. For each voice, a likelihood of the observed channel set  $P_{cn}$  [Eq. (4)] given the oracle  $F_0$  value is computed via the likelihood function. The likelihoods are compared and the set  $P_{cn}$  is assigned to the voice for which the likelihood is larger. Method B.2.2. is also based on  $F_0$ , and the only difference with B.2.1. is that the segregation is performed based on the estimated (instead of ground truth)  $F_0$ s from the preceding time step. Method B.2.3. is substantially different from the first two methods. In addition to the estimated  $F_0$ , it uses the ground truth information about the formants to segregate the voices. For each voice, a channel-dependent weight is computed. This reflects the energy distribution over the frequency channels for a given combination of  $\widehat{F_0}$ ,  $F_1$ , and  $F_2$ . The  $F_0$  estimate is used in the encoding of weights but is not explicitly used for the segregation as in the first two methods.

distributions, the tracking cannot be performed until these distributions are known (see also Sec. II B 4). For the three-dimensional voice state, which was used for the signal generation (Sec. II A), the *state transition* distribution, which describes the evolution of the state, could be derived easily: The transitions of the parameters can be modeled with a linear motion model. However, the *observation statistics* distribution, describing the relationship between the three-dimensional state and the sPAF—sparse features with a changing magnitude—is a much more complex problem. In the current work, we derive this observation model only for the fundamental frequency, and we validate our framework in the one-dimensional case.

Specifically, the system tracks a single dimension of the multidimensional generative state:  $F_0$  (see the output of the computational model in Fig. 1). Thus, when referring to the estimation process, we replace the vector notation  $\vec{s}$  with a one-dimensional state  $s$ ,

$$\begin{aligned}\widehat{s}_F(n) &= \widehat{F_0}(n), \\ \widehat{s}_B(n) &= \widehat{F_0}_B(n).\end{aligned}\quad (5)$$

The symbol “ $\widehat{\phantom{x}}$ ” is also used to differentiate between the ground truth and estimated quantities. Tracking yields the one-dimensional estimated state trajectories,

$$\begin{aligned}\widehat{T}_F &= \{\widehat{s}_F(n)|n=0, \dots, N\}, \\ \widehat{T}_B &= \{\widehat{s}_B(n)|n=0, \dots, N\}.\end{aligned}\quad (6)$$

We use two parallel particle filters: one for the foreground and one for the background. Each particle filter consists of a finite set of 300 particles, i.e., the hypothetical  $F_0$  values with weights assigned to them. Below, we shortly review the processing steps for the foreground particle filter.

The background particle filter executes the same operations for the background stream.

- (1) At the system onset, a particle filter is *initialized* with the available prior knowledge (Fig. 1, B.3.1.). The initial expectation is created. The hypotheses are sampled from the *attention prior* probability distribution:  $p(s_F(0))$ . All of the particles are given the same weight.
- (2) Next, the iteration of a particle filter begins with the *prediction step* (Fig. 1, B.3.2.). The system at the current time  $n$  is expected to have changed since the last time instance  $n-1$ . Hence, the new expectation is formed. The new particles are predicted given the previous particles. This is performed by drawing the samples from the *state transition* distribution,  $p(s_F(n)|s_F(n-1))$ .
- (3) In the next step, the expectation is confronted with the the observation (Fig. 1, B.3.3.). The weights of the particles are updated. For each hypothesis in the particle set, the weight is computed by evaluating the *observation statistics* distribution  $p(O_F(n)|s_F(n))$ . Additionally, the weight is incrementally updated by multiplying the previous weight with the observation statistics and renormalizing across the particles.
- (4) The hypothetical states together with the normalized weights assigned to them constitute the approximate discrete posterior distribution for the foreground voice. The final *state estimate*—the most likely hypothesis—is the expected value of the approximate posterior (Fig. 1, B.3.4.).
- (5) The iteration finishes with a *resampling step* (Fig. 1, B.3.5.), which is executed to focus the limited computational resources (finite particle set) on the regions of high importance. The particles with small weights are eliminated, and the particles with large weights are duplicated. The resampled particle set is generated by

drawing samples from the most recent approximate discrete posterior distribution. Resampling reduces the problem of *weight degeneracy*, which is a consequence of the particles being distributed too widely. However, it can lead to the overconcentration of particles, which, on the other hand, leads to *sample impoverishment* (Li *et al.*, 2014). To overcome this trade-off, the resampling step is performed only at time steps when the foreground observation  $O_F(n)$  is not empty and the particle diversity, measured in terms of the *effective sample size*, is lower than a predetermined threshold.

For the mathematical details of the particle-filtering approach, the reader is referred to the literature (Arulampalam *et al.*, 2002; Chen *et al.*, 2003).

#### 4. Top-down knowledge

The probability models describing the properties of a considered auditory scene are required to solve the sequential Bayesian estimation. They are included in the *top-down knowledge* stage of the model (Fig. 1, B.4.). The following probabilistic functions are used:

- (1) The *attention prior*  $p(s_F(0))$  (Fig. 1, B.4.1.) describes the initial expectation about the state. In the attentive tracking task, the listener is instructed to focus on a particular stream by playing back the beginning of that stream alone. Having heard this cue, the listener knows the initial parameters associated with that stream. The foreground tracking, in contrast to the background tracking, is initialized in an informed way. This is illustrated in Fig. 4.
- (2) The *state transition probability*  $p(s_F(n)|s_F(n-1))$  (Fig. 1, B.4.2.) describes the dynamics of the state and is responsible for the particle development: predicting new hypotheses based on the previous hypotheses set. In this

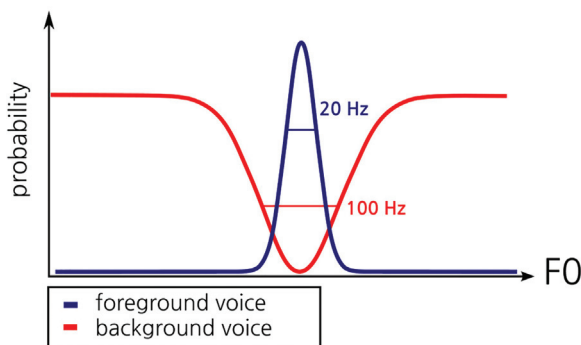


FIG. 4. The attention prior probability model. We simulate how the listener's expectation is guided in a selective auditory attention task. The foreground particle filter is initialized around the ground truth value and the background particle filter everywhere else. This was performed by sampling from the discrete distributions defined in the range between 100 and 400 Hz with the shapes as shown in this figure and the weights normalized to one. The shape of the distribution was derived using a normal distribution with a 10 Hz standard deviation for the foreground and a 50 Hz standard deviation for the background.

study, we use the  $F0$  transition model described in detail in Sec. II C 1.

- (3) The *observation statistics*  $p(O_F(n)|s_F(n))$  (Fig. 1, B.4.3.) describes the relationship between the observation and the state. For this study, it is a function which quantifies the likelihood that the observed and segregated sPAF originate from a given  $F0$ . It is computed by integrating the likelihood of all of the sPAF channel sets  $P_{cn}$  contributing to the observation  $O_F(n)$ .

The above distributions describe the properties of a single voice in isolation but are later applied to the segregated mixture of two voices. We consider an auditory scene comprised of two voices: The foreground and background have the same properties. Hence, apart from the *attention prior*, the models for the foreground and background are the same. For a more detailed mathematical description, please refer to Sec. II C.

#### C. Single voice $F0$ models

This section contains the implementation details of the probabilistic models required to track  $F0$  based on the sPAF. Section II C 1 reviews the  $F0$  transition model, and Sec. II C 2 reviews the  $F0$  observation model.

##### 1. $F0$ transition model

The  $F0$  transition model describes the temporal evolution of  $F0$ , which is naturally limited due to the physical constraints of the speech production. This continuity is conveyed in the transition model. To predict the next value for a given hypothetical  $F0$ , the trend  $\Delta\widehat{F0}(n) = \widehat{F0}(n-2) - \widehat{F0}(n-1)$  between two previous estimates is computed, the next value according to that trend  $F0 + \Delta\widehat{F0}(n)$  is predicted, and, finally, the Gaussian noise is added to this value,

$$p(F0(n)|F0(n-1)) = \mathcal{N}(F0(n-1) + \Delta\widehat{F0}(n), \sigma_{\text{trans}}),$$

$$\Delta\widehat{F0} = \widehat{F0}(n-2) - \widehat{F0}(n-1), \quad (7)$$

where  $\sigma_{\text{trans}} = 1$  Hz is the standard deviation of the Gaussian distribution centered at  $F0(n-1) + \Delta\widehat{F0}(n)$ . In addition, we make sure that the difference between the two previous estimates  $\Delta\widehat{F0}(n)$  does not exceed the largest allowed step of  $10\sigma_{\text{trans}}$  and the extrapolated value  $F0 + \Delta\widehat{F0}(n)$  does not exceed a typical pitch value range [100, 400]. We can describe this procedure with the following pseudocode:

```

WHILE  $|\Delta\widehat{F0}(n)| > 10\sigma_{\text{trans}}$  or  $F0$ 
     $+ \Delta\widehat{F0}(n) \notin [100, 400]$ 
     $\Delta\widehat{F0}(n) := 0.8\Delta\widehat{F0}(n)$ 
END WHILE
    
```

The process of predicting a new  $F0$  value based on the hypothetical previous  $F0$  is illustrated in Fig. 5.



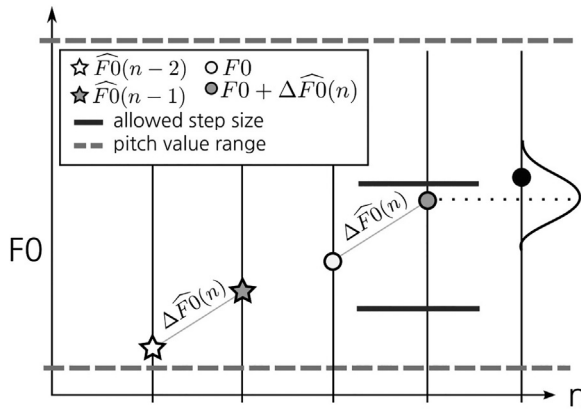


FIG. 5. The state transition probability model predicting the next state value.  $F_0$ , the hypothetical fundamental frequency;  $\widehat{F}_0(n-1)$ , the  $F_0$  estimate in the last time step;  $\widehat{F}_0(n-2)$ , the  $F_0$  estimate in the second-to-last time step;  $\Delta\widehat{F}_0(n)$ , the difference between the last two  $F_0$  estimates.

## 2. $F_0$ observation model

The  $F_0$  observation model describes the relationship between the observed (and segregated) sPAF and the underlying  $F_0$ . It quantifies the likelihood of the segregated sPAF given a hypothetical  $F_0$  value.

A major difficulty in designing the likelihood function for the sPAF is that the magnitude of the observation varies largely over time. Depending on the temporary properties of the sound mixture, the segregated observation  $O_F(n)$  can include a different number of channel sets  $P_{cn}$  and within the channel sets, a varying number of period glimpses  $P_{cnm}$  can be observed (details about the sPAF can be found in Appendix A). The observation statistics function has to deal with this changing magnitude of the observation, which is a typical problem for the models with sparse observation. We solve this problem in the following way. First, the likelihood of every observed period glimpse  $P_{cnm}$  is computed individually. Next, in each non-empty channel set  $P_{cn}$ , the likelihood is integrated by computing a product across the likelihoods of the elements of the channel set,

$$p(P_{cn}|F_0) = \prod_m p(P_{cnm}|F_0). \quad (9)$$

We assume the mutual independence of the period glimpses within a channel set  $P_{cn}$ . The product ensures that the high likelihood for a given  $F_0$  only occurs when there is a good match for all of the period glimpses detected in this frequency channel.

Each channel set contributing to the segregated observation  $O_F(n)$  provides more evidence for the considered stream. Therefore, the likelihood is summed across the frequency channels,

$$p(O_F(n)|F_0) = \sum_c p(P_{cn}|F_0). \quad (10)$$

Below we explain the motivation and implementation of the function  $p(P_{cnm}|F_0)$ , which evaluates the likelihood of a single period glimpse  $P_{cnm}$ . Section II C 2 a discusses

the distribution of the period glimpses and its relation to the period histogram by Schroeder (1968). Section II C 2 b introduces the notion of a *relative period glimpse* and comments on the distribution of the relative period glimpses. Section II C 2 c presents our approach for modeling these data as a mixture of circular von-Mises distributions.

**a. Distribution of period glimpses.** The period glimpses  $P_{cnm}$  do not always represent the period of  $F_0$  itself. An example of the sPAF extracted from a voice with  $F_0 = 124$  Hz is shown in Fig. 6(A). Bandpass filtering influences the periodicity of the signal at the output of each frequency channel [see Fig. 6(B)]. In the low-frequency channels, there is typically only one resolved harmonic of  $F_0$  per band; the periodicity in these channels is related to the dominant harmonic. The high-frequency channels are broad enough to fit several harmonics, which interact with each other; in these channels, the period is related to the difference frequency, which is  $F_0$  itself (a similar nature of the periodicity at the output of the cochlear-inspired filterbank was also described in Shamma and Dutta, 2019). Furthermore, a signal with a period  $P$  is also periodic at  $2P$ ,  $3P$ ,  $4P$ , etc., therefore, multiples of the period are also detected as glimpses. Altogether, a single period glimpse  $P_{cnm}$  can assume any value equal to  $i/jF_0$ , where  $j = 1, 2, \dots$  is a harmonic number and  $i = 1, 2, \dots$  is a period multiple number.

The principle has already been described by Schroeder (1968), who used the notion of the *period histogram* to estimate the fundamental frequency. Schroeder (1968) reported that the instantaneous periods, which can be detected at the output of a filterbank for a signal with a given  $F_0$ , can be related to  $F_0$ , harmonics of  $F_0$  ( $1/2F_0$ ,  $1/3F_0$ ,  $1/4F_0$ , etc.), or their multiples ( $2/2F_0$ ,  $3/2F_0$ ,  $4/2F_0$ , etc.). A schematic example of such a period histogram is shown in Fig. 6(C). Without the knowledge of the harmonic order, it is not possible to derive the original  $F_0$  directly from a detected period. However, for a given  $F_0$ , certain period values are more likely to occur than others. Various  $F_0$ -estimation techniques were based on this observation: The histogram of the instantaneous periods has a peak at the fundamental period.

In our model, the period histogram cannot be used directly to estimate the  $F_0$ . The signal contains two simultaneously active voices sharing the frequency space: Some frequency channels show evidence of one voice, some of another voice, and there can be also channels in which no salient periodicity is found. Altogether, there are not enough period glimpses in a single time frame to construct a meaningful period histogram with a clear peak at a  $1/F_0$ . However, we propose to use the concept of the period histogram to derive the probability model describing period glimpse likelihood for a given  $F_0$ ,  $p(P_{cnm}|F_0)$ .

**b. Distribution of relative period glimpses.** To design the probability model  $p(P_{cnm}|F_0)$ , we analyzed the empirical distribution of the period glimpses  $P_{cnm}$ , which was extracted from a single synthesized voice with varying  $F_0$ ,

# Period glimpses for a voice with $F_0=124$ Hz:

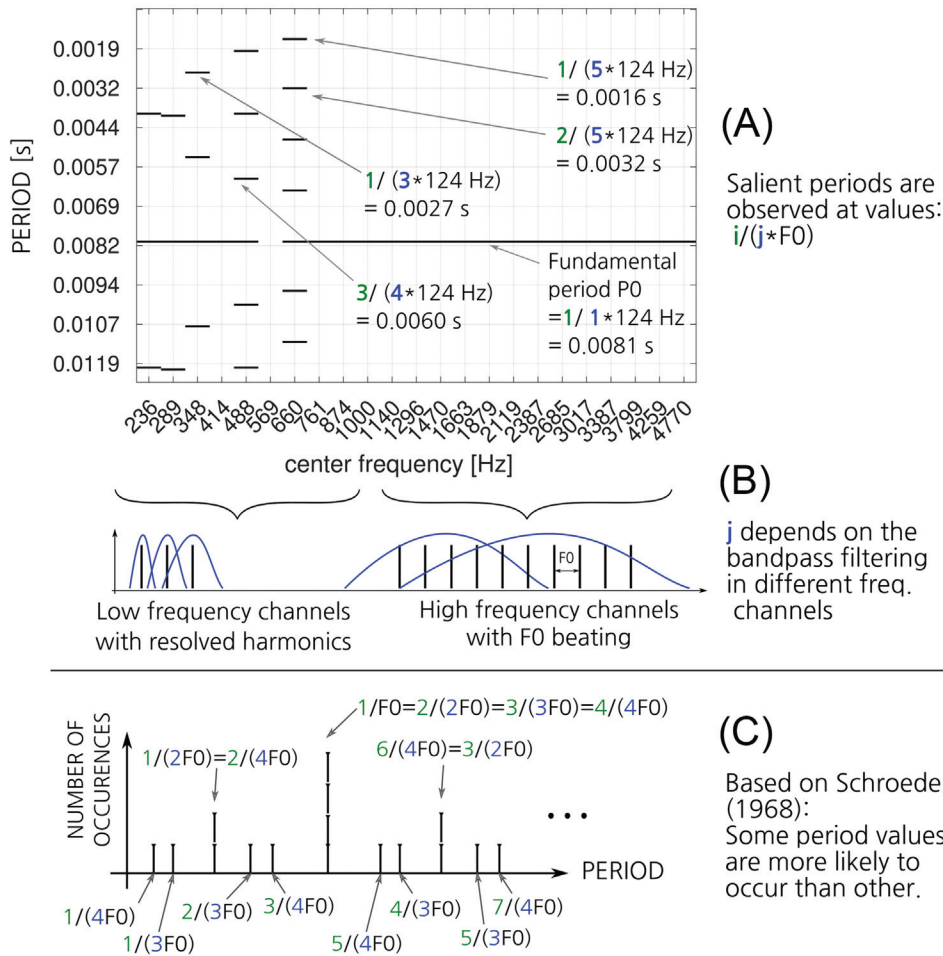


FIG. 6. The distribution of period glimpses. (A) The period glimpses extracted from a voice signal with  $F_0 = 124$  Hz (accumulated over 1 s) are shown. The observed values are related to the underlying  $F_0$  by the formula  $i/jF_0$ , where  $j$  is a harmonic number and  $i$  is an integer multiplier. (B) Which  $j$  is observed in a given frequency channel depends on the bandpass filtering. (C) As in the original plot by Schroeder, here, the period histogram for a signal consisting of the fundamental and first three harmonics is plotted. The same period values can originate from different harmonics of  $F_0$ . This makes some period values more likely to occur than others.

$F_1$ , and  $F_2$ . Plotting the histogram of the absolute period glimpses in seconds is difficult to interpret. There is an ambiguity related to the fact that  $P_{cnn}$  can assume any value equal to  $i/jF_0$ . Although all period multiples are the evidence of the same  $F_0$ , they can have different values in ms. To resolve the ambiguity with respect to the period multiple  $i$ , we can transform the period glimpses in seconds to the relative period glimpses, computed as

$$R_{cnn}(F_0) = \text{rem}\left(\frac{P_{cnn}}{P_0}\right) = \text{rem}(P_{cnn}F_0), \quad (11)$$

where  $P_0 = F_0^{-1}$  is the period of the hypothetical  $F_0$ , and  $\text{rem}(\cdot)$  is the remainder from the division. The relative period glimpses  $R_{cnn}(F_0)$  range from zero to one. After this transformation, all multiples  $i$  of the same period are mapped to the same value. What is remaining is the ambiguity about the harmonic number  $j$ .

Figure 7(A) shows the relative period histogram for a synthetic voice with varying  $F_0$ . The majority of the relative period values  $R_{cnn}(F_0)$  lie around zero and one. The next highest peak is at 0.5, followed by smaller peaks at 0.33 and 0.66 and further smaller peaks at 0.25 and 0.75. The interpretation is that the observed period glimpse is most likely

to be an integer multiple of the fundamental period:  $i/F_0$ , contributing to the peaks at zero and one. The second most likely value is a multiple of the period of the second harmonic of  $F_0$ :  $i/2F_0$ , contributing to the peaks at 0, 0.5, and 1. The next peak can be found for the multiples of the third harmonic,  $i/3F_0$ , contributing to the peaks at 0, 1/3, 2/3, and 1.

The advantage of the relative period histogram proposed here is that it summarizes the distribution for all of the  $F_0$  values, resolves the ambiguity about the period multiple  $i$ , and shows the probability of observing a glimpse from the harmonic number  $j$ . The form of the distribution was empirically found to be independent of  $F_0$  and the formant frequencies. The distribution has a different form in each frequency channel; however, in this work, we do not exploit these differences in the modeling (data not shown here).

**c. Mixture of von-Mises.** The distribution of the period glimpses can be modeled analytically as a mixture of 11 circular von-Mises distributions. The number 11 comes from the highest reported number of resolved harmonics (Bernstein and Oxenham, 2003). Each element of the sum represents a different harmonic  $j$  of  $F_0$ ,

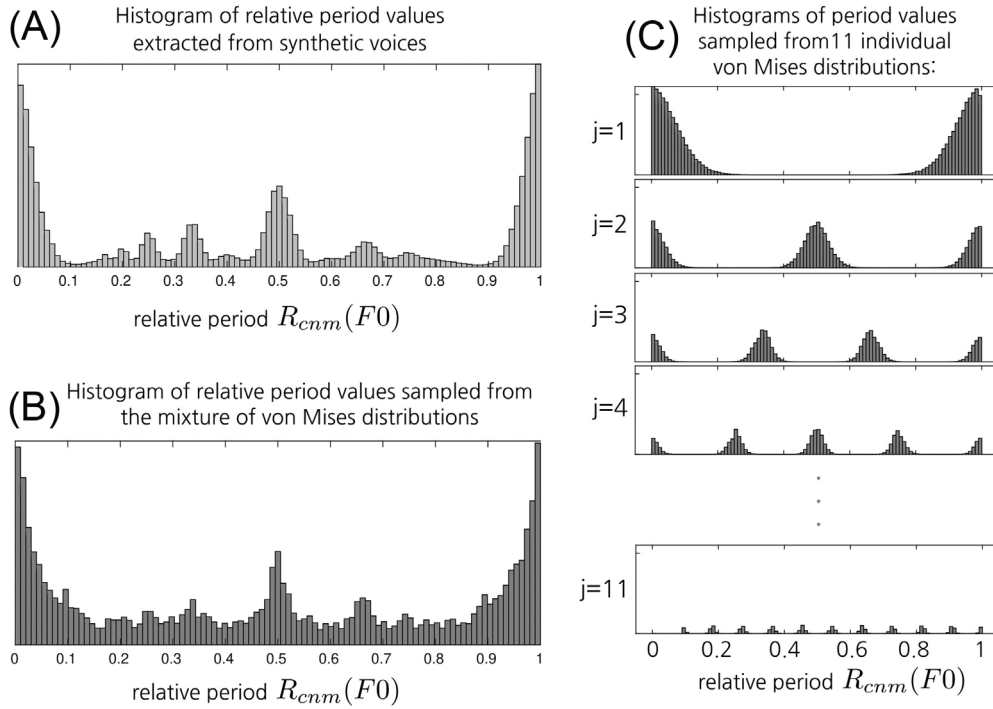


FIG. 7. The distribution of relative period glimpses. (A) The histogram of all relative period glimpses (relative to the underlying  $F_0$ ) in a 20 s synthetic voice signal with varying  $F_0$ ,  $F_1$ , and  $F_2$  is depicted. (B) The histogram of 1000 relative period glimpses sampled from the mixture of von-Mises distribution are shown. (C) Eleven histograms of 1000 relative period glimpses each, sampled from the individual von-Mises distributions contributing to the mixture, are shown.

$$p(P_{cnm}|F_0) = \sum_j^{11} C_j \mathcal{M}(R_{cnm}(jF_0)2\pi; \mu, \kappa), \quad (12)$$

where  $F_0$  is the hypothetical fundamental frequency,  $R_{cnm}(jF_0)$  is the relative period value with respect to the  $j$ th harmonic of the hypothetical  $F_0$ ,  $\mathcal{M}$  denotes the von-Mises distribution with the mean  $\mu = 0$  and concentration parameter  $\kappa = 5$ .  $C_j = j^{-1} / \sum_j^{11} j^{-1}$  is the normalizing constant for the  $j$ th harmonic. It is reciprocal to the harmonic number: The higher the harmonic number is, the lower the probability is of the period glimpse originating from that harmonic. Figure 8 explains the procedure of evaluating a single glimpse period.

Figure 7(C) shows the histograms of the periods sampled from each of the 11 von-Mises distributions contributing to the mixture. Figure 7(B) shows the samples from the mixture distribution in which every harmonic has a different prior [Eq. (12)]. The sampled values correspond well to the histogram from Fig. 7(A).

### III. MODEL EVALUATION

To evaluate the model, we simulated two conditions from the psychoacoustic study by Woods and McDermott (2015) and compared the model performance with the human performance. In this section, we explain how we obtain responses of the model in a psychoacoustic task, and we discuss the simulated experiments in more detail.

Furthermore, we simulated several additional conditions, which have not been tested by Woods and McDermott

(2015), on the human listeners. The results can be found in Appendix C.

#### A. Psychoacoustic study design

In the study by Woods and McDermott (2015), the participants were given the following task: After hearing a 500 ms *cue* signal indicating which voice should be attended, the 2 s long signal, containing two competing voices, was presented. At the end of a trial, a 500 ms *probe* signal, coming from one of the two voices, was presented, and the listeners had to decide whether or not the probe came from the attended voice (see Fig. 9, top). The performance was measured in terms of the sensitivity index ( $d'$ ).

We simulated this experimental procedure using the attentive tracking model. Each trial consisted of a pair of competing voices synthesized based on the state trajectories  $\mathcal{T}_F$  and  $\mathcal{T}_B$  [see Eqs. (1) and (2)]. The trajectory sampling rate was  $F_S = 50$  Hz, and  $N$  was set to 101, which corresponds to a signal length of 2 s.

The model's task was to track the fundamental frequency of the cued voice in an online manner, using the sPAF. Instead of presenting the *cue* signal explicitly to the model, we simulated the additional information that the cue delivers to the listeners. The tracking was initialized closely around the ground truth value  $F_{0F}(0)$  for the foreground voice and everywhere besides that value for the background voice (see Sec. II B 4 for details). Tracking yielded the estimated one-dimensional state trajectories  $\hat{\mathcal{T}}_F$  and  $\hat{\mathcal{T}}_B$  [see Eq. (6)].

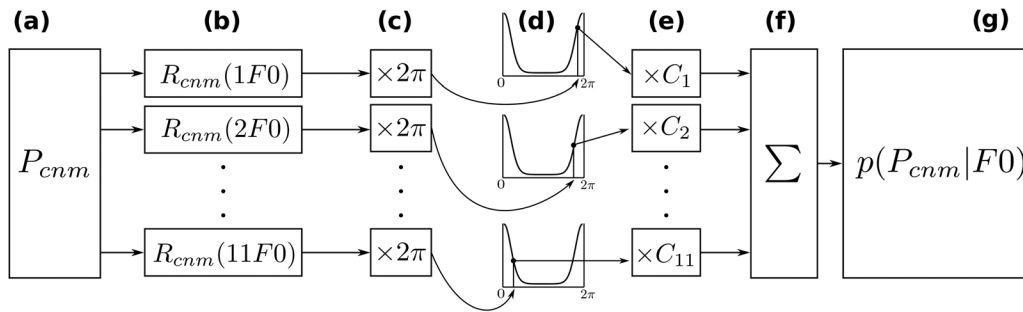


FIG. 8. The procedure for evaluating the observed glimpsed period values with a mixture of circular von-Mises distributions. Based on one observed period glimpse value (a), we compute 11 relative period values  $R_{cnm}(jF0)$  [Eq. (11)], where  $j$  is the harmonic number (b). Next, we multiply by  $2\pi$  to obtain a circular variable (c). The resulting 11 values are evaluated with the circular von-Mises distribution centered at 0 (d). Each likelihood is multiplied with a normalizing constant, which depends on harmonic number (e). The values are added (f), and the final result is the likelihood of a single period glimpse given hypothetical  $F0$  [see Eq. (12)] (g).

The next step was to obtain the model's response to a probe. We did not present the probe signal to the model. Instead, we compared the last 25 values of the estimated  $F0$  of the foreground (corresponding to the last 500 ms of the signal) to the last 25 values of the ground truth  $F0$  of the foreground voice and to the last 25 values of the ground truth  $F0$  of the background voice. We used the root mean square error as a distance measure,

$$\begin{aligned} \text{RMSE}^+ &= \sqrt{\frac{1}{25} \sum_{n=76}^{101} (\widehat{F0}_F(n) - F0_F(n))^2}, \\ \text{RMSE}^- &= \sqrt{\frac{1}{25} \sum_{n=76}^{101} (\widehat{F0}_F(n) - F0_B(n))^2}, \end{aligned} \quad (13)$$

where  $\text{RMSE}^+$  was a distance from a positive probe and  $\text{RMSE}^-$  was a distance from a negative probe. A  $\text{RMSE}^+$  or  $\text{RMSE}^-$  value within a tolerance range  $r$  was considered as

a model's positive response to a (positive or negative) probe. Otherwise, it was considered to be a negative response. We varied the criterion  $r$  and obtained the percentage of the true positive (TP) and false positive (FP) responses across all of the trials for each value of  $r$ :  $\text{TP}(r)$  and  $\text{FP}(r)$ . Plotting the  $\text{TP}(r)$  against  $\text{FP}(r)$  responses yields the receiver operating characteristics (ROC) curve, which illustrates the performance of a binary classifier for different discrimination thresholds. From the ROC curve, we could derive the  $d'$  value,

$$d' = \sqrt{2Z(\text{AUC})}, \quad (14)$$

where AUC is the area under the ROC curve, computed with the trapezoidal approximation and function  $Z(p)$ , where  $p \in [0, 1]$  is the inverse of a cumulative Gaussian distribution. The lower panel of Fig. 9 schematically shows a single trial of the simulated experiment.

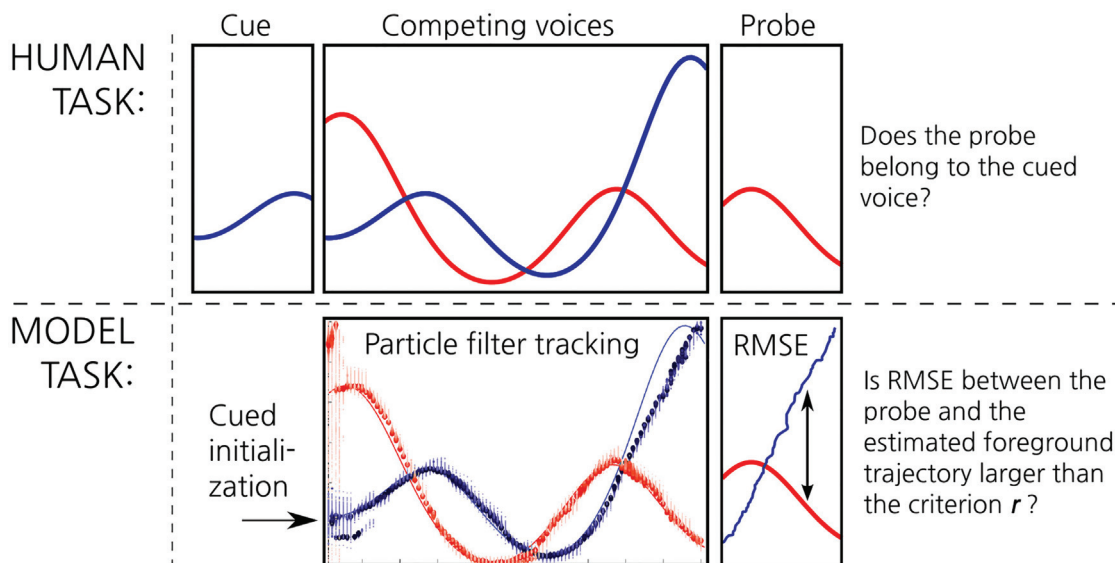


FIG. 9. The simulation of the psychoacoustic study with the attentive tracking model. (Top) The schematic view of one trial of the attentive tracking experiment performed with human listeners is shown. (Bottom) The method for simulating this experiment using the attentive tracking model is depicted. The size and color saturation of the dots correspond to the particle weights.



## B. Computational simulations

The following numerical experiments were performed in the scope of this paper (cf. Table I).

### 1. Simulation 1: Stream segregation of sources varying in just one feature

The goal of this computational simulation was to reproduce the listening experiment from Woods and McDermott (2015), originally called “stream segregation of sources varying in just one feature.” This experiment compared the attentive tracking performance for two types of stimuli, competing voices with trajectories varying in only one dimension ( $F0$ ) and competing voices varying in all three dimensions (see Fig. 10, rightmost panel). The results showed that the discrimination between the attended and unattended voice is well above a chance level when the parameters of the voices vary in all three dimensions, but solving the same task is not possible if the voices vary in only one dimension ( $F0$ ): The mean  $d'$  values dropped from  $d' \approx 1.2$  to  $d' \approx 0.2$  (see the magenta crosses in Fig. 10).

In this simulation, we used 100 random trajectory pairs varying in all three dimensions and 100 random trajectory pairs varying only in  $F0$ . For the trajectory generation, we used the same procedure as was used in Woods and McDermott (2015): The trajectory of each varying parameter ( $F0, F1, F2$ ) was generated independently by picking a random excerpt of Gaussian noise (500 Hz sampling rate), filtering it between 0.05 Hz and 0.6 Hz, and adjusting the value range. The trajectories crossed at least once in each varying dimension. To avoid the initialization conflicts, we

restricted the trajectories of the competing voices to start in different  $F0$  ranges (either 100–250 Hz or 250–400 Hz). Each trajectory pair was used twice as a trial. Each trajectory was once assigned the roles of attended and unattended voice, which resulted in 200 trials per condition. Based on the trials, a  $d'$  value was computed in each condition as described in Sec. III A. The simulation was repeated 20 times to account for the randomness in the initialization, prediction, and resampling steps of the particle filter (cf. Sec. II B 3). For each condition, 20  $d'$  values were obtained.

We obtained the model’s results in this task for three model variants.

*a. Simulation 1.a.  $F0$ -guided tracking.* In this condition, the foreground-background segregation was guided with the ground truth  $F0$  value used to synthesize the voices (see method B.2.1. in Fig. 3 from Sec. II C). This condition was employed to quantify the upper performance limit for the  $F0$ -guided feature segregation. We posed the following question: If the perfect estimations of  $F0_F(n-1)$  and  $F0_B(n-1)$  were available in every time step, would the glimpses decomposed based on this information be sufficient for the system to track the  $F0$  of the voices? Earlier results by Josupeit and Hohmann (2017) showed that sparse periodicity-based features were distinctive for all of the speakers in a multi-talker setup. Therefore, we expected that the sPAF would encode the information about both voices in a competing voices scenario (at least for the trajectories varying in all dimensions). We also wanted to test how much the performance can drop in this optimal case when the voices vary only in  $F0$  and have identical formants.

TABLE I. Overview of the computational simulations in Sec. III.

Simulation number	Corresponding human experiment	Foreground-background segregation method	Experiment conditions	Nr trials/condition	Nr runs/condition
1.a.	Stream segregation of sources varying in just one feature	$F0$ -guided segregation (Fig. 3, B.2.1.)	(1) Competing voices varying only in $F0$ (2) Competing voices varying in $F0, F1, F2$	200	20
1.b.	Stream segregation of sources varying in just one feature	Segregation without oracle information (Fig. 3, B.2.2.)	(1) Competing voices varying only in $F0$	200	20
1.c.	Stream segregation of sources varying in just one feature	Formant-guided segregation (Fig. 3, B.2.3.)	(2) Competing voices varying in $F0, F1, F2$	200	20
2.a.	Effect of source proximity	$F0$ -guided segregation (Fig. 3, B.2.1.)	Competing voices with minimum distance of (1) 0.5 semitones, (2) 2.5 semitones, (3) 5.5 semitones, (4) 7.5 semitones	100	10
2.b.	Effect of source proximity	Segreg. without oracle information (Fig. 3, B.2.2.)	Competing voices with minimum distance of (1) 0.5 semitones, (2) 2.5 semitones, (3) 5.5 semitones, (4) 7.5 semitones	100	10
2.c.	Effect of source proximity	Formant-guided segregation (Fig. 3, B.2.3.)	Competing voices with minimum distance of (1) 0.5 semitones, (2) 2.5 semitones, (3) 5.5 semitones, (4) 7.5 semitones	100	10

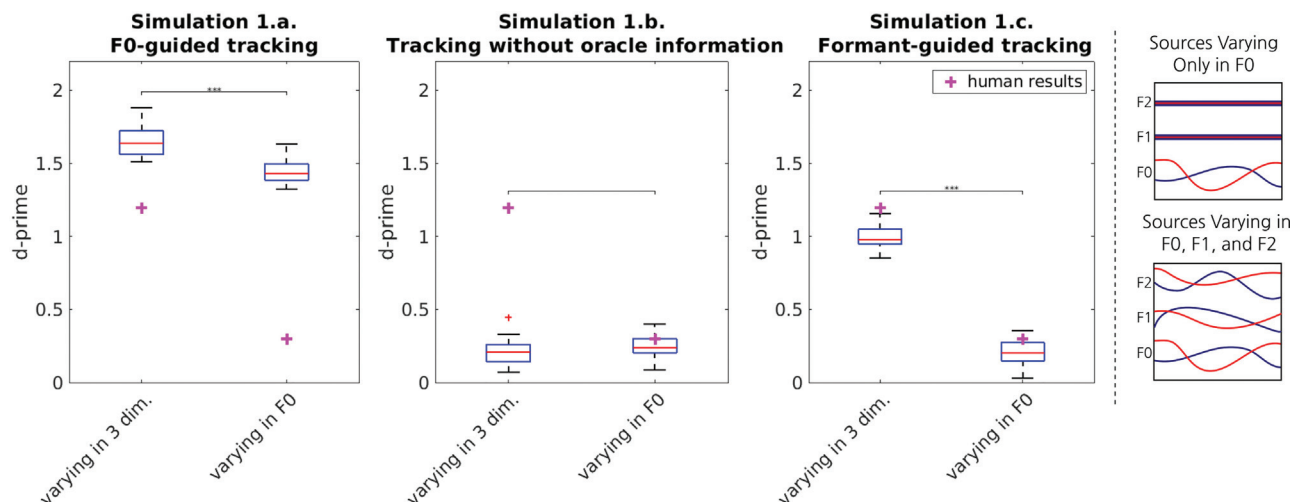


FIG. 10. Simulation 1: The stream segregation of sources varying in just one feature. (Left) The results of simulations 1.a–1.c. in terms of  $d'$ . (Right) The stimulus conditions used in simulation 1 in which the crosses in magenta denote the mean human results from Woods and McDermott (2015), the boxplots summarize the distribution across the 20  $d'$  values in each condition. The statistical significance of the difference between the tested conditions is measured with a  $t$ -test and represented in the plot with the symbols \*,  $p \leq 0.05$ ; \*\*,  $p \leq 0.01$ ; or \*\*\*,  $p \leq 0.001$ .

**b. Simulation 1.b. Tracking without oracle information.** This condition was used to evaluate the actual performance of the model with  $F0$ -based feature segregation. Here, we used the foreground-background segregation method based on the previous  $F0$  estimate (see method B.2.2. in Fig. 3 from Sec. II C). We quantified how robust the  $F0$ -tracking could be if we used the estimated  $\widehat{F0}_F(n-1)$  and  $\widehat{F0}_B(n-1)$  to decompose the glimpses. We expected the discrimination performance to drop in comparison to the  $F0$ -guided method from simulation 1.a. but to be above the chance level for the trajectories varying in three dimensions.

**c. Simulation 1.c. Formant-guided tracking.** With this condition, we investigated how much improvement, in comparison to simulation 1.b., could be gained by exploiting the oracle information about the formant frequencies in the foreground-background segregation stage. Here, the foreground-background segregation was based on the previous  $F0$  estimate and ground truth formant values used to synthesize the voices (see method B.2.3. in Fig. 3 from Sec. II C). We investigated how much the tracking performance would improve, if the perfect estimates of the formants  $F1_F(n-1)$ ,  $F2_F(n-1)$ ,  $F1_B(n-1)$  and  $F2_B(n-1)$  were available in each time step. We expected the discrimination performance in this condition to improve significantly in comparison to simulation 1.b. but not to exceed the model's performance from simulation 1.a.

## 2. Simulation 2: Effect of source proximity

The goal of this computational simulation was to reproduce the listening experiment from Woods and McDermott (2015), originally called “effect of source proximity.” In this experiment, the authors examined the influence of the proximity of competing voices on the attentive tracking performance. They compared eight conditions. In each condition, the trajectories were restricted to pass each other with a

different minimum distance in the three-dimensional feature space. The results showed that the discrimination between the attended and unattended voices improved continuously as the voice distance was increased (see Fig. 12).

We tested four conditions with different minimum distances. In each condition, the minimum three-dimensional Euclidean distance between the trajectories (in semitones) was restricted to fall within the designated bin limits. The in limits were 0–1, 2–3, 5–6, and 7–8 semitones for the minimum distance conditions 0.5, 2.5, 5.5, and 7.5 semitones, respectively. For the trajectory generation, we used the same procedure as in simulation 1 with the difference that the Gaussian noise was filtered between 0.05 and 0.3 Hz as in Woods and McDermott (2015). All of the trajectories crossed at least once in each dimension. We initialized the trajectories of the competing voices in different  $F0$  ranges (either 100–250 Hz or 250–400 Hz). Per each condition, 50 random trajectory pairs were generated. Each trajectory pair was used twice as a trial (each trajectory was once assigned the roles of attended and unattended voice), which resulted in 100 trials per condition. Based on the trials, a  $d'$  value was computed in each condition as described in Sec. III A. The simulation was repeated ten times to account for the randomness in the initialization, prediction, and resampling steps of the particle filter (cf. Sec. II B 3). For each condition, ten  $d'$  values were obtained.

Following the concept from simulation 1, in this task, we also obtained the model's results for three model variants.

**a. Simulation 2.a.  $F0$ -guided tracking.** We expected to show the potential of the correctly segregated sparse periodicity-based features.

**b. Simulation 2.b. Tracking without oracle information.** We analyzed the limitations, which the model

encounters when attempting to solve the attentive tracking task, only based on the estimated  $F_0$ .

*c. Simulation 2.c. Formant-guided tracking.* We performed this simulation to test how much improvement could be brought to the model if the segregation stage was based on the perfect formant information.

## IV. RESULTS AND DISCUSSION

### A. Simulation 1

In this section, the results of simulation 1 are presented, which modeled the listening experiment *stream segregation of sources varying in just one feature* (see Sec. III B 1).

#### 1. Simulation 1.a

With this simulation, we aimed at validating the periodicity-based glimpses in the context of the  $F_0$  tracking. The median  $d'$  value across all runs of the simulation was  $d' = 1.64$  for the voices varying in three dimensions and  $d' = 1.42$  for the voices varying only in  $F_0$  (see Fig. 10, left). A good discrimination performance indicates that the system could track the  $F_0$  trajectories of both voices. These results prove that the glimpses segregated based on oracle information about  $F_0$  in the preceding time step contain sufficient information to segregate the foreground and background glimpses and estimate the pitch of two simultaneous voices. In this simulation, the model outperforms the human listeners. It uses oracle information, which was not available to the listeners. This shows that the information content of the sPAF is well above what is needed to explain the human performance.

#### 2. Simulation 1.b

We expected that the discrimination performance would decrease when we replace the oracle  $F_0$  with the estimated  $F_0$ . Nevertheless, after seeing the successful discrimination results in simulation 1.a., we expected it to remain above the chance level for most of the conditions. However, as shown in Fig. 10 (middle), the results dropped significantly in comparison to simulation 1.a. The median  $d'$  value across all runs of the simulation was  $d' = 0.39$  for the voices varying in three dimensions and  $d' = 0.21$  for the voices varying only in  $F_0$ .

The foreground-background segregation in this model version depends on the  $F_0$  estimates of the foreground and background voices. When at least one of the estimates is not correct, a particle filter responsible for one voice may receive the channel sets  $P_{cn}$ , which originated from the other voice. Because the model assumes that the observation is reliable, the particle filter always treats the incoming data as valid evidence and updates its particles based on it, leading to a false estimation. As in the subsequent step, the segregation again depends on the estimated  $F_0$ , and the error propagates potentially until there are no more particles covering

the true  $F_0$  region. When this happens, it is virtually impossible for the particle filter to get back on the correct track.

Looking more closely at the tracking results, we detected different ways in which the algorithm without any oracle information could typically be misled.

- identity switches at the  $F_0$  crossings [see Fig. 11(A)]. At the  $F_0$  crossings, the foreground particle filter could take over the tracking of the background voice and/or vice versa. When the  $F_0$  of both voices is the same, the channel sets  $P_{cn}$  of the foreground voice are equally likely to be forwarded to the foreground as to the background particle filter. This problem cannot be resolved by the continuity of the model. Apart from the state transition model, which is responsible for redistributing the particles at every time step based on two previous time steps, there is no additional mechanism that would prevent the estimated tracks from changing direction.
- Tracking the (sub)harmonics of the correct  $F_0$  [see Fig. 11(B)]. The second problem was related to tracking the harmonics or subharmonics of the correct  $F_0$ . For example, the period glimpses originating from  $F_0 = 115$  Hz, in general, convey a relatively high likelihood for the hypothesis that  $F_0 = 330$  Hz. At some time instances, where there were only a few period glimpses available, the likelihood for the hypothesis  $F_0 = 330$  Hz might have even exceeded the likelihood for  $F_0 = 115$  Hz. In that case, the particle set concentrated around the incorrect  $F_0$  region. Without any particles left in the region of the true  $F_0$ , a particle filter could only track the closest harmonic.
- Identity the switch at the sub(harmonics) of the correct  $F_0$  [see Fig. 11(C)].

The third reason was a combination of the two aspects mentioned above. A particle filter could start to track a harmonic or subharmonic of a competing voice. This can be seen as an identity switch at the places where the  $F_0$  of one voice crosses with a (sub)harmonic of the second voice.

In summary, the tracking can potentially be misled at every point where the  $F_0$  trajectories or their harmonics or subharmonics cross [see Fig. 11(D)]. We concluded that using solely  $F_0$  estimates to segregate the observation extracted from the competing voices was not sufficient and did not reproduce the human results.

#### 3. Simulation 1.c

In this simulation, we investigated whether the additional information about the formant frequencies could possibly prevent the period glimpses from being assigned to the wrong stream. As expected, with the formant-guided tracking, we obtained a significant improvement in the discrimination performance in comparison to simulation 1.b. without oracle information. The median  $d'$  value across all runs of the simulation was  $d' = 0.98$  for the voices varying

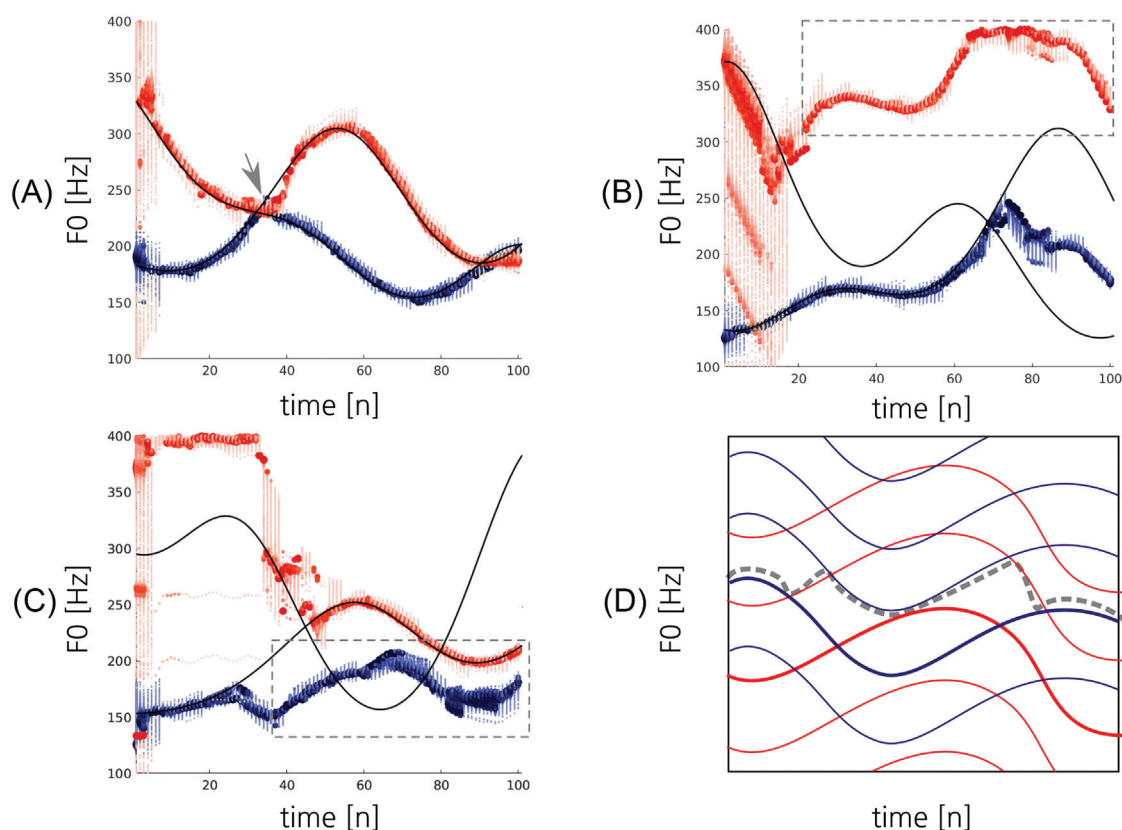


FIG. 11. The typical tracking errors encountered in simulation 1.b. Each plot presents a single trial of the attentive tracking experiment. The blue color indicates the foreground voice, and the red color indicates the background voice. The size and color saturation of the dots in plots (A)–(C) correspond to the particle weights. The gray dashed rectangles help to locate the discussed error in the image. (A) the identity switch at the  $F_0$  crossing, (B) following a harmonic of a  $F_0$  of the competing voice, and (C) following a subharmonic of the correct  $F_0$ . (D) The solid lines illustrate the trajectories of the harmonics or subharmonics of the true  $F_0$ , and the dashed line represents the potential estimated trajectory. The tracking can potentially be misled at every intersection point.

in three dimensions and  $d' = 0.2$  for the voices varying only in  $F_0$  (see Fig. 10).

Based on the instantaneous energy distribution across the frequency channels, this foreground-background segregation method determined which channels were more likely to contain period glimpses of which voice. This additional prerequisite prevented the period glimpses from the wrong channel from leaking into the observation. Even if both voices temporarily shared the same  $F_0$ , a difference in  $F_1$  and  $F_2$  helped to disentangle the origin of the period glimpses in a given channel. An advantage of this method was not expected for the trajectories varying only in one dimension. Lacking the separation in the  $F_1$  and  $F_2$  dimensions, the model had to rely purely on the estimated  $F_0$  as in simulation 1.b. The simulation results were in agreement with this expectation—the discrimination performance for the voices varying only in  $F_0$  was at the chance level. The same effect was observed for humans. Hence, the results obtained in this simulation were in the best agreement with the human results.

## B. Simulation 2

In this section, the results of simulation 2 are presented (see Fig. 12). This simulation modeled the listening experiment *effect of source proximity* (see Sec. III B 2).

### 1. Simulation 2.a

In this simulation, the oracle  $F_0$  tracks were used to segregate the sPAF. The median  $d'$  values were (from small to large minimum distance) 2.07, 1.99, 2.35, and 2.36. The corresponding  $d'$  values for humans were 0, 0.3, 1.3, and 2. The model clearly outperforms the human listeners. Even in conditions with a low minimum distance, it reaches a very good discrimination performance. This shows that the sPAF, even when extracted from a mixture of two voices lying very close to one another in the feature space, contain sufficient information to segregate and track them. Because the feature segregation stage in this simulation is  $F_0$ -guided, the results represent an optimal case. The increase in  $d'$  as a function of the minimum distance was not observed. For most of the conditions, the model is much better than the listeners, who do not use any oracle information to solve this task. For the highest minimum distance,  $-7.5$  semitones, humans reach almost optimal (according to the model) performance, meaning that they have no difficulty in segregating the voices, as if they were able to perfectly estimate the  $F_0$  tracks.

### 2. Simulation 2.b

In this simulation, no oracle knowledge was used to segregate the sPAF. The median  $d'$  values were (from small



to large minimum distance) 0.16, 0.45, 0.55, and 0.52. The corresponding  $d'$  values for humans were 0, 0.3, 1.3, and 2. The results for the first two conditions—minimum distances of 0.5 and 2.5 semitones—are in a good agreement with the human results. The discrimination performance is low, indicating that the task is difficult. Nevertheless, increasing the minimum distance between the trajectories from 0.5 to 2.5 semitones results in a significant improvement. For the lowest minimum distances between the trajectories, humans lose the ability to distinguish between the sources. The reason might be the limited resolution of attention or the perceptual fusion of the voices when they take on similar feature values. Our model without any oracle information seems to represent well these adverse conditions in which the resolution of the auditory system is limited. It might indicate that for the low minimum distances, the attentive tracking in humans is prone to similar errors as for the model without oracle information (see Fig. 11).

In the next two conditions with a minimum distance of 5.5 and 7.5 semitones, the model is not capable of reproducing the human results. At the distance of 5.5 semitones, the model cannot reach the human performance, which is consistent with simulation 1.b. Increasing the distance from 2.5 to 5.5 semitones results in a huge improvement for the human listeners, and only a slight, although significant, improvement for the model. Somewhere between the distance of 2.5 and 5.5 semitones, humans start to use the available information to separate the features and segregate the voices. The model, on the other hand, is still erroneous and not capable to segregate the features at the crossings.

Interestingly, the model results for the minimum distance of 7.5 semitones are slightly worse than those for 5.5 semitones. A possible explanation could be that increasing the distance raises the likelihood that the model will follow a track one octave away from the correct  $F_0$ .

### 3. Simulation 2.c

In this simulation, the oracle information about the formants of the voices was used to segregate the sPAF. The median  $d'$  values were (from small to large minimum distance) 0.46, 0.83, 1.02, and 1.16. The corresponding  $d'$  values for humans were 0, 0.3, 1.3, and 2. In the first two conditions—minimum distances of 0.5 and 2.5 semitones—the model outperforms the human listeners. The model is given more information than required to explain the human performance. Contrary to the model, the listeners did not know the formant tracks and had to estimate them to segregate the voices.

For the minimum distance of 5.5 semitones, the model coincides with the human results. This minimum distance provides enough separation in the feature space, allowing humans and the model to use the formant frequencies to segregate the voices. However, a further increase in the minimum distance does not make the model perform better in this task. This shows a limited resolution of the formant-guided segregation method. Even with the perfect knowledge of the formants, which allows us to predict the energy distribution across the frequency channels, the sPAF cannot be segregated in an optimal way. We know from the results of simulation 2.a. that the model's performance could be further increased if the likelihood of the correctly estimated  $F_0$  was used in the foreground-background segregation stage. This requires a model of the joint likelihood for  $F_0$ ,  $F_1$ , and  $F_2$  and three-dimensional tracking.

## V. GENERAL DISCUSSION

In the present study, we combined the concepts of the auditory glimpses, statistical top-down knowledge, and Bayesian inference to model the attentive tracking of the

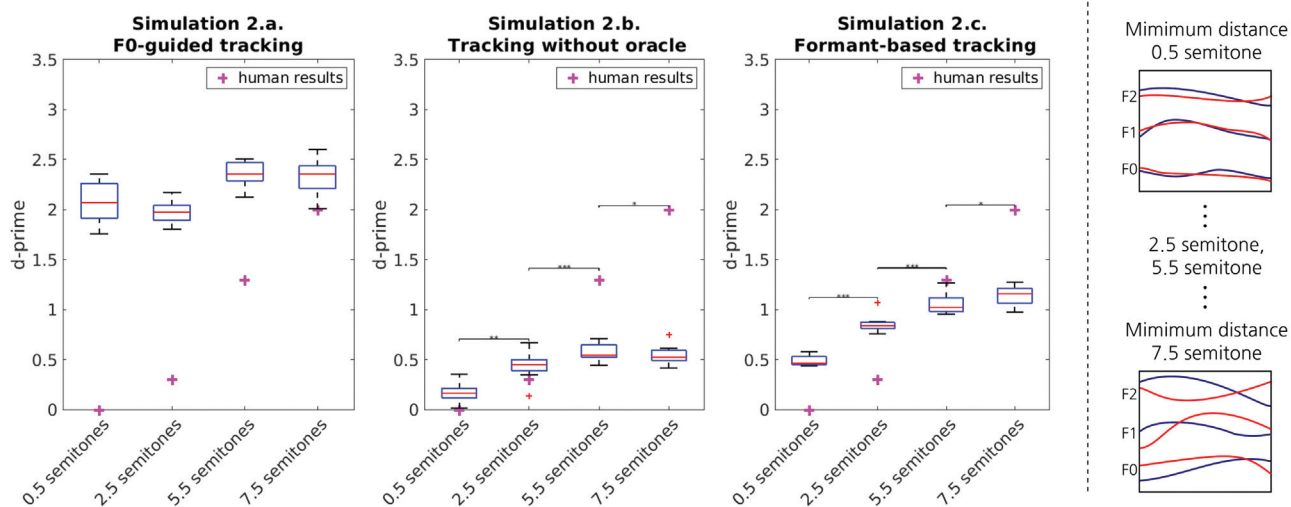


FIG. 12. Simulation 2: The effect of the source proximity. (Left) The results of simulations 2.a.–2.c. in terms of  $d'$  are shown. (Right) The stimulus conditions used in simulation 2 are shown. The crosses in magenta denote the mean human results from Woods and McDermott (2015). The boxplots summarize the distribution across ten  $d'$  values in each condition. The statistical significance of the difference between the tested conditions is measured with a  $t$ -test and represented in the plot with the symbols \*,  $p \leq 0.05$ ; \*\*,  $p \leq 0.01$ ; or \*\*\*,  $p \leq 0.001$ .

voices. We developed a unified computational framework, consisting of the sparse periodicity-based feature extraction, single voice probability models, and particle filter tracking. We evaluated the model using the attentive tracking paradigm proposed by Woods and McDermott (2015).

#### A. Attentive tracking model in relation to human data

Our aim was to reproduce two experiments from Woods and McDermott (2015), which showed that the human listeners can attentively track a voice in the presence of a second voice if their parameters maintain sufficient separation in the parameter space. The experiment *effect of source proximity* showed that attentive tracking gets worse as the distance between the varying voice parameters decreases, and the experiment *stream segregation of sources varying in just one feature* showed that it fails when the voices vary only in one dimension.

Whether we could predict these aspects of attentive tracking with the computational framework proposed here depends on the foreground-background segregation stage of the model. The foreground-background segregation based on the oracle  $F0$  leads to a performance much better than that of the human listeners. This showed that even for conditions with only one varying feature or very close parameter transitions, there is enough information in the acoustic signal to solve the attentive tracking task. The sparse periodicity-based features extracted from the mixture signal provide evidence of two distinct voices, which is sufficient to track these voices. However, because of the limited resolution of the auditory system, the listeners have trouble encoding this information when the voices pass close in the parameter space.

The adverse conditions, in which humans had difficulties attentively tracking a voice, were most accurately reproduced by the model with the foreground-background segregation based on the estimated  $F0$  without any oracle knowledge. In these conditions, our model suffers from the confusions at the crossings of  $F0$  or crossings of the harmonics of  $F0$ . In the absence of the information about the formants of the voices, the ambiguity at the crossing cannot be resolved. Two voices share the frequency space, and after extracting and segregating the sPAF, the remaining evidence for one voice is typically incomplete. This incomplete pattern can produce a higher likelihood for a wrong hypothesis. Various studies mention the multimodal distributions of the pitch matches for the complex tones with few harmonics (Boer, 1956; Cariani and Delgutte, 1996; Schouten *et al.*, 1962; Terhardt, 1989). Additionally, models of the pitch of concurrent sounds also suffer from the ambiguous  $F0$  estimates (Assmann and Summerfield, 1990; Saddler *et al.*, 2021; Schouten *et al.*, 1962). Ambiguity does not mean randomness; in all of the mentioned studies, the pitches evoked by a stimulus are in systematic relationships with each other (they lie at the harmonics and their submultiples). In conclusion, any tonal sound other than a pure tone, especially complex tones lacking some harmonics, are more or less

ambiguous in pitch. Hence, the errors of the model are consistent with the observations in humans. According to the simulations, the results segregation seems to be the major limitation, not the masking.

Considering this inherent ambiguity of the pitch, there must be additional mechanisms that help the listeners solve the attentive tracking task. The foreground-background segregation with oracle information about the preceding formants allowed us to reproduce the human results in the conditions in which human performance is good but still not optimal (not reaching the results of the  $F0$ -guided model). This supports the conclusion of Woods and McDermott (2015), which to attentively track a voice, the auditory system binds several task-related qualities together—in this case, all of the varying dimensions  $F0$ ,  $F1$ , and  $F2$ . The results confirmed that it may be possible to improve the model's performance using a multidimensional likelihood model comprising  $F0$ ,  $F1$ , and  $F2$ . To achieve this, the interdependencies between these parameters and their impact on the sPAF have to be investigated. However, an extension of the model to blindly track the information in multiple dimensions is beyond the scope of the current study and subject to future work (see Sec. V B).

Apart from the simulated experiments, there are other studies with relevance in the context of modeling the attentive tracking.

Madsen *et al.* (2019) found differences in the attentive tracking between musicians and nonmusicians, which were not found for the speech perception in noise. The authors argued that the advantage of the musicians can be due to their experience in making fine-grained auditory discrimination judgments. In our model, the experience of the musicians could be reflected in the computational block *top-down knowledge*. For the musicians, the probabilistic voice models—*state transition* and *observation statistics*—might be more accurate than that for the nonmusicians. With an accurate model, the foreground can be separated from the background in a more optimal way, which eventually leads to an improved ability of following one of two competing voices.

Attentive tracking in a more ecologically valid scenario has recently been investigated by Siedenburg *et al.* (2021). The task of normal hearing and hearing impaired participants was to track the individual musical voices in JS Bach's *The Art of the Fugue*. The performance depended on the degree of hearing impairment, number of voices in the mixture, and timbral heterogeneity between the voices. Simulating the timbral heterogeneity might require tracking more than the parameters  $F0$ ,  $F1$ , and  $F2$ , but, at least on a conceptual level, this task can be well defined in the proposed computational framework. To simulate the effect of the hearing impairment, the parameters of the sPAF extraction could be modified such that the features reflect the defective auditory resolution. The influence of the number of voices in the mixture would be mirrored in the feature segregation stage. The more voices there are in the background stream, the harder it is to separate them from the foreground.

Another aspect of attentive tracking was presented in a follow-up study by Woods and McDermott (2018). The results of this study showed that within a relatively short time, humans can assimilate a repetitive schema in the attended sound and use it to solve the task. Schema learning occurred even when sources never appeared in isolation and despite the fact that the schema appeared transposed or dilated/compressed to varying degrees. In our model, the probability distributions of the *top-down knowledge* are related to the short-term voice changes. The schema learning seems to be related to a different time scale, in which the statistical regularities across the trials can be observed. To simulate this phenomenon, additional modality, responsible for the adaptation of the top-down knowledge, could be implemented. The adaptation component seems natural to add in the further versions of this model because it would be required for simulating other tasks, including memory, novelty processing, or bottom-up attention.

With the examples above, we demonstrate that our model has a generic structure, which could be used to simulate experiments other than the experiment simulated in this study. Our framework is comprised of various sub-tasks, which could be adapted or extended depending on the complexity of the simulated auditory scene.

## B. Attention in the computational framework

Attention is a process of focusing limited neural resources on the region of perceptual interest. Perceptual interest can be chosen consciously (top-down attention) or evoked unintentionally as a consequence of the unexpected external stimulus (bottom-up attention). The attentive tracking model presented in the current study simulates the top-down attention and how it is maintained over time. The limited neural resources are represented as a finite set of sampled hypotheses in the particle filter. The listener's conscious choice to focus the attention on the cued voice is simulated by initializing the particles in an informed way. Maintaining the attentive resources on the voice of interest throughout the duration of the signal is represented by the resampling step in a particle filter.

Algorithmically, the resampling eliminates the particles with small weights and duplicates the particles with high weights. This way the hypotheses keep concentrating around the region of the highest importance. Without the resampling step, the hypotheses would freely evolve in the state space, according to the state transition model. After several iterations, the particles would spread, and they would not be concentrated around the voice of interest any longer. In that scenario, initializing the particles around the correct value would only be helpful in the first few iterations. However, when the resampling is active, the effect of proper initialization lasts throughout the whole signal: The particles maintain their focus on the voice of interest.

## C. Relation to previous work

This article builds on the previous work in the field of auditory signal processing. Especially worth mentioning is

the work on sequential estimation in the auditory scene (Nix and Hohmann, 2007; Nix *et al.*, 2003; Spille *et al.*, 2013) and a series of modeling studies related to the sPAF (Josupeit and Hohmann, 2017; Josupeit *et al.*, 2016; Josupeit *et al.*, 2020).

In the study by Nix and Hohmann (2007), multidimensional, nonlinear statistical filtering was proposed as a tool for filtering the disturbed speech as well as a plausible model of feature binding in the auditory scene. Particle filters with a codebook-based prediction and update step and observation in the form of short-time speech spectra were used to track the location and spectral shape. The second line of research investigated the role of the periodicity in the auditory scene and developed the sparse periodicity-based feature extraction method (Josupeit and Hohmann, 2017). In this second study (Josupeit and Hohmann, 2017), an auditory model based on these features was used to predict the human results of a multi-talker communication performance test. The model of Josupeit *et al.* (2020) was used to predict the spatial release from masking in humans.

Although both models mentioned above were shown to be powerful in illustrating the principles of the auditory scene, the first model lacked a more specialized feature space and the latter model did not include sequential processing. With this study, we bridge these ideas. We combine the statistical filtering with the periodicity-based features to reproduce the results of yet another psychoacoustic task—following a voice.

## D. Model limitations and future development

There are several directions in which we will develop the current model in the future.

### 1. Multidimensional tracking

To demonstrate the feasibility of the approach, we used the one-dimensional state space. We modeled the attentive tracking task as tracking of the fundamental frequency. Although many studies show that vowel segregation relies on the fundamental frequency estimation, we have shown that more dimensions are required to segregate the observed features. To improve the model of the acoustic feature binding, high-dimensional state tracking should be performed. This requires knowledge of the nonlinear relationship between the multidimensional state and the observation space. In the future, we will learn these features from the data, for example, using deep learning techniques. In a recent study (Luberadzka *et al.*, 2020), we showed that using a deep regression network, we can learn the mapping between the three-dimensional state space and sPAF of a single voice. Incorporating the learned mapping into the tracking system would enable the full use of the potential of the particle-filtering- and periodicity-based features, which was only partly exploited in this study.

### 2. Background representation and attention modeling

The model presented in this study assumes that the auditory scene consists of two streams: the attended



foreground and unattended background. We believe that this distinction is valid in the majority of the auditory scenes. However, the current model is limited by the fact that there is no difference in the tracking of the foreground and background. Apart from the initialization with the attention prior model, the particle filters have the identical properties. Both execute the resampling step, which mimics the attentive tracking. In other words, the model attentively tracks the foreground and background. This is a simplified approach, which does not reflect the differences between the foreground and background processing in the auditory system. Moreover, the unattended stream is assumed to have the exact same properties as the attended stream, which in the model is reflected by the identical top-down models for the foreground and background. This is true for the competing voices scenario but does not hold for a scenario with different background sounds.

When attentively tracking the target voice, humans still perceive the sounds from the background. However, if a person focuses the attention on the foreground, the perception of the background stream is less sharp than that of the foreground stream. It was demonstrated in Woods and McDermott (2015) by the difference in the detection of vibrato between the attended and unattended voices. This difference was only found for good streamers, who were able to successfully perform the attentive tracking task.

Although, at this point, the attention-driven differences in the perception of the foreground and background could not be simulated with our model, this limitation should be tackled in the future research, for example, by reproducing the experiment with the vibrato detection. Instead of using two particle filters with resampling, the resampling could be used in the foreground particle filter alone. The background particle filter could be updated without resampling, providing the information about the background statistics but keeping the hypotheses broadly distributed.

### 3. Realistic stimuli

The continuous competing voice signals with the time-varying parameters, which are used to evaluate the current model, are simple but challenging stimuli. One voice is assigned to the attended foreground, and the other voice is assigned to the background. On the one hand, the acoustic scene seems to be simple. There are no background noises or reverberation. Both voices are continuously active and do not contain any consonants, the information about the periodicity is available in every time instance, and is not disturbed by any additional signals or pauses. On the other hand, the voices are simultaneously active throughout the whole stimulus duration, meaning that they always “share” the frequency space. In a more realistic scenario, we would expect not only much more disturbance from the acoustic environment but also, due to the sparsity of the speech signal, many more *T-F* windows with one clearly dominating voice. In future work, we plan to test the model using more realistic signals containing speech.

## E. Auditory model among machine learning advances

In recent years, machine learning has gained a huge interest in various scientific fields, including audio signal processing. Deep supervised networks are used directly on the audio time signal to perform all sorts of tasks, including source segregation and tracking (Purwins *et al.*, 2019). These techniques have also been used to predict human performance in psychoacoustic tasks (Kondo *et al.*, 2018; Spille *et al.*, 2018). These “end-to-end” methods have their merits and show an impressive performance when it comes to applications. However, one big criticism is that they provide hardly any insight into how the problem has been solved by the network. In this work, we focus on the explanatory rather than predictive power of auditory modeling. Our modeling framework was developed to provide a plausible computational illustration of how the auditory system might analyze the acoustic scene. It provides insight into the different sub-tasks of auditory scene analysis. All of the building blocks of the model can be extended toward deep learning. This way, the modeling framework could benefit from the current technological advances without sacrificing the existing conceptual structure.

## VI. CONCLUSIONS

In this paper, we presented a computational model of attentive voice tracking, which unifies the concepts of auditory glimpses, sequential Bayesian inference, and perceptual organization of the auditory scene into foreground and background streams. We used an acoustic scene containing two competing voices with time-varying parameters to demonstrate the processing steps. We implemented the model as a combination of the sPAF, sequential Monte Carlo sampling, and single voice probability models. We proposed an *F0* observation model, which describes the statistical relationship between the generative fundamental frequency and observed period glimpses. A comparison of the model with the human performance showed that although the optimally segregated sPAF convey sufficient information to perform the attentive tracking task, humans are not always able to decode this information. The joint modeling of *F0* and the spectral profile (formants) is required to reach the human performance levels. This shows that a combination of features may be used by the auditory system to correctly segregate the spectro-temporal glimpses and attentively track a voice through acoustic space.

## ACKNOWLEDGMENTS

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project No. 352015383, SFB 1330. The research reported in this publication was supported by the National Institute On Deafness and Other Communication Disorders of the National Institutes of Health under Award No. R01DC015429. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.



## APPENDIX: A SPARSE PERIODICITY-BASED FEATURE EXTRACTION

We use the feature extraction scheme developed by Josupeit and Hohmann (2017). In the original approach, the four types of periodicity-based auditory features were extracted separately for a number of frequency bands from the multi-talker input signal: period  $P_{cns}$ , related to the pitch of a sound; periodic energy  $E_{cns}$ , related to the spectral shape of the periodic sound; and periodicity-based interaural time and level differences,  $T_{cns}$  and  $L_{cns}$ , related to the azimuthal sound source location.  $c$  indicates the frequency band number,  $n$  is the time index, and  $s$  is the channel of a binaural signal. In the current study, we investigated whether these features can be used to track the  $F0$  of the voices. Hence, from all four of the periodicity-based features proposed by Josupeit and Hohmann (2017),  $P_{cns}$ ,  $E_{cns}$ ,  $T_{cns}$ , and  $L_{cns}$ , we use only the pitch-related  $P_{cns}$ . Because we use the single-channel input, we can get rid of the index  $s$  and adopt the notation  $P_{cn}$  into the current modeling framework.  $P_{cn}$  is a *channel set*, which contains a varying number of *period glimpses*  $P_{cnm}$ .

The extraction of the period glimpses  $P_{cnm}$  consists of the following steps [see Fig. 13(A)].

### 1. Auditory preprocessing

One channel of a binaural input signal is passed through a middle ear bandpass filter (500–2000 Hz) and gammatone filterbank with 23 channels  $c = 1, \dots, 23$  and center frequencies between  $f_c = 200$  and  $f_c = 5000$  Hz with 1 ERB distance and a filter width of 1 ERB. Next, the cochlea power-law compression with an exponent of 0.4 and hair-cell processing, using a half-wave rectification and a 770 Hz low-pass filter, is applied in each frequency channel (Dietz et al., 2011). The waveform and spectral shape of the signal processed by the hair cells are altered by half-wave rectification: A hair-cell-processed band has a broadened spectrum, including a direct current (DC) component, the demodulated envelope, and usually energy in the frequency region of the original band limited signal (Dietz et al., 2008). Hence, an additional spectral limitation of the hair cell stage output is required. Josupeit and Hohmann (2017) obtained this by introducing an additional band-specific filtering (gammatone filters with center frequency equal to  $f_c$  and a bandwidth of  $f_c/3$  for the *fine structure channels* with center frequencies  $f_c < 1400$  Hz and a gammatone filter with a constant center frequency of 135 Hz and a bandwidth of 16.9 Hz for the *envelope channels* with center frequencies  $f_c > 1400$  Hz) and differentiation in the time domain. Here, we replace these steps by filtering the half-wave rectified signals in all of the frequency bands with a 40 Hz high-pass filter, which removes the spectral components below the typical range of the pitch frequency, including the DC component. The advantage of this modification is that there is no hard division between the fine structure and modulation channels: All of the frequency channels undergo the same transformations, which reduces the number of free parameters in the

model. Not separating between the envelope and fine structure means that the beating from unresolved harmonics and resolved harmonics contribute with the same weight to the  $F0$  estimate. This is a simplification as these two components may be weighed differently in the auditory system (Dietz et al., 2008).

### 2. Periodicity analysis

Next, the periodic structure of the waveform is analyzed with the normalized synchrogram technique (Hohmann, 2006). The preprocessed waveforms of each frequency channel are analyzed every 20 ms as follows: Around each considered time step  $n$ , eight signal segments of duration  $P'$  are formed as depicted in Fig. 13(A)(2).  $P'$  is varied from 1/700 to 1/80 Hz in 1/16000 Hz steps, resulting in 178 *tested periods*. The eight signal segments are averaged, yielding a *base function*  $v_{cn}(P')$ . The energy of the waveform that spans all eight signal segments is termed the total energy. It is calculated as the mean square amplitude. The energy of the base function is called the *periodic energy*  $E_{P,cn}(P')$  and corresponds to the mean square amplitudes of the base function. Last, for each point in time  $n$ , channel  $c$  and each tested period  $P'$ , the *normalized periodic energy*  $\text{synch}_{cn}(P')$ , defined as the ratio of the periodic energy and total energy, is computed, which is called the *synchronism*,

$$\text{synch}_{cn}(P') = \frac{E_{P,cn}(P')}{E_{\text{tot},cn}(P')}. \quad (\text{A1})$$

A synchronism value equal to one means that the signal is fully repeating itself with a period corresponding to the window length  $P'$ ; the values of zero indicate no similarity between the signals in the eight segments of length  $P'$ , meaning that the signal is not at all periodic with that period length.

### 3. Glimpsing

An example synchronism  $\text{synch}_{cn}(P')$  is plotted in Fig. 13(A)(3). The period glimpses  $P_{cnm}$ , where  $m$  is a glimpse index, are defined as the period values  $P'$ , corresponding to the local maxima (values larger than the neighboring values) of the synchronism, which meet the following two criteria:

- The first criterion filters out the noisy  $T$ - $F$  ( $cn$ ) bins, which have a lower degree of periodicity than the speech  $cn$  bins. If the maximum value of the synchronism exceeds the threshold  $T_1$ , then the  $T$ - $F$  bin  $cn$  is considered to originate from the speech and the glimpses are extracted from the synchronism. Otherwise, no glimpses are extracted for that  $cn$  bin.
- The second criterion is used to find the salient local maxima of the considered synchronism, which carry the information about the period of the speech signal. Every local maximum that exceeds the threshold  $T_2$  is considered to be a period glimpse  $P_{cnm}$  [blue circles in Fig. 13(A)(3)].

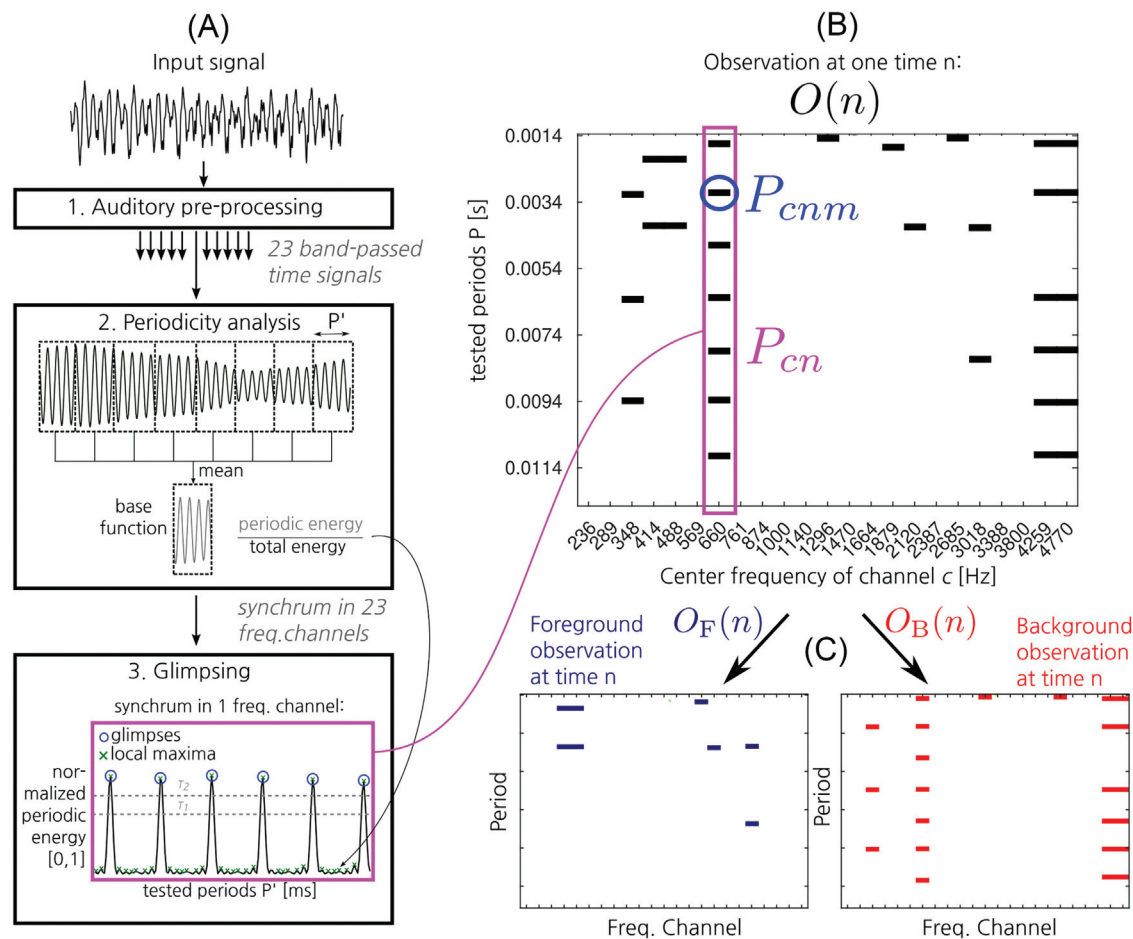


FIG. 13. The glimpsed feature extraction and segregation. (A) The processing stages of the glimpsed feature extraction and (B) the glimpsed observation in a one time instance are depicted.

(c) All of the period glimpses  $P_{cnm}$  are included in the channel set  $P_{cn}$ . The channel set may be empty in the case in which the first criterion is not met.

Josupeit and Hohmann (2017) defined the thresholds for the fine structure channels ( $T_1 = 0.9$  and  $T_2 = 0.8$ ) and envelope channels ( $T_1 = 0.5$  and  $T_2 = 0.4$ ) individually. We performed an analysis to determine the optimal threshold values in each frequency band. We consider the glimpsing thresholds to be a fixed property of the modeled auditory system, which we assume to be generalizable across different acoustic scenarios. Although the competing voice signals are used as stimuli in this study, the optimization procedure was performed using speech shaped noise (cf. Appendix B). The resulting  $T_1$  decreases with the increasing channel center frequency, and  $T_2$  is set relative to the threshold  $T_1$  such that  $T_2 = 0.9 \cdot T_1$  (see Fig. 14). The values are in good agreement with the thresholds from Josupeit and Hohmann (2017).

In summary, to extract the period glimpses, we use a procedure identical to the feature extraction procedure from Josupeit and Hohmann (2017), except for the following:

- instead of performing the periodicity analysis every 10 ms, we do it every 20 ms;

- instead of a hard distinction between the fine structure and envelope channels, we process all 23 channels in the same way;
- instead of the differentiation stage and additional gamma-tone filtering after the half-wave rectification, we use the 40 Hz high-pass filter;
- instead of analyzing the periods  $P'$  in a range between 1/1400 and 1/80 Hz, we do it in a range between 1/700 and 1/80 Hz;
- instead of fixed threshold values, we use frequency-dependent threshold values; and
- instead of extracting all of the periodicity-based features  $P_{cn}$ ,  $E_{cn}$ ,  $T_{cn}$ , and  $L_{cn}$ , we extract only the period glimpses  $P_{cn}$ .

## APPENDIX B: GLIMPSE THRESHOLD ANALYSIS

As described in Appendix A, the glimpses in each  $T$ - $F$  bin are extracted from the synchronism, an example of which is depicted in Fig. 15. If the global maximum exceeds  $T_1$ , the glimpses are extracted. This is performed using the second threshold  $T_2$ —each local maximum above  $T_2$  defines a period glimpse.

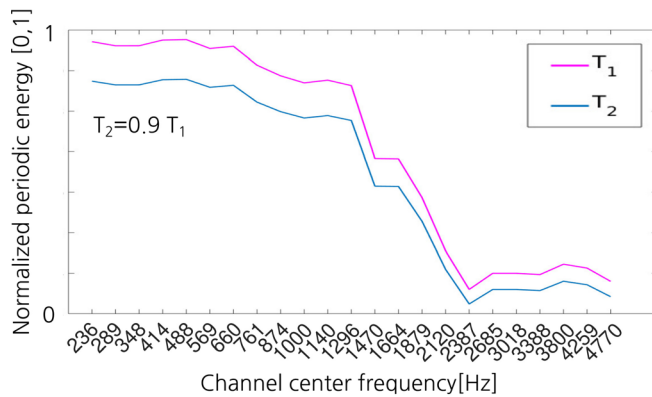


FIG. 14. (Color online) The glimpse extraction threshold values are shown.

The goal of this analysis was to determine the thresholds, which will lead to the best discrimination between the two categories: the eriodicity of a single voiced source and the periodicity of a sound mixture. We used the synthetic voice signal with a random parameter trajectory to create the data for the first category (cf. Sec. II A) and a speech shaped noise to create the data in the second category. Both of the signals were 200 s long with a sampling rate of  $f_s = 16\,000$  Hz. For both signals, we extracted and stored all of the local maxima of the synchra for each  $T$ - $F$  bin ( $cn$ ). The thresholds that lead to the best discrimination between the voice and speech shaped noise are chosen as the optimal thresholds. We consider the glimpsing thresholds to be a fixed property of the auditory system, therefore, we assume that the optimal thresholds determined in this analysis will translate well to any other acoustic scenario. The glimpses extracted from any other signal based on these thresholds are most likely to originate from a single voiced source and carry robust information related to its pitch.

### 1. Finding the optimal $T_1$

First, we search for the  $T_1$  value in each channel, which leads to the best discrimination between the voice and noise

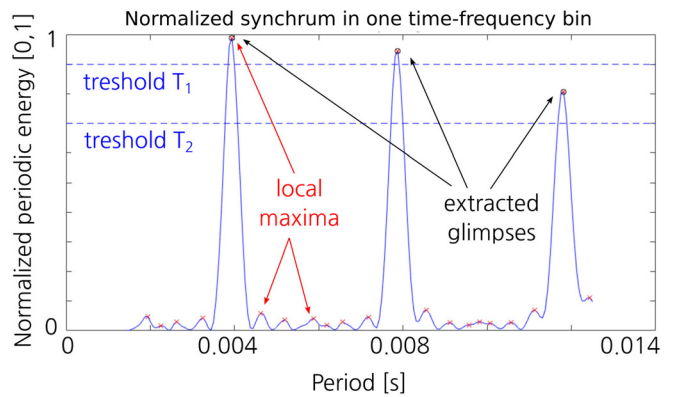


FIG. 15. (Color online) The synchrum in one  $T$ - $F$  bin is shown.

frames. We make an assumption that all of the bins  $cn$  in which the global maximum of the synchrum exceeds  $T_1$  are speech frames. In this way, we classify the frames as either noise or speech. Using the ground truth information, we perform a statistical analysis of the voiced frame detection task. For the various  $T_1$  values, we obtain the percentage of noise  $T$ - $F$  bins that were incorrectly classified as speech  $T$ - $F$  bins, the false positive rate (FPR), and the percentage of speech  $T$ - $F$  bins that were correctly classified as speech, the true positive rate (TPR). By relating those two measures, we obtain a ROC curve. For each channel, we obtain one ROC curve (see Fig. 16), as well as the area under this curve (AUC; see Fig. 17), which tells us about the performance of the detection task.

The ROC and AUCs show that the discrimination task is possible (clearly better than the chance level). The next step is to find a cut-off point on the curve, which best suits the task, and the threshold that corresponds to that cut-off point. The main constraint of the glimpse extraction task is to extract as few noise glimpses as possible. This means that to find an optimal cut-off point, we focus on limiting the FPR. On the other hand, we do not want to get rid of the glimpses originating from the speech, thus, we make sure

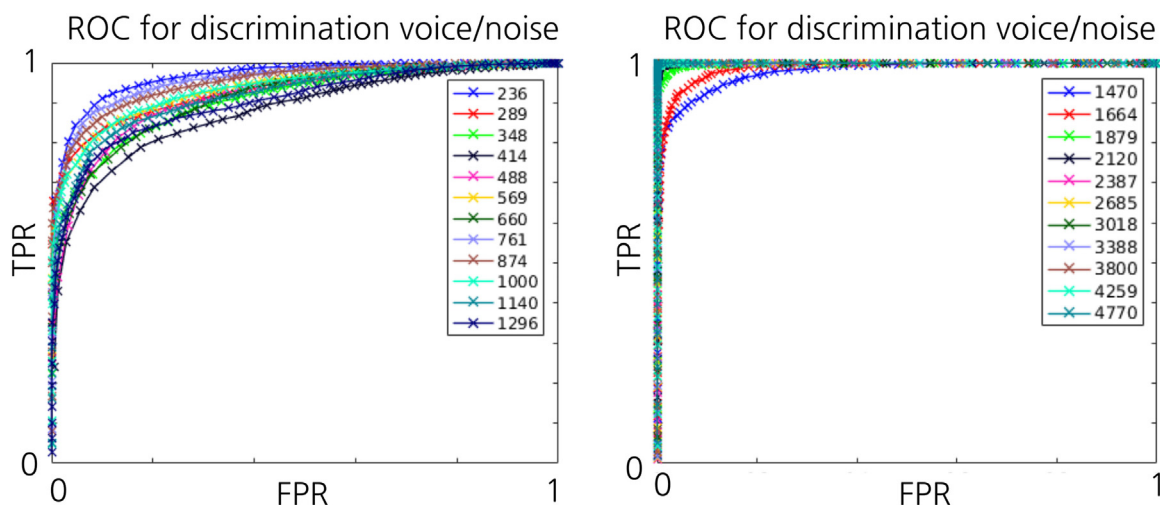


FIG. 16. (Color online) The ROC curves for the speech frames detection task. (Left) The channels 1:12 and (right) channels 13:23 are depicted.

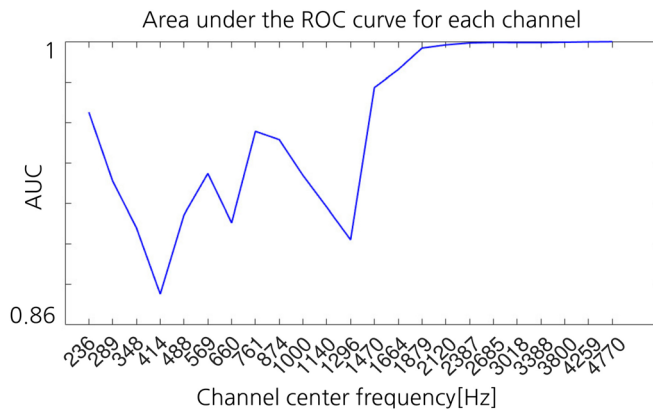


FIG. 17. (Color online) The area under the ROC curve is shown. It exceeds 0.5 in all of the channels, which means that the classifier is working properly.

that the TPR is sufficiently high. We decided to set the permissible FPR at 2%. The threshold is found by finding a value of the threshold that corresponds to the 2% FPR on the ROC curve. If the ROC curve does not contain a point that lies at exactly 2%, the threshold is interpolated between the two closest values. Figure 18 presents the thresholds chosen this way for 23 frequency channels and the corresponding TPRs, which can be achieved with the chosen thresholds.

## 2. Finding the optimal $T_2$

We want to find the second threshold value  $T_2$ , which will lead to the exclusion of the peaks in the synchrum that do not come from the correct periodicity. Here, we call them the *spurious peaks*. For that, we plot the distribution of the local maxima in all of the  $T$ - $F$  bins, which are classified as speech bins (using the first threshold  $T_1$ ). In each bin, we normalize all of the local maxima to the global maximum value in this bin. This way, we obtain the distribution, which

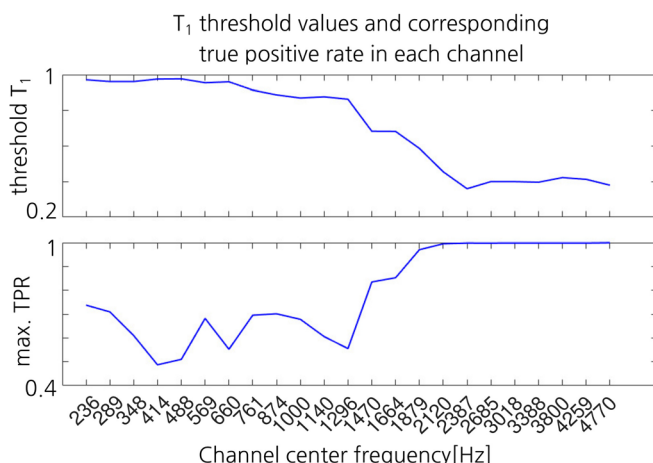


FIG. 18. (Color online) (Top) The thresholds chosen with a criterion of 2% permissible false positive rate. (Bottom) The TPR, which can be achieved using the chosen threshold, is shown. In the worst case, the TPR is about 50%.

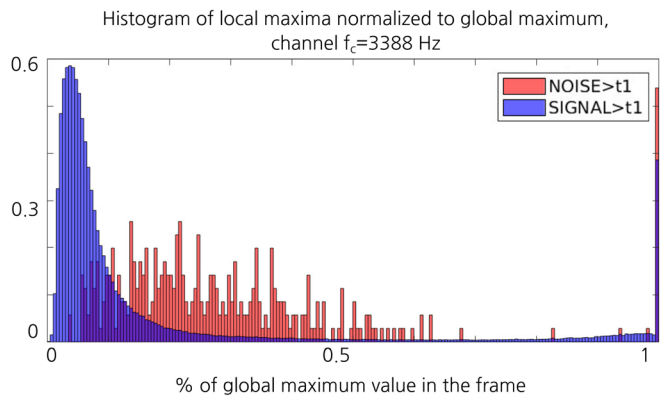


FIG. 19. (Color online) The distribution of the local maxima in the speech frames relative to the global maximum are shown. The gammatone channels  $20f_c = 3388$  Hz.

shows how much the local maxima, on average, deviate from the global maxima.

We observe a bimodal distribution with high counts at both ends of the range (see Fig. 19). We interpret this in the following way. The values close to zero are very frequent and they correspond to the spurious peaks, which do not carry information about the actual periodicity of the signal. The values that are very close to one correspond to the peaks that are almost as high as the global maximum and give us information about the real periodicity of the signal. These distributions indicate that the local maxima of the synchrum are either as high as the global maximum (those are the salient ones that are of our interest) or much lower than the global maximum (the spurious peaks). This means that it is not necessary to look for the salient local maxima within the values that are much lower than the global maximum of the frame. Thus, finally we decided to set the relative threshold  $T_2$  to be 0.9 of the threshold  $T_1$ .

## APPENDIX C: SUPPLEMENTARY INFORMATION

Apart from computational experiments described in this paper, we simulated several additional trajectory conditions, which have not yet been tested on human listeners and might be a valuable resource for further studies.

We simulated, in total, six possible combinations of the conditions, depicted and described in Fig. 20. In each of these conditions, we used 100 random trajectory pairs. Each trajectory pair was used twice as a trial. We initialized the trajectories of the competing voices in different  $F_0$  ranges (either 100–250 Hz or 250–400 Hz). The simulation was repeated 20 times. For each run, a  $d'$  value was computed. This computational experiment was repeated for three settings of the model: with formant-guided tracking, without using any oracle information, and with  $F_0$ -guided tracking.

There are two main effects reflected in Appendix C.

### 1. Influence of formant dynamics

In Woods and McDermott (2015), the authors compared the attentive tracking of voices varying and crossing in all



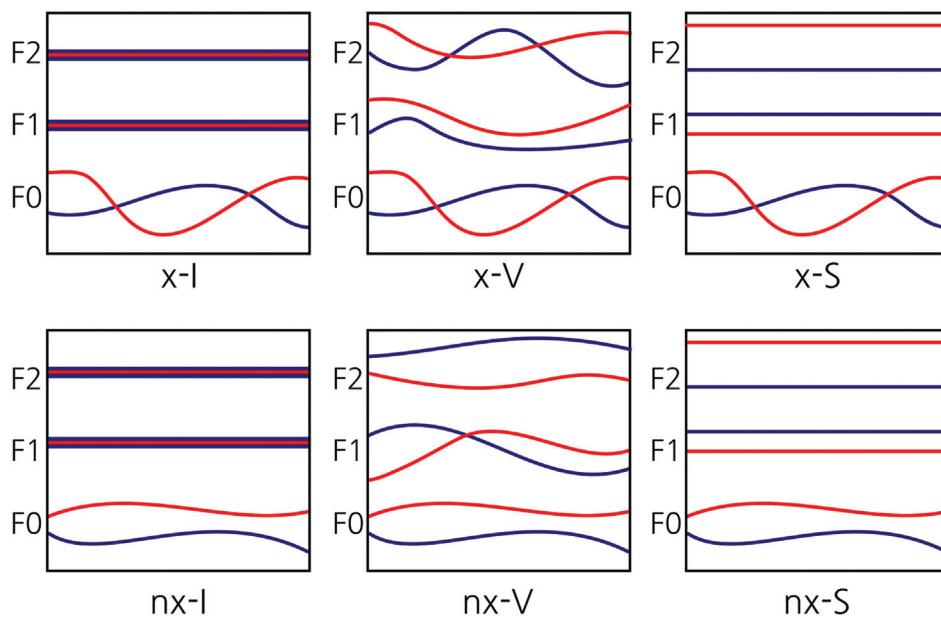


FIG. 20. The six supplementary conditions simulated by the model. There are two different types of  $F_0$  trajectories: (1) crossing  $F_0$  (x) and (2) non-crossing  $F_0$  (nx). There are three different types of formant trajectories: (1) constant over time and identical between the voices (I), (2) varying over time and different between the voices (V), and (3) constant over time but different between the voices (S).

three dimensions to attentive tracking of voices varying and crossing only in  $F_0$ . The voices with identical and constant formants could not be discriminated by the human listeners. These results suggested that multiple features allow accurate streaming where single features cannot. However, the experiment did not quantify the influence of the feature dynamics on the segregation. To examine this, we added additional conditions with the trajectories varying and crossing in  $F_0$  and with the formant frequencies, which were static over time but different between the voices. The results, depicted in Fig. 21, showed that constant separation between the

formant frequencies aids the performance in the attentive tracking. It is easier to distinguish the voices if their formants do not vary over time, which proves that, at least for the model, there is a significant influence of the formant dynamics on the attentive tracking.

## 2. Influence of trajectory crossing

In Woods and McDermott (2015), the trajectory pairs crossed at least once in each feature dimension. This way the authors ensured that the identification of the voices was

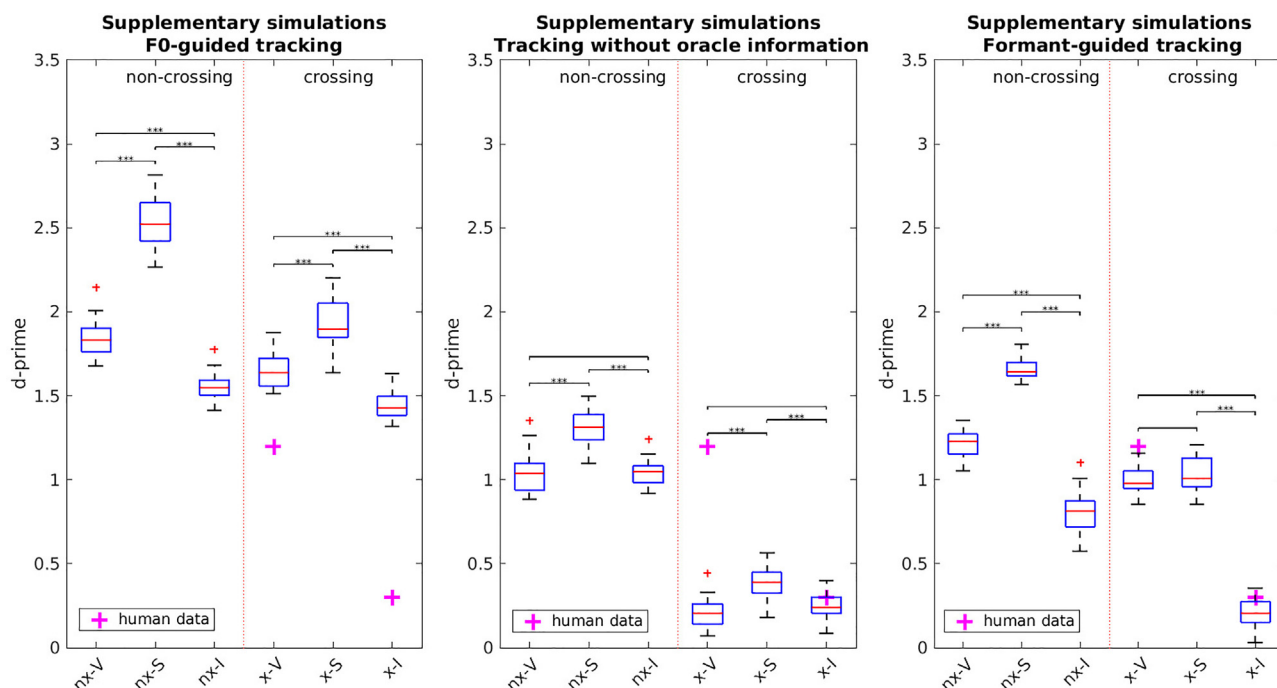


FIG. 21. (Color online) The additional experiments, where the abbreviations for the condition names are x, crossing  $F_0$ ; nx, non-crossing  $F_0$ ; the I, the formants are constant and identical between the voices; V, the formants are varying and different between the voices; S, the formants are static but different between the voices.

not performed on the basis of one single feature. However, it was not tested if the voices with the varying features trajectories, which do not cross in time, are indeed easier to discriminate than when the trajectories cross. To examine the model's response to this task, we performed the additional simulations for the stimuli with non-crossing trajectories. The results, depicted in Fig. 21, showed that the crossings in the  $F0$  trajectories have a significant influence on the attentive tracking performance. Even for the model without any oracle information, the discrimination performance for the non-crossing trajectories is very good. This proves that, for the model, it is easier to track the noncrossing trajectories, and the  $F0$  crossings are the main reason for the model to fail in the tracking task.

Aitchison, L., and Lengyel, M. (2017). "With or without you: Predictive coding and Bayesian inference in the brain," *Curr. Opin. Neurobiol.* **46**, 219–227.

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). "A tutorial on particle filters for online nonlinear/non-gaussian Bayesian tracking," *IEEE Trans. Signal Process.* **50**(2), 174–188.

Assmann, P. F., and Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," *J. Acoust. Soc. Am.* **88**(2), 680–697.

Bernstein, J. G., and Oxenham, A. J. (2003). "Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number?," *J. Acoust. Soc. Am.* **113**(6), 3323–3334.

Best, V., Mason, C. R., Swaminathan, J., Kidd, G., Jakien, K. M., Kampel, S. D., Gallun, F. J., Buchholz, J. M., and Glyde, H. (2016). "On the contribution of target audibility to performance in spatialized speech mixtures," in *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing* (Springer, Cham), pp. 83–91.

Best, V., Mason, C. R., Swaminathan, J., Roverud, E., and Kidd, G., Jr. (2017). "Use of a glimpsing model to understand the performance of listeners with and without hearing loss in spatialized speech mixtures," *J. Acoust. Soc. Am.* **141**(1), 81–91.

Best, V., Ozmeral, E. J., Kopčo, N., and Shinn-Cunningham, B. G. (2008). "Object continuity enhances selective auditory attention," *Proc. Natl. Acad. Sci. U.S.A.* **105**(35), 13174–13178.

Boer, E. D. (1956). "Pitch of inharmonic signals," *Nature* **178**(4532), 535–536.

Bregman, A. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound* (Bradford/MIT Press, Cambridge, MA).

Bressler, S., Masud, S., Bharadwaj, H., and Shinn-Cunningham, B. (2014). "Bottom-up influences of voice continuity in focusing selective auditory attention," *Psychol. Res.* **78**(3), 349–360.

Cariani, P. A., and Delgutte, B. (1996). "Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch," *J. Neurophysiol.* **76**(3), 1717–1734.

Carlyon, R. P. (2004). "How the brain separates sounds," *Trends Cognit. Sci.* **8**(10), 465–471.

Carlyon, R. P., Cusack, R., Foxton, J. M., and Robertson, I. H. (2001). "Effects of attention and unilateral neglect on auditory stream segregation," *J. Exp. Psychol.: Hum. Percept. Perform.* **27**(1), 115–127.

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). "Probabilistic models of cognition: Conceptual foundations," *Trends Cogn. Sci.* **10**(7), 287–291.

Chen, Z. (2003). "Bayesian filtering: From Kalman filters to particle filters, and beyond," *Statistics* **182**(1), 1–69.

Cherry, E. C. (1953). "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.* **25**(5), 975–979.

Cohen-Lhyver, B., Argentieri, S., and Gas, B. (2018). "The head turning modulation system: An active multimodal paradigm for intrinsically motivated exploration of unknown environments," *Front. Neurobot.* **12**, 60.

Cooke, M. (2006). "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.* **119**(3), 1562–1573.

Darwin, C. (2008). "Listening to speech in the presence of other sounds," *Philosoph. Trans. R. Soc. B: Biol. Sci.* **363**(1493), 1011–1021.

Dietz, M., Ewert, S. D., and Hohmann, V. (2011). "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Commun.* **53**(5), 592–605.

Dietz, M., Ewert, S. D., Hohmann, V., and Kollmeier, B. (2008). "Coding of temporally fluctuating interaural timing disparities in a binaural processing model based on phase differences," *Brain Res.* **1220**, 234–245.

Di Fu, C. W., Yang, G., Kerzel, M., Nan, W., Barros, P., Wu, H., Liu, X., and Wermter, S. (2020). "What can computational models learn from human selective attention? a review from an audiovisual unimodal and crossmodal perspective," *Front. Integr. Neurosci.* **14**, 10.

Elhilali, M. (2013). "Bayesian inference in auditory scenes," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, New York, pp. 2792–2795.

Elhilali, M., and Shamma, S. A. (2008). "A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation," *J. Acoust. Soc. Am.* **124**(6), 3751–3771.

Elhilali, M., Xiang, J., Shamma, S. A., and Simon, J. Z. (2009). "Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene," *PLoS Biol.* **7**(6), e1000129.

Ellis, D. P. (1999). "Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures," *Speech Commun.* **27**(3–4), 281–298.

Friston, K., Adams, R., Perrinet, L., and Breakspear, M. (2012). "Perceptions as hypotheses: Saccades as experiments," *Front. Psychol.* **3**, 151.

Garrido, M. I., Kilner, J. M., Stephan, K. E., and Friston, K. J. (2009). "The mismatch negativity: A review of underlying mechanisms," *Clin. Neurophysiol.* **120**(3), 453–463.

Gregory, R. L. (1980). "Perceptions as hypotheses," *Philosoph. Trans. R. Soc. London. B, Biol. Sci.* **290**(1038), 181–197.

Gregory, R. L. (1997). "Knowledge in perception and illusion," *Philosoph. Trans. R. Soc. London. Ser. B: Biol. Sci.* **352**(1358), 1121–1127.

Haftner, E. R., Sarampalis, A., and Loui, P. (2008). "Auditory attention and filters," in *Auditory Perception of Sound Sources* (Springer, Cham), pp. 115–142.

Heilbron, M., and Chait, M. (2018). "Great expectations: Is there evidence for predictive coding in auditory cortex?," *Neuroscience* **389**, 54–73.

Helmholtz, H. von. (1897). "The facts in perception," in *Helmholtz on perception: Its physiology and development*, edited by R. M. Warren and R. P. Warren (Wiley, New York).

Hohmann, V. (2006). "Method for extracting periodic signal components, and apparatus for this purpose," U.S. patent application 11/223,125 (April 6, 2006).

Josupeit, A., and Hohmann, V. (2017). "Modeling speech localization, talker identification, and word recognition in a multi-talker setting," *J. Acoust. Soc. Am.* **142**(1), 35–54.

Josupeit, A., Kopčo, N., and Hohmann, V. (2016). "Modeling of speech localization in a multi-talker mixture using periodicity and energy-based auditory features," *J. Acoust. Soc. Am.* **139**(5), 2911–2923.

Josupeit, A., Schoenmaker, E., van de Par, S., and Hohmann, V. (2020). "Sparse periodicity-based auditory features explain human performance in a spatial multitalker auditory scene analysis task," *Eur. J. Neurosci.* **51**(5), 1353–1363.

Kaya, E. M., and Elhilali, M. (2017). "Modelling auditory attention," *Philosoph. Trans. R. Soc. B: Biol. Sci.* **372**(1714), 20160101.

Koch, I., Lawo, V., Fels, J., and Vorländer, M. (2011). "Switching in the cocktail party: Exploring intentional control of auditory selective attention," *J. Exp. Psychol.: Hum. Percept. Perform.* **37**(4), 1140–1147.

Kondo, K., Taira, K., and Kobayashi, Y. (2018). "Binaural speech intelligibility estimation using deep neural networks," in *Interspeech*, pp. 1858–1862.

Li, T., Sun, S., Sattar, T. P., and Corchado, J. M. (2014). "Fight sample degeneracy and impoverishment in particle filters: A review of intelligent approaches," *Expert Syst. Appl.* **41**(8), 3944–3954.

Luberadzka, J., Kayser, H., and Hohmann, V. (2020). "Estimating fundamental frequency and formants based on periodicity glimpses: A deep learning approach," in *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, New York, pp. 1–6.

Madsen, S. M., Marschall, M., Dau, T., and Oxenham, A. J. (2019). "Speech perception is similar for musicians and non-musicians across a wide range of conditions," *Sci. Rep.* **9**(1), 1–10.

- McDermott, J. H. (2009). "The cocktail party problem," *Curr. Biol.* **19**(22), R1024–R1027.
- Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (2014). "Mechanisms of noise robust representation of speech in primary auditory cortex," *Proc. Natl. Acad. Sci. U.S.A.* **111**(18), 6792–6797.
- Näätänen, R., Gaillard, A. W., and Mäntysalo, S. (1978). "Early selective-attention effect on evoked potential reinterpreted," *Acta Psychol.* **42**(4), 313–329.
- Nix, J., and Hohmann, V. (2007). "Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering," *IEEE Trans. Audio, Speech, Lang. Process.* **15**(3), 995–1008.
- Nix, J., Kleinschmidt, M., and Hohmann, V. (2003). "Computational auditory scene analysis by using statistics of high-dimensional speech dynamics and sound source direction," in *Eighth European Conference on Speech Communication and Technology*.
- Popham, S., Boebinger, D., Ellis, D. P., Kawahara, H., and McDermott, J. H. (2018). "Inharmonic speech reveals the role of harmonicity in the cocktail party problem," *Nat. Commun.* **9**(1), 2122.
- Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). "Probabilistic brains: Knowns and unknowns," *Nat. Neurosci.* **16**(9), 1170–1178.
- Purwins, H., Sturm, B., Li, B., Nam, J., and Alwan, A. (2019). "Introduction to the issue on data science: Machine learning for audio signal processing," *IEEE J. Sel. Top. Signal Process.* **13**(2), 203–205.
- Saddler, M. R., Gonzalez, R., and McDermott, J. H. (2021). "Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception," *Nat. Commun.* **12**(1), 1–25.
- Sanborn, A. N., and Chater, N. (2016). "Bayesian brains without probabilities," *Trends Cognit. Sci.* **20**(12), 883–893.
- Schoenmaker, E., and van de Par, S. (2016). "Intelligibility for binaural speech with discarded low-SNR speech components," in *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing* (Springer, Cham), pp. 73–81.
- Schouten, J. F., Ritsma, R., and Cardozo, B. L. (1962). "Pitch of the residue," *J. Acoust. Soc. Am.* **34**(9B), 1418–1424.
- Schroeder, M. R. (1968). "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acoust. Soc. Am.* **43**(4), 829–834.
- Schröger, E., Marzecová, A., and SanMiguel, I. (2015). "Attention and prediction in human audition: A lesson from cognitive psychophysiology," *Eur. J. Neurosci.* **41**(5), 641–664.
- Shamma, S., and Dutta, K. (2019). "Spectro-temporal templates unify the pitch percepts of resolved and unresolved harmonics," *J. Acoust. Soc. Am.* **145**(2), 615–629.
- Shamma, S. A., and Micheyl, C. (2010). "Behind the scenes of auditory perception," *Curr. Opin. Neurobiol.* **20**(3), 361–366.
- Shi, L., and Griffiths, T. L. (2009). "Neural implementation of hierarchical bayesian inference by importance sampling," in *Advances in Neural Information Processing Systems*, pp. 1669–1677.
- Shinn-Cunningham, B. G. (2008). "Object-based auditory and visual attention," *Trends Cognit. Sci.* **12**(5), 182–186.
- Siedenburg, K., Goldmann, K., and Van De Par, S. (2021). "Tracking musical voices in Bach's The Art of the Fugue: Timbral heterogeneity differentially affects younger normal-hearing listeners and older hearing-aid users," *Front. Psychol.* **12**, 608684.
- Snyder, J. S., Gregg, M. K., Weintraub, D. M., and Alain, C. (2012). "Attention, awareness, and the perception of auditory scenes," *Front. Psychol.* **3**, 15.
- Spille, C., Ewert, S. D., Kollmeier, B., and Meyer, B. T. (2018). "Predicting speech intelligibility with deep neural networks," *Comput. Speech Lang.* **48**, 51–66.
- Spille, C., Meyer, B., Dietz, M., and Hohmann, V. (2013). "Binaural scene analysis with multidimensional statistical filters," in *The Technology of Binaural Listening* (Springer, Cham), pp. 145–170.
- Szabó, B. T., Denham, S. L., and Winkler, I. (2016). "Computational models of auditory scene analysis: A review," *Front. Neurosci.* **10**, 524.
- Terhardt, E. (1989). "On the role of ambiguity of perceived pitch in music," in *Proceedings of the 13th ICA Belgrade*, pp. 35–38.
- Woods, K. J., and McDermott, J. H. (2015). "Attentive tracking of sound sources," *Curr. Biol.* **25**(17), 2238–2246.
- Woods, K. J., and McDermott, J. H. (2018). "Schema learning for the cocktail party problem," *Proc. Natl. Acad. Sci. U.S.A.* **115**(14), E3313–E3322.
- Wrigley, S. N., and Brown, G. J. (2004). "A computational model of auditory selective attention," *IEEE Trans. Neural Networks* **15**(5), 1151–1163.
- Xiang, J., Simon, J., and Elhilali, M. (2010). "Competing streams at the cocktail party: Exploring the mechanisms of attention and temporal integration," *J. Neurosci.* **30**(36), 12084–12093.