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

Joanna Luberadzka,¹ Hendrik Kayser,¹ and Volker Hohmann¹

Auditory Signal Processing and Cluster of Excellence Hearing4all,
Department of Medical Physics and Acoustics, University of Oldenburg, Germany,
joanna.luberadzka@uni-oldenburg.de

(Dated: 11 November 2021)
Humans are able to follow a speaker even in challenging acoustic conditions. The perceptual mechanisms underlying this ability remain unclear. We present a computational model of attentive voice tracking, consisting of four computational blocks: 1) sparse periodicity-based auditory feature (sPAF) extraction, 2) foreground-background segregation, 3) state estimation, and 4) top-down knowledge. The model connects theories about auditory glimpses, foreground-background segregation and Bayesian inference. It is implemented with sPAF, sequential Monte Carlo sampling and probabilistic voice models. We evaluate the model by comparing it with the human data obtained in the study of Woods and McDermott (2015), which measured the ability to track of one of two competing voices with time-varying parameters (fundamental frequency (F0) and formants (F1, F2)). We test three model versions, which differ in the type of information used for the segregation: version (a) uses oracle F0, version (b) uses estimated F0 and version (c) uses spectral shape derived from estimated F0 and oracle F1 and F2. Version (a) simulates optimal human performance in conditions with the largest separation between the voices, version (b) simulates conditions where the separation in not sufficient to follow the voices, and version (c) is closest to human performance for moderate voice separation.
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 listeners 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 first two formants — varied over time. There were no constant, distinctive features between the voices that could facilitate stream formation (for example, direction of arrival or timbre) and the multidimensional parameter trajectories crossed over time. They concluded that listeners can successfully distinguish the attended voice from the background voice if they focus their attention on one of the voices and if the parameter trajectories maintain a sufficient separation in the feature space.

Exactly how the mixture of acoustic signals is decomposed into attended foreground and residual background remains unclear (Carlyon, 2004). Over the last few years, several re-
searchers have addressed this and other questions related to 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 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 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 arrows in the upper left and bottom right corner of Figure 1).

The scale starts with the fully stimulus-driven processing, responsible for extracting the basic auditory features (like harmonic structure, time onset and spatial direction (Darwin, 2007)). 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, energetic and informational masking makes it difficult to extract reliable information about individual sources.

In such a scenario, perception can be characterized as glimpsing (Darwin, 2007): Reliable auditory features are provided primarily by the time-frequency units that 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 a time, rather than the superposed features from several sound objects
(Best et al., 2016, 2017; Cooke, 2006; Schoenmaker and van de Par, 2016). The studies
mentioned above usually prove the relevance of the time-frequency (T-F) bins dominated
by a single source by comparing the speech intelligibility for signals resynthesized from
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, 2018)
developed an approach capable of blindly extracting speech glimpses from the mixture of
signals, called the *sparse periodicity-based auditory features* (sPAF). Speech is a highly pe-
riodic signal and the harmonic structure plays a crucial role in voiced speech segregation
(Popham et al., 2018). The core idea of 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. (2018) was used to model several aspects of auditory perception in a multi-
talker scenario. Results suggested that, despite their sparsity, sPAF contain the bulk of the
information needed to decode a complex auditory scene. The *sPAF extraction* stage (see
Fig. 1, B.1. and Section 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 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 that 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 time-frequency 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 the remaining talkers (Darwin, 2007).

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 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 Section II B 2) of the current model, the sPAF at a given time instance are decomposed into fore-
ground 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 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 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 Section II B 4) of the model. The implementation details of these models are presented in Section II C.

The top-down and bottom-up processes are not mutually independent: they are engaged in a machinery of interactions generating perception (Bressler et al., 2014). Many recent studies propose Bayesian inference as an elegant computational illustration of this synergy. Bayesian estimation is granted as a plausible model of the optimal inference both in a general view on cognition (Aitchison and Lengyel, 2017; Chater et al., 2006; Helmholtz and Von, 1878; Pouget et al., 2013) as well as in the context of auditory perception (Elhilali, 2013;
In a nutshell, Bayesian models of 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, sequential Bayesian estimation — finding the posterior state distribution given a series of observations — does not have a single universal solution. There is a variety of methods, which 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 idea has been revisited in recent studies comparing 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 Section II.B.3) integrates the components of our model into a sequential Bayesian inference framework. We use the Monte Carlo sam-
pling approach, i.e., particle filtering (Arulampalam et al., 2002) to simultaneously track the state of the foreground and the background stream. 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 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 $F_0$, $F_1$, and $F_2$, whose parameter trajectories cross in time. This type of stimulus is new among sPAF-based modeling approaches, which previously worked with multitalker speech sets (Josupeit and Hohmann, 2017; Josupeit et al., 2018).

Instead of the template-matching approach, which evaluated sPAF on the word scale, the proposed model performs sequential processing. It tracks fundamental frequency of competing voices using instantaneous sPAF and probabilistic $F_0$ models (Section II C). The sPAF and $F_0$ 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 sparse periodicity-based auditory features (Josupeit and Hohmann, 2017; Josupeit et al., 2018) 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.

3. A comparison of a single-dimensional version of our model (F0 tracking) with human results in 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 Figure 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. II B 1), where the salient auditory features are extracted from the input mixture, Foreground-background segregation (Sec. II B 2), where the sPAF are segregated into foreground and background segments, State estimation (Sec. II B 3), where the particle filter estimates the voice state based on the segregated sPAFs, and Top-down knowledge (Sec. II B 4), where the probability distributions required for the state estimation are briefly reviewed. Section II C presents implementation details of the probability distributions related to voice fundamental frequency, which are used in this study: F0 transition and F0 observation model.
Figure 1. Modeling framework. A. Signal generation: Two synthetic voiced signals are created based on randomized state trajectories, which comprise the time series of the voice parameters. The mixture of the foreground (F) and the background (B) signal is the input to the model. B. Computational framework: Attentive tracking of a foreground sound source in an auditory scene is implemented with several computational sub-tasks, placed on a top-down-bottom-up scale (see arrows in the top-left and bottom-right corner). 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 the background voice.
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: 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 $F_0$ and first two formants $F_1$ and $F_2$ vary over time, as in the study by Woods and McDermott (2015). One voice is considered the foreground and the other voice is considered the background. In each time instance, the signals are defined by the 3-dimensional state vectors $\vec{s}_F$ and $\vec{s}_B$ containing the parameter values:

$$\vec{s}_F(n) = \begin{bmatrix} F_0F(n) \\ F_1F(n) \\ F_2F(n) \end{bmatrix}, \quad \vec{s}_B(n) = \begin{bmatrix} F_0B(n) \\ F_1B(n) \\ F_2B(n) \end{bmatrix},$$

(1)

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

$$\mathcal{T}_F = \{\vec{s}_F(n)|n = 0, ..., N\}$$

$$\mathcal{T}_B = \{\vec{s}_B(n)|n = 0, ..., N\},$$

(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 input to a formant synthesizer, which generates synthetic foreground and background acoustic signals. The signals are summed and the resulting mixture is the input to the model. Fig. 1, A. illustrates the signal generation procedure.
B. Computational framework

1. Sparse periodicity-based auditory features (sPAF) extraction

The signal containing mixture of the foreground and the background is forwarded to the feature extraction stage (See Fig. 1, B.1.). The sparse periodicity-based auditory feature (sPAF) extraction stage is based on a series of studies by Josupeit et al. (2016), Josupeit and Hohmann (2017) and Josupeit et al. (2018). 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 the Appendix (Sec.VII A).

Figure 2. sPAF represent robust tonal components of the auditory scene. In contrast to a one-dimensional time representation (a) or two-dimensional time-frequency representation (b), the sPAF analyze the sound also in a 3rd dimension: period (c). Moreover, sPAF allow only the salient, robust components originating from a single sound source. Noisy T-F bins, as well as the bins containing superposition of many sound sources are eliminated by the glimpsing process (d).

sPAF extraction consists of thee main steps (Fig. 1, B.):
1. *Auditory pre-processing*, which provides the auditory-inspired time-frequency 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 time-frequency-period representation (Fig.1, B.1.2.).

3. *Glimpsing*, which removes all the non-salient information from the time-frequency-period representation and extracts the salient period values called here *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.):

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

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

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

$P_{cn}$ consists of the salient period values — *period glimpses* — denoted as $P_{cnm}$. Indices $c, n, m$ 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 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 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 3 alternative methods, in order to evaluate various aspects of the model. Figure 3 shows the pseudocode for each of these methods. The purpose of each method is reviewed in more detail in Section 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 the background voice, given the segregated sPAF. For tracking, we use particle filtering, combined with probability distributions from the top-down knowledge stage (Fig. 1, B.4.). 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 to iteratively compute the state estimate based on these samples and weights.
Figure 3. Foreground-background segregation methods. Method B.2.1. uses the ground truth F0 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 F0 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 F0, the only difference with B.2.1. is that the segregation is done based on the estimated (instead of ground truth) F0s from the preceding time step. Method B.2.3. is substantially different from the first two methods: In addition to the estimated F0, it uses the ground truth information about the formants to segregate voices. For each voice, a channel-dependent weight is computed. This reflects the energy distribution over frequency channels for a given combination of $\hat{F}0$, $F1$, and $F2$. The F0 estimate is used in the encoding of weights, but is not explicitly used for the segregation as in the first two methods.

Although particle filtering allows for multi-dimensional tracking of events with arbitrary probability distributions, the tracking cannot be performed until these distributions are known (See also Sec. II B 4). For the 3-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: Transitions of parameters can be modeled with a linear motion model. However, the observation statistics distribution, describing the relationship between the 3-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 output of the computational model in Fig. 1). Thus, when refering to the estimation process we replace the vector notation \( \vec{s} \) with a one-dimensional state \( s \):

\[
\hat{s}_F(n) = \hat{F}_0F(n)
\]

\[
\hat{s}_B(n) = \hat{F}_0B(n).
\] (5)

The symbol \( \hat{\cdot} \) is also used to differentiate between the ground truth and the estimated quantities. Tracking yields one-dimensional estimated state trajectories:

\[
\hat{T}_F = \{\hat{s}_F(n) | n = 0, ..., N\}
\]

\[
\hat{T}_B = \{\hat{s}_B(n) | n = 0, ..., N\}.
\] (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., 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 system onset, a particle filter is initialized with the available prior knowledge (Fig. 1, B.3.1.). The initial expectation is created. Hypotheses are sampled from the attention prior probability distribution: \( p(s_F(0)) \). All 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. New particles are predicted, given the previous particles. This is done by drawing 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 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 particles.

4. 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 hypotheses – 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 regions of high importance. The particles with small weights are eliminated and 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 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 carried out only at time steps when the foreground observation \( O_F(n) \) is not empty and when the particle diversity, measured in terms of effective sample size is lower than a predetermined threshold.

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

4. Top-down knowledge

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. 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. Foreground tracking, in contrast to the background tracking, is initialized in an informed way. This is illustrated in Figure 4.
We simulate how 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 done by sampling from discrete distributions defined in the range between 100 - 400 Hz with the shapes as shown in this figure, with weights normalized to 1. Shape of the distribution was derived using a normal distributions, with 10 Hz standard deviation for the foreground and 50 Hz standard deviation for the background.

2. 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 particle development: predicting new hypotheses based on the previous hypotheses set. In this study, we use the F0 transition model described in detail in Sec. II C 1.

3. 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 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 the background have the same properties. Hence, apart from the attention prior, the models for foreground and background are the same. For a more detailed mathematical description, please refer to the Section II C.

### C. Single voice F0 models

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

#### 1. F0 transition model

The F0 transition model describes temporal evolution of F0, which is naturally limited due to 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 $\Delta \hat{F0}(n) =$
Figure 5. State transition probability model predicting the next state value. \( F_0 \): hypothetical fundamental frequency, \( \hat{F}_0(n - 1) \): \( F_0 \) estimate in the last time step, \( \hat{F}_0(n - 2) \): \( F_0 \) estimate in the second-to-last time step, \( \Delta \hat{F}_0(n) \): difference between the last two \( F_0 \) estimates.

\( \hat{F}_0(n - 2) - \hat{F}_0(n - 1) \) between two previous estimates is computed, the next value according to that trend \( F_0 + \Delta \hat{F}_0(n) \) is predicted, and finally, the gaussian noise is added to this value:

\[
p(F_0(n)|F_0(n - 1)) = \mathcal{N}(F_0(n - 1) + \Delta \hat{F}_0(n), \sigma_{\text{trans}}),
\]

where \( \sigma_{\text{trans}} = 1 \) Hz is the standard deviation of the gaussian distribution centered at \( F_0(n - 1) + \Delta \hat{F}_0(n) \). In addition, we make sure that the difference between two previous estimates \( \Delta \hat{F}_0(n) \) does not exceed the largest allowed step of \( 10 \times \sigma_{\text{trans}} \) and that the extrapolated value \( F_0 + \Delta \hat{F}_0(n) \) does not exceed a typical pitch value range \([100, 400]\). We
can describe this procedure with the following pseudocode:

\[
\text{WHILE } |\Delta \hat{F}_0(n)| > 10 \cdot \sigma_{\text{trans}} \text{ or } F_0 + \Delta \hat{F}_0(n) \notin [100, 400] \\
\Delta \hat{F}_0(n) := 0.8 \cdot \Delta \hat{F}_0(n). \\
\text{END WHILE}
\]

The process of predicting a new \(F_0\) value based on the hypothetical previous \(F_0\) is illustrated in Fig. 5.

2. \textit{F0 observation model}

\(F_0\) observation model describes the relationship between 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 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 the Appendix VII 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:

$$p(P_{cn}|F0) = \prod_{m} p(P_{cnm}|F0). \quad (9)$$

We assume the mutual independence of the period glimpses within a channel set $P_{cn}$. The product ensures that high likelihood for a given $F0$ only occurs when there is a good match for all 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 frequency channels:

$$p(O_F(n)|F0) = \sum_{c} p(P_{cn}|F0). \quad (10)$$

Below we explain the motivation and implementation of the function $p(P_{cnm}|F0)$, which evaluates the likelihood of a single period glimpse $P_{cnm}$. Section II C 2a discusses the distribution of period glimpses and its relation to the period histogram by Schroeder (1968). Section II C 2b introduces the notion of a relative period glimpse and comments on the distribution of relative period glimpses. Section II C 2c presents our approach for modeling this data as a mixture of circular von-Mises distributions.

a. Distribution of period glimpses. Period glimpses $P_{cnm}$ do not always represent the period of $F0$ itself. An example of sPAF extracted from a voice with $F0 = 124$ Hz is shown in Fig. 6.A.. Bandpass filtering influences the periodicity of the signal at the output of each frequency channel (see Fig. 6B). In low-frequency channels, there is typically only one resolved harmonic of $F0$ per band; the periodicity in these channels is related to the
dominant harmonic. 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 (similar nature of the periodicity at the output of cochlear-inspired filterbank was also described in Shamma and Dutta (2019)). Furthermore, a signal with a period $P$, is also periodic at $2 \cdot P$, $3 \cdot P$, $4 \cdot P$, etc., therefore multiples of the period are also detected as glimpses. Altogether, a single period glimpse $P_{\text{cnm}}$ can assume any value equal to $\frac{i}{j \cdot F_0}$, where $j = 1, 2, \ldots$ is a harmonic number and $i = 1, 2, \ldots$ 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 $F_0$ ($\frac{1}{2} F_0$, $\frac{1}{3} F_0$, $\frac{1}{4} F_0$, etc.) or their multiples ($\frac{2}{2} F_0$, $\frac{3}{2} F_0$, $\frac{4}{2} F_0$, etc.). A schematic example of such a period histogram is shown in Figure 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 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, 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 $1/F_0$. 

25
Figure 6. Distribution of period glimpses. A. Period glimpses extracted from a voice signal with $F0 = 124$ Hz (accumulated over 1 second). Observed values are related to the underlying $F0$ by the formula $\frac{i}{j\cdot F0}$, 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 band pass filtering. C. Schroeder’s period histogram (as in the original plot by Schroeder, here the period histogram for a signal consisting of the fundamental and the first three harmonics is plotted.). The same period values can originate from different harmonics of $F0$. This makes some period values more likely to occur than others.
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 $P_{cnm}$ extracted from a single synthesized voice with varying $F_0$, $F_1$, and $F_2$. Plotting the histogram of absolute period glimpses in seconds is difficult to interpret: There is an ambiguity related to the fact that $P_{cnm}$ can assume any value equal to $\frac{i}{F_0}$: Even though 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 relative period glimpses, computed as follows:

$$R_{cnm}(F_0) = \text{rem}(\frac{P_{cnm}}{P_0}) = \text{rem}(P_{cnm} \cdot F_0),$$  \hspace{1cm} (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 $R_{cnm}(F_0)$ range from 0 to 1. 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 $R_{cnm}(F_0)$ lie around 0 and 1. Next highest peak is at 0.5, followed by smaller peaks at 0.33 and 0.66, and further 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: $\frac{i}{F_0}$, contributing to the peaks at 0 and 1. Second most likely value is a multiple of the period of the second harmonic of $F_0$: $\frac{i}{2F_0}$, contributing to the
peaks at 0, 0.5 and 1. Next peak can be found for the multiples of the third harmonic: $\frac{i}{3}F_0$, 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 $F_0$ values, resolves the ambiguity about the period multiple $i$, and shows the probability of observing a glimpse from harmonic number $j$. The form of the distribution was empirically found to be both independent of $F_0$ and of 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 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 \cite{Bernstein and Oxenham, 2003}. Each element of the sum represents a different harmonic $j$ of $F_0$:

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

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

Figure 7.C. shows histograms of the periods sampled from each of 11 von Mises distribution contributing to the mixture. Figure 7.B. shows samples from the mixture distribution,
Figure 7. Distribution of relative period glimpses. A. Histogram of all relative period glimpses (relative to the underlying F0) in a 20 second synthetic voice signal with varying F0, F1, and F2. B. Histogram of 1000 relative period glimpses sampled from the mixture of von-Mises distribution. C. 11 histograms of 1000 relative period glimpses each, sampled from the individual von-Mises distributions contributing to the mixture.

in which every harmonic has a different prior (Eq. 12). The sampled values correspond well to the histogram from Figure 7.A.
Figure 8. Procedure for evaluating 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).

III. MODEL EVALUATION

To evaluate the model, we simulated two conditions from the psychoacoustic study by Woods and McDermott (2015) and compared model performance with 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 human listeners. Results can be found in the supplementary material to this manuscript.
Figure 9. Simulation of the psychoacoustic study with the attentive tracking model. Upper panel: Schematic view of one trial of the attentive tracking experiment performed with human listeners. Lower panel: Method for simulating this experiment using the attentive tracking model. The size and color saturation of the dots correspond to particle weights.

A. Psychoacoustic study design

In the study by Woods and McDermott (2015) participants were given the following task:

After hearing a 500 ms cue signal indicating which voice should be attended, the 2s 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, upper panel). Performance was measured in terms of sensitivity index ($d'$).
We simulated this experimental procedure using the attentive tracking model. Each trial consisted of a pair of competing voices synthetized based on state trajectories $\mathcal{T}_F$ and $\mathcal{T}_B$ (see Eq. (1) and (2)). The trajectory sampling rate was $F_S = 50\, \text{Hz}$, and $N$ was set to 101, which corresponds to a signal length of 2s.

The model’s task was to track the fundamental frequency of the cued voice in an online manner, using sparse periodicity-based auditory features (sPAF). Instead of presenting the cue signal explicitly to the model, we simulated the additional information that the cue delivers to the listeners. Tracking was initialized closely around the ground truth value $F_0(0)$ for the foreground voice and everywhere besides that value for the background voice (See Sec. II B 4 for details). Tracking yielded the estimated 1-dimensional state trajectories $\hat{\mathcal{T}}_F$ and $\hat{\mathcal{T}}_B$ (see Eq. (6)).

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 $F_0$ of the foreground (corresponding to the last 500 ms of the signal) to the last 25 values of the ground truth $F_0$ of the foreground voice and to the last 25 values of the ground truth $F_0$ of the background voice. We used the root mean square error as a distance measure:

$$RMSE^+ = \sqrt{\frac{1}{25} \sum_{n=76}^{101} (\hat{F}_0(n) - F_0(n))^2}$$

$$RMSE^- = \sqrt{\frac{1}{25} \sum_{n=76}^{101} (\hat{F}_0(n) - F_0(n))^2},$$

where $RMSE^+$ was a distance from a positive probe and $RMSE^-$ was a distance from a negative probe. A $RMSE^+$ or $RMSE^-$ value within a tolerance range $r$ was considered as a model’s positive response to a (positive or negative) probe. Otherwise it was consid-
ered a negative response. We varied the criterion \( r \) and obtained the percentage of the true positive (TP) and false positive (FP) responses across all trials, for each value of \( r : TP(r) \) and \( FP(r) \). Plotting \( TP(r) \) against \( FP(r) \) 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:

\[
d' = \sqrt{2Z(AUC)},
\]

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 shows schematically a single trial of the simulated experiment.

### B. Computational simulations

| Simulation number | Corresponding human experiment | Foregr.-backgr. 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.             | "                              | Segreg. without oracle information (Fig. 3, B.2.2) | "                     | "                  | "                |
| 1.c.             | "                              | Formant-guided segregation (Fig. 3, B.2.3) | "                     | "                  | "                |
| 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.             | "                              | Segreg. without oracle information (Fig. 3, B.2.2) | "                     | "                  | "                |
| 2.c.             | "                              | Formant-guided segregation (Fig. 3, B.2.3) | "                     | "                  | "                |

Table I. Overview of the computational simulations in the model evaluation section.

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 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, right most panel). The results showed that discrimination between the attended and the 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-prime values dropped from $d' \approx 1.2$ to $d' \approx 0.2$ (see 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 trajectory generation we used the same procedure as 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. Trajectories crossed at least once in each varying dimension. To avoid 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-prime 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 step of the particle filter (cf. Sec. II B 3). 20 d-prime values for each condition were obtained.

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

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 estimation of \( F_0^F(n-1) \) and \( F_0^B(n-1) \) was available in every time step, would the glimpses decomposed based on this information be sufficient for the system to track \( F_0 \) of the voices? Earlier results by Josupeit and Hohmann (2017) showed that sparse periodicity-based features were distinctive for all speakers in a multi-talker set-up. Therefore, we expected that sPAF would encode the information about both voices in a competing voices scenario (at least for 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.

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 \( \hat{F}_0^F(n-1) \) and \( \hat{F}_0^B(n-1) \) to decompose the glimpses. We expected the discrimination performance to drop in comparison to F0-guided method from
Simulation 1.a., but to be above chance level for the trajectories varying in three dimensions.

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 previous F0 estimate and ground truth formant values used to synthesize 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 formants $F_1(n-1)$, $F_2(n-1)$, $F_1B(n-1)$ and $F_2B(n-1)$ 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 proximity of competing voices on attentive tracking performance. They compared 8 conditions: In each condition the trajectories were restricted to pass each other with a different minimum distance in the 3-dimensional feature space. Results showed that discrimination between the attended and the unattended voice improved continuously as the voice distance was increased (see Fig. 12).
We tested 4 conditions with different minimum distances. In each condition, the minimum 3-dimensional Euclidean distance between trajectories (in semitones) was restricted to fall within designated bin limits. Bin 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. For 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 trajectories crossed at least once in each dimension. We initialized trajectories of the competing voices in different $F_0$ ranges (either 100 – 250 Hz or 250 – 400 Hz). 50 random trajectory pairs per conditions 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-prime value was computed in each condition, as described in Sec. III A. The simulation was repeated 10 times to account for the randomness in the initialization, prediction and resampling step of the particle filter (cf. Sec. II B 3). 10 d-prime values for each condition were obtained.

Following the concept from Simulation 1, in this task we also obtained the model’s results for three model variants:

Simulation 2.a. F0-guided tracking, in which we expecteded to show the potential of correctly segregated sparse periodicity-based features,

Simulation 2.b. Tracking without oracle information, in which we analyzed limitations, which the model encounters when attempting to solve the attentive tracking task only based on the estimated $F_0$, and
Simulation 2.c. Formant-guided tracking, which we performed to test how much improvement could be brought to the model if the segregation stage was based on 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).

Simulation 1.a.

With this simulation we aimed at validating the periodicity-based glimpses in the context of $F_0$ tracking. The median d-prime value across all runs of the simulation was $d' = 1.64$ for voices varying in three dimensions and $d' = 1.42$ for voices varying only in $F_0$ (see Fig. 10, left panel). A good discrimination performance indicates that the system could track the $F_0$ trajectories of both voices. These results prove that glimpses segregated based on oracle information about $F_0$ in the preceding time step contain sufficient information to segregate foreground and background glimpses and to 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 human performance.

Simulation 1.b.
Figure 10. Simulation 1: Stream Segregation of Sources Varying in Just One Feature. Left panel: Results of Simulations 1.a. - 1.c. in terms of d-prime. Right panel: Stimulus conditions used in Simulation 1. Crosses in magenta denote the mean human results from Woods and McDermott (2015). Boxplots summarize the distribution across 20 d-prime values in each condition. Statistical significance of the difference between tested conditions is measured with a t-test, and is represented in the plot with symbols * (for $p \leq 0.05$), ** (for $p \leq 0.01$), or *** (for $p \leq 0.001$).

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 chance level for most conditions. However, as shown in Fig. 10 (middle panel), the results dropped significantly in comparison to Simulation 1.a.. The median d-prime value across all runs of the simulation was $d' = 0.39$ for voices varying in three dimensions and $d' = 0.21$ for 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 voice. When at least one of the estimates is not
correct, a particle filter responsible for one voice may receive channel sets $P_{cn}$ that originated from the other voice. Since 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. Since in the subsequent step the segregation again depends on the estimated $F0$, the error propagates potentially until there are no more particles covering the true $F0$ 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 $F0$ crossings (see Fig. 11.A.)**

  At the $F0$ crossings, the foreground particle filter could take over the tracking of the background voice and/or vice versa. When the $F0$ 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, 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 $F0$ (see Fig. 11.B.)**

  The second problem was related to tracking the harmonics or subharmonics of the correct $F0$. For example, period glimpses originating from $F0 = 115\text{ Hz}$ in general convey a relatively high likelihood for the hypothesis that $F0 = 330\text{ Hz}$. At some time
Figure 11. Typical tracking errors encountered in Simulation 1.b. Each plot presents a single trial of the attentive tracking experiment. Blue color indicates foreground voice, red color indicates background voice. The size and color saturation of the dots in plots A-C correspond to particle weights. Gray dashed rectangles help to locate the discussed error in the image. A. Identity switch at the $F_0$ crossing, B. Following a harmonic of a $F_0$ of the competing voice. C. Following a subharmonic of the correct $F_0$, D. Solid lines illustrate the trajectories of harmonics or subharmonics of the true $F_0$, dashed line represents potential estimated trajectory: The tracking can potentially be misled at every intersection point.
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 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, 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 observation extracted from competing voices was not sufficient and did not reproduce human results.

**Simulation 1.c.**

In this simulation, we investigated whether additional information about the formant frequencies could possibly prevent period glimpses from being assigned to the wrong stream. As expected, with the formant-guided tracking we obtained a significant improvement in 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 voices varying in three dimensions and $d' = 0.2$ for voices varying only in $F_0$ (see Fig. 10).
Based on the instantaneous energy distribution across 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 $F0$, a difference in $F1$ and $F2$ helped to disentangle the origin of 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 $F1$ and $F2$ dimensions, the model had to rely purely on the estimated $F0$ as in Simulation 1.b. Simulation results were in agreement with this expectation - discrimination performance for voices varying only in $F0$ 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 human results.

**B. Simulation 2**

In this Section the results of Simulation 2 are presented. This simulation modeled the listening experiment *Effect of Source Proximity* (see Sec. III B 2).

**Simulation 2.a.**

In this simulation oracle $F0$ tracks were used to segregate sPAF. The median d-prime values were (from small to large minimum distance): 2.07, 1.99, 2.35, and 2.36. Corresponding d-prime values for humans were 0, 0.3, 1.3, 2. The model clearly outperforms human listeners. Even in conditions with low minimum distance it reaches very good discrimination performance. This shows that 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
Figure 12. Simulation 2: Effect of Source Proximity. Left panel: Results of Simulations 2.a. - 2.c. in terms of d-prime. Right panel: Stimulus conditions used in Simulation 2. Crosses in magenta denote the mean human results from Woods and McDermott (2015). Boxplots summarize the distribution across 10 d-prime values in each condition. Statistical significance of the difference between tested conditions is measured with a t-test, and is represented in the plot with symbols * (for $p \leq 0.05$), ** (for $p \leq 0.01$), or *** (for $p \leq 0.001$).

and track them. Since the the feature segregation stage in this simulation is F0-guided, the results represent an optimal case. The increase of d-prime as a function of the minimum distance was not observed. For most conditions the model is much better than 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 F0 tracks.

Simulation 2.b.
In this simulation no oracle knowledge was used to segregate sPAF. The median d-
prime values were (from small to large minimum distance): 0.16, 0.45, 0.55, and 0.52.
Corresponding d-prime values for humans were 0, 0.3, 1.3, 2. The results for the first two
conditions – minimum distance of 0.5 semitones and 2.5 semitones – are in a good agreement
with human results. Discrimination performance is low, indicating that the task is difficult.
Nevertheless, increasing the minimum distance between trajectories from 0.5 to 2.5 semitones
results in a significant improvement. For the lowest minimum distances between trajectories
humans lose the ability to distinguish between the sources. The reason might be the limited
resolution of attention or the perceptual fusion of 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 low minimum distances, the attentive tracking in humans is prone to similar errors as
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 human results. At the distance of 5.5 semitones the model
cannot reach 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 human listeners,
and only a slight, though significant, improvement for the model. Somewhere between the
distance of 2.5 and 5.5 humans start to use the available information to separate the features
and segregate voices. The model, on the other hand, is still erroneous and not capable to
segregate features at the crossings.
Interestingly, model results for the minimum distance of 7.5 semitones are slightly worse than for 5.5 semitones. A possible explanation could be that increasing distance raises the likelihood that the model will follow a track one octave away from the correct $F_0$.

Simulation 2.c.

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

For the minimum distance of 5.5 semitones the model coincides with human results. This minimum distance provides enough separation in the feature space, allowing both humans and the model to use formant frequencies to segregate voices. However, further increase of 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 formants, which allows to predict the energy distribution across frequency channels, 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 correctly estimated $F_0$ was used it in the foreground-background segregation stage. This requires a model of the joint likelihood for $F_0$, $F_1$, and $F_2$ and 3-dimensional tracking.
V. GENERAL DISCUSSION

In the present study, we combined the concepts of auditory glimpses, statistical top-down knowledge, and Bayesian inference to model the attentive tracking of voices. We developed a unified computational framework consisting of sparse periodicity-based feature extraction, single-voice probability models, and particle filter tracking. We evaluated the model using 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 human listeners can attentively track a voice in the presence of a second voice if their parameters maintain sufficient separation in the parameter space. Experiment Effect of Source Proximity showed that attentive tracking gets worse as the distance between the varying voice parameters decreases, and 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 oracle $F0$ leads to a performance much better than that of human listeners. This showed that, even for conditions with only one varying feature or with very close parameter transitions, there is enough information in the acoustic signal to solve the attentive tracking task. 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 to encode this information when the voices pass close in the parameter
space.

Adverse conditions, in which humans had difficulties to attentively track a voice, were
most accurately reproduced by the model with the foreground-background segregation based
on the estimated $F_0$, without any oracle knowledge. In these conditions our model suffers
from the confusions at the crossings of $F_0$ or crossings of the harmonics of $F_0$. In the
absence of information about the formants of voices, the ambiguity at the crossing cannot
be resolved. Two voices share the frequency space, and after extracting and segregating
sPAF, the remaining evidence for one voice is typically incomplete. This incomplete pattern,
can produce higher likelihood for a wrong hypothesis. Various studies mention multimodal
distributions of pitch matches for complex tones with few harmonics (Boer, 1956; Cariani
and Delgutte, 1996; Schouten et al., 1962; Terhardt, 1989). Additionally, models of pitch of
concurrent sounds also suffer from the ambiguous $F_0$ estimates (Assmann and Summerfield,
1990; Saddler et al., 2020; Schouten et al., 1962). Ambiguity does not mean randomness;
in all mentioned studies the pitches evoked by a stimulus are in systematic relationships
to 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 observations in
humans. According to the simulations results segregation seems to be the major limitation,
not masking.
Considering this inherent ambiguity of pitch, there must be additional mechanisms that help listeners solve the attentive tracking task. 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) that to attentively track a voice, the auditory system binds several task-related qualities together - in this case all varying dimensions: F0, F1, and F2. The results confirmed that it may be possible to improve the model’s performance using a multi-dimensional likelihood model comprising F0, F1, and F2. To achieve this, the interdependencies between these parameters and and their impact on the sparse periodicity-based auditory features have to be investigated. However, an extension of the model to blindly track information in multiple dimensions is beyond the scope of the current study and subject to future work (see Sec. V D).

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

Madsen et al. (2019) found differences in attentive tracking between musicians and non-musicians, which were not found for speech perception in noise. The authors argued that the advantage of musicians can be due to their experience in making fine-grained auditory discrimination judgments. In our model the experience of musicians could be reflected in the computational block top-down knowledge. For musicians, the probabilistic voice models – state transition and observation statistics – might be more accurate than for non-musicians. 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 individual musical voices in the JS Bach’s *The Art of the Fugue*. Performance depended on the degree of hearing impairment, number of voices in the mixture, and on the timbral heterogeneity between the voices. Simulating timbral heterogeneity might require tracking more than parameters $F_0$, $F_1$, and $F_2$, but, at least on a conceptual level, this task can be well defined in the proposed computational framework. To simulate the effect of hearing impairment, the parameters of the sPAF extraction could be modified, so 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 in the background stream, the harder it is to separate it 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 statistical regularities across trials can be observed. To simulate this phenomenon, additional modality responsible for the adaptation of top-down knowledge could be implemented. The
adaptation component seems natural to add in the further versions of this model, since 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 other experiments 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 can be 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 particles with small weights and duplicates 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 anymore. In that scenario, initializing 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

The present article builds upon 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, on the other hand, a series of modeling studies related to sparse periodicity-based auditory features (Josupeit and Hohmann, 2017; Josupeit et al., 2016, 2018).

In the study by (Nix and Hohmann, 2007) multidimensional, nonlinear statistical filtering was proposed as a tool for filtering 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 location and spectral shape. The second line of research investigated the role of 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. (2018) 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 one lacked a more specialized feature space and the latter did not include sequential processing. With this study, we bridge these ideas: We combine 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 acoustic feature binding, high-dimensional state tracking should be performed. This requires knowledge of the non-linear 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 the 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: 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 tracking of foreground and background: Apart from the initialization with the attention prior model, the particle filters have identical properties. Both execute resampling step, which mimics the attentive tracking. In other words, the model attentively tracks both the foreground and the background. This is a simplified approach, which does not reflect the differences between 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 identical top-down models for both the foreground and the 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 detection of vibrato between the attended and unattended voice. 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 foreground and background could not be simulated with our model, this limitation should be tackled in
the further 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 time-varying parameters, used to evaluate
the current model, are simple but challenging stimuli. One voice is assigned to the attended
foreground and the other 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 time-frequency 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 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 building blocks of the model can be extended towards 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 stream. 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 sparse periodicity-based auditory features, sequential Monte Carlo sampling, and single voice probability models. We proposed a $F_0$ observation model, which describes the statistical relationship between generative fundamental frequency and observed period glimpses. Comparison of the model with human performance showed that although optimally segregated sPAF convey sufficient information to perform the attentive tracking task, humans are not always able to decode this information. Joint modeling of $F_0$ and spectral profile (formants) is required to reach human performance levels. This shows that a combination of features may be used by the auditory system in order 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-ID 352015383 – SFB 1330.

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 Number R01DC015429. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Aitchison, L., and Lengyel, M. (2017). “With or without you: predictive coding and bayesian inference in the brain,” Current opinion in neurobiology 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 Transactions on signal processing 50(2), 174–188.

Assmann, P. F., and Summerfield, Q. (1990). “Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies,” The Journal of the Acoustical Society of America 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?,” The Journal of the Acoustical Society of America 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 Jr, G. (2017). “Use of a glimpsing model to understand the performance of listeners with and without hearing loss in spatialized speech mixtures,” The Journal of the Acoustical Society of America 141(1), 81–91.

Best, V., Ozmeral, E. J., Kopčo, N., and Shinn-Cunningham, B. G. (2008). “Object continuity enhances selective auditory attention,” Proceedings of the National Academy of Sciences 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. cam-
bridge, ma, us”.

Bressler, S., Masud, S., Bharadwaj, H., and Shinn-Cunningham, B. (2014). “Bottom-up in-
fluences of voice continuity in focusing selective auditory attention,” Psychological research
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,” Journal of neurophysiology 76(3), 1717–1734.

Carlyon, R. P. (2004). “How the brain separates sounds,” Trends in cognitive sciences 8(10),
465–471.

Carlyon, R. P., Cusack, R., Foxton, J. M., and Robertson, I. H. (2001). “Effects of at-
tention and unilateral neglect on auditory stream segregation.,” Journal of Experimental
Psychology: Human Perception and Performance 27(1), 115.

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). “Probabilistic models of cognition:
Conceptual foundations”.

Chen, Z. et al. (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,” The Journal of the acoustical society of America 25(5), 975–979.

Cohen-Lhyver, B., Argentieri, S., and Gas, B. (2018). “The head turning modulation sys-
tem: An active multimodal paradigm for intrinsically motivated exploration of unknown
environments,” Frontiers in neurorobotics 12, 60.
Cooke, M. (2006). “A glimpsing model of speech perception in noise,” The Journal of the Acoustical Society of America 119(3), 1562–1573.

Darwin, C. (2007). “Listening to speech in the presence of other sounds,” Philosophical Transactions of the Royal Society B: Biological Sciences 363(1493), 1011–1021.

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,” Frontiers in integrative neuroscience 14.

Dietz, M., Ewert, S. D., and Hohmann, V. (2011). “Auditory model based direction estimation of concurrent speakers from binaural signals,” Speech Communication 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 research 1220, 234–245.

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, pp. 2792–2795.

Elhilali, M., and Shamma, S. A. (2008). “A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation,” The Journal of the Acoustical Society of America 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 biology 7(6).
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 Communication 27(3-4), 281–298.

Friston, K., Adams, R., Perrinet, L., and Breakspear, M. (2012). “Perceptions as hypotheses: Saccades as experiments,” Frontiers in psychology 3, 151.

Garrido, M. I., Kilner, J. M., Stephan, K. E., and Friston, K. J. (2009). “The mismatch negativity: a review of underlying mechanisms,” Clinical neurophysiology 120(3), 453–463.

Gregory, R. L. (1980). “Perceptions as hypotheses,” Philosophical Transactions of the Royal Society of London. B, Biological Sciences 290(1038), 181–197.

Gregory, R. L. (1997). “Knowledge in perception and illusion,” Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 352(1358), 1121–1127.

Hafter, E. R., Sarampalis, A., and Loui, P. (2008). “Auditory attention and filters,” in Auditory perception of sound sources (Springer), pp. 115–142.

Heilbron, M., and Chait, M. (2017). “Great expectations: is there evidence for predictive coding in auditory cortex?,” Neuroscience.

Helmholtz, H. v., and Von, H. (1878). “The facts in perception,” Helmholtz, Epistemological Writings, tr. M. Lowe, eds. R. Cohen and Y. Elkana. D. Reidel Publishing.

Hohmann, V. (2006). “Method for extracting periodic signal components, and apparatus for this purpose” US Patent App. 11/223,125.

Josupeit, A., and Hohmann, V. (2017). “Modeling speech localization, talker identification, and word recognition in a multi-talker setting,” The Journal of the Acoustical Society of America 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,” The Journal of the Acoustical Society of America 139(5), 2911–2923.

Josupeit, A., Schoenmaker, E., van de Par, S., and Hohmann, V. (2018). “Sparse periodicity-based auditory features explain human performance in a spatial multitalker auditory scene analysis task,” European Journal of Neuroscience.

Kaya, E. M., and Elhilali, M. (2017). “Modelling auditory attention,” Philosophical Transactions of the Royal Society B: Biological Sciences 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,” Journal of Experimental Psychology: Human Perception and Performance 37(4), 1140.

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 Systems with applications 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, 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,” Scientific reports 9(1), 1–10.
McDermott, J. H. (2009). “The cocktail party problem,” Current Biology 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,” Proceedings of the National Academy of Sciences 111(18), 6792–6797.

Näätänen, R., Gaillard, A. W., and Mäntysalo, S. (1978). “Early selective-attention effect on evoked potential reinterpreted,” Acta psychologica 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 transactions on audio, speech, and language processing 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,” Nature communications 9(1), 2122.

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). “Probabilistic brains: knowns and unknowns,” Nature neuroscience 16(9), 1170.

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 Journal of Selected Topics in Signal Processing 13(2), 203–205.
Saddler, M. R., Gonzalez, R., and McDermott, J. H. (2020). “Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception,” bioRxiv.

Sanborn, A. N., and Chater, N. (2016). “Bayesian brains without probabilities,” Trends in cognitive sciences 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,” The Journal of the Acoustical Society of America 34(9B), 1418–1424.

Schroeder, M. R. (1968). “Period histogram and product spectrum: New methods for fundamental-frequency measurement,” The Journal of the Acoustical Society of America 43(4), 829–834.

Schröger, E., Marzecová, A., and SanMiguel, I. (2015). “Attention and prediction in human audition: a lesson from cognitive psychophysiology,” European Journal of Neuroscience 41(5), 641–664.

Shamma, S., and Dutta, K. (2019). “Spectro-temporal templates unify the pitch percepts of resolved and unresolved harmonics,” The Journal of the Acoustical Society of America 145(2), 615–629.

Shamma, S. A., and Micheyl, C. (2010). “Behind the scenes of auditory perception,” Current opinion in neurobiology 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 in cognitive sciences* **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,” *Frontiers in Psychology* **12**.

Snyder, J. S., Gregg, M. K., Weintraub, D. M., and Alain, C. (2012). “Attention, awareness, and the perception of auditory scenes,” *Frontiers in Psychology* **3**, 15.

Spille, C., Ewert, S. D., Kollmeier, B., and Meyer, B. T. (2018). “Predicting speech intelligibility with deep neural networks,” *Computer Speech & Language* **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), pp. 145–170.

Szabó, B. T., Denham, S. L., and Winkler, I. (2016). “Computational models of auditory scene analysis: a review,” *Frontiers in Neuroscience* **10**, 524.

Terhardt, E. (1989). “On the role of ambiguity of perceived pitch in music,” in *Proc. 13th ICA Belgrade*, pp. 35–38.

Woods, K. J., and McDermott, J. H. (2015). “Attentive tracking of sound sources,” *Current Biology* **25**(17), 2238–2246.
Woods, K. J., and McDermott, J. H. (2018). “Schema learning for the cocktail party problem,” Proceedings of the National Academy of Sciences 115(14), E3313–E3322.

Wrigley, S. N., and Brown, G. J. (2004). “A computational model of auditory selective attention,” IEEE Transactions on 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,” Journal of Neuroscience 30(36), 12084–12093.

VII. 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, $n$ the time index and $s$ the channel of a binaural signal. In the current study, we investigated whether these features can be used to track $F0$ of the voices. Hence, from all four periodicity-based features proposed by Josupeit and Hohmann (2017), $P_{cns}$, $E_{cns}$, $T_{cns}$, $L_{cns}$, we use only the pitch-related $P_{cns}$. Since we use 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 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 band-pass filter...
(500 – 2000 Hz) and gammatone filterbank with 23 channels \( c = 1, \ldots, 23 \) 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 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 fine structure channels with centre frequencies \( f_c < 1400 \) Hz and gammatone filter with a constant center frequency of 135 Hz and a bandwidth of 16.9 Hz for envelope channels with centre 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 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 fine structure and modulation channels: all frequency channels undergo the same transformations, which reduces the number of free parameters in the model. Not separating between envelope and fine structure means that beating from unresolved harmonics and the resolved harmonics contribute with the same weight to
the $F0$ estimate. This is a simplification, as these two components may be weighted 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 Figure 13.A.2. $P'$ is varied from $1/700$ Hz 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 total energy. It is calculated as the mean square amplitude. The energy of the base function is termed periodic energy $E_{P,cn}(P')$ and corresponds to the mean square amplitudes of the base function. Lastly, 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 the total energy is computed, which is called the synchrum:

$$\text{synch}_{cn}(P') = \frac{E_{P,cn}(P')}{E_{\text{tot},cn}(P')}.$$  \hspace{1cm} (15)

A synchrum value equal to 1 means that the signal is fully repeating itself with a period corresponding to the window length $P'$; 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 synchrum $\text{synch}_{cn}(P')$ is plotted in Figure 13.A.3. Period glimpses $P_{cnm}$, where $m$ is a glimpse index, are defined as the period values $P'$ corresponding to local maxima (values larger than the neighboring values) of the synchrum, which meet the following two criteria:

(a) The first criterion filters out the noisy time-frequency ($cn$) bins, which have a lower degree of periodicity than the speech $cn$ bins. If the maximum value of the synchrum exceeds the threshold $T_1$, then the time-frequency bin $cn$ is considered to originate from the speech and the glimpses are extracted from the synchrum. Otherwise no glimpses are extracted for that $cn$ bin.

(b) The second criterion is used to find the salient local maxima of the considered synchrum that carry the information about the period of the speech signal. Every local maximum that exceeds the threshold $T_2$ is considered a period glimpse $P_{cnm}$ (blue circles in Figure 13.A.3.).

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

Josupeit and Hohmann (2017) defined the thresholds for 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 glimpsing thresholds 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. Sec. VII.B). The resulting $T_1$ decreases with increasing channel center frequency, and $T_2$ is set relatively to the threshold $T_1$: $T_2 = 0.9 \cdot T_1$ (see Figure 14). The values are in good agreement with the thresholds from Josupeit and Hohmann (2017).

![Figure 14. Glimpse extraction threshold values.](image)

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

- Instead of performing periodicity analysis every 10 ms, we do it every 20 ms
- Instead of a hard distinction between fine structure and envelope channels, we process all 23 channels in the same way.
- Instead of the differentiation stage and additional gammatone filtering after half-wave rectification, we use the 40 Hz high-pass filter.
- Instead of analyzing the periods $P'$ in a range between $1/1400$ Hz to $1/80$ Hz, we do it in a range between $1/700$ Hz to $1/80$ Hz.
- Instead of fixed threshold values, we use frequency-dependent threshold values.
Instead of extracting all periodicity-based features $P_{cn}$, $E_{cn}$, $T_{cn}$ and $L_{cn}$, we extract only the period glimpses $P_{cn}$.

**B. Glimpse threshold analysis**

As described in the Appendix (Sec. VII A), glimpses in each time-frequency bin are extracted from the synchrum, an example of which is depicted in Figure 15. If the global maximum exceeds $T_1$, glimpses are extracted. This is done using the second threshold $T_2$ - each local maximum above $T_2$ defines a period glimpse.

The goal of this analysis was to determine thresholds, which will lead to the best discrimination between two categories: periodicity of a single voiced source and periodicity of a sound mixture. We used 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

![Normalized synchrum in one time-frequency bin](image-url)
in the second category. Both signals were 200 s long with a sampling rate of $f_s = 16000$ Hz. For both signals, we extracted and stored all the local maxima of the synchra for each time-frequency bin ($c_n$). Thresholds that lead to the best discrimination between the voice and the speech shaped noise are chosen as optimal thresholds. We consider glimpsing thresholds 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 that leads to the best discrimination between voice and noise frames. We make an assumption that all bins $c_n$ 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 various $T_1$ values, we obtain the percentage of noise T-F bins that were incorrectly classified as speech T-F bins – false positive rate (FPR), and the percentage of speech T-F bins that were correctly classified as speech – true positive rate (TPR). By relating those two measures, we obtain a receiver operating characteristic (ROC) curve. For each channel, we obtain one ROC curve (see Figure 16), as well as the area under this curve (AUC) (see Figure 17), which tell us about the performance of the detection task.
The ROCs and AUCs show that the discrimination task is possible (clearly better than chance level). The next step is to find a cut-off point on the curve that 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 false positive rate (FPR). On the other hand, we do not want to get rid of the glimpses originating from speech, so we make sure that the true positive rate (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 thresholds chosen this way for 23 frequency channels and the corresponding TPR that can be achieved with the chosen thresholds.
Figure 17. Area under the ROC curve. It exceeds 0.5 in all channels, which means the classifier is working properly.

2. **Finding the optimal $T_2$**

We want to find the second threshold value $T_2$ that will lead to 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 time-frequency bins classified as speech bins (using the first threshold $T_1$). In each bin, we normalize all the local maxima to the global maximum value in this bin. This way, we obtain the distribution, which shows how much the local maxima on average deviate from the global maxima.

We observe a bimodal distribution, with the high counts at both ends of the range (see Fig. 19). We interpret this in the following way: Values close to 0 are very frequent and they...
correspond to spurious peaks, which do not carry information about the actual periodicity of the signal. Values that are very close to 1 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 $T_2$ as 0.9 of the threshold $T_1$. 

Figure 18. Upper panel: Thresholds chosen with a criterion of 2% permissible false positive rate. Lower panel: TPR that can be achieved using the chosen threshold. In the worst case, the TPR is about 50%.
Figure 19. Distribution of the local maxima in speech frames relative to the global maximum.

Gammatone channels 20 $f_c = 3388$ Hz

VIII. SUPPLEMENTARY MATERIAL

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 the further studies.

We simulated in total 6 possible combinations of conditions, depicted and described in Figure 20. In each of these conditions we used 100 random trajectory pairs. Each trajectory pair was used twice as a trial. We initialized 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-prime value was computed. This computational experiment was repeated for 3
settings of the model: with formant-guided tracking, without using any oracle information, and with F0-guided tracking.

There are two main effects reflected in this supplementary data:

**Influence of formant dynamics.** In (Woods and McDermott, 2015) the authors compared attentive tracking of voices varying and crossing in all three dimensions to attentive tracking of voices varying and crossing only in F0. The voices with identical and constant formants could not be discriminated by human listeners. These results suggested that multiple features allow accurate streaming where single features cannot. However the experiment did not quantify the influence of feature dynamics on the segregation. To examine this, we added additional conditions with trajectories varying and crossing in F0 and with formant frequencies, which were static over time, but different between the voices. Results, depicted in Figure 21 showed that constant separation between the formant frequencies aids performance in the attentive tracking. It is easier to distinguish voices if their formants do not vary over time, which proves that, at least for the model, there is a significant influence of formant dynamics on the attentive tracking.

**Influence of trajectory crossing.** In (Woods and McDermott, 2015) trajectory pairs crossed at least once in each feature dimension. This way the authors ensured that the identification of the voices was not done on the basis of one single feature. However, it was not tested if the voices with 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 tasks, we performed the additional simulations for stimuli with non-crossing trajectories. Results, depicted in Figure 21 showed that 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 non-crossing trajectories is very
good. This proves that, for the model, it is easier to track the noncrossing trajectories and
that the F0 crossings are the main reason for the model to fail in the tracking task.

Figure 20. Six supplementary conditions simulated by the model. There are two different types of
F0 trajectories: 1) crossing F0 (x) and 2) non-crossing F0 (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 voices
(S).
Figure 21. Additional experiments: Abbreviations for condition names: x: crossing F0, nx: non-crossing F0, I- formants are constant and identical between the voices, V: formants are varying and different between the voices, S - formants are static, but different between voices.