PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Glimpsed periodicity features and recursive Bayesian estimation for modeling attentive voice tracking

Joanna LUBERADZKA, Hendrik KAYSER, Volker HOHMANN

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

Abstract

Computational models are a way of approaching research questions related to auditory perception. One relevant question is how we are able to follow and understand speech in complex acoustic scenes. Previous studies suggest that for tracking a speaker in such conditions, humans use (1) sparse, speaker-related bits of robust information - 'auditory glimpses' and (2) a mechanism of predictive coding with a movable locus of attention. The goal of the present study is to develop a computational model for attentive tracking of voices, which takes these two aspects into account. We model auditory glimpses using Glimpsed Periodicity Features and predictive coding using Recursive Bayesian Estimation. We assume that perception is organized into an attended foreground and unattended background. We propose parallel particle filters - one for each category - to track the concurrent events. In this approach, each incoming glimpse is associated with either foreground or background based on accumulated evidence. Simulations with artificially generated data of a 'glimpsing' nature (sparse, robust) showed that this approach is suitable to track multidimensional parameter trajectories of two competing sources. This suggests the potential of the method to track simultaneously active voices based on the Glimpsed Periodicity Features.

Keywords: Computational Auditory Scene Analysis, Particle Filtering

INTRODUCTION

Human listeners can segregate the sound mixture into streams and attentively follow a desired voice. This perceptual ability can be described as a combination of the top-down and bottom-up processes: On one hand, it is a result of bottom-up processing, i.e. simultaneous grouping ([3]) of the sound components available at a time instance by the similarities in their physical properties like harmonic structure, time onset or spatial direction ([7]). Previous studies suggest that in complex acoustic conditions the auditory system may use primarily undisrupted pieces of information - *auditory glimpses*, which provide reliable cues for simultaneous grouping and can be processed with single-voice models ([14, 19, 2, 6]).

On the other hand, to attentively track a voice the auditory system has to decompose those sparsely occurring salient glimpses into streams of information and then maintain them over time. We know that sequential grouping ([3]) is possible when the perceptual distance between the components of the individual streams is large enough ([22]). However, it has been furthermore shown, that attentive tracking is possible even if there are no constant dissimilarities between the competing voices and their properties vary in time ([24]). This speaks for the theory that sequential grouping uses the contextual knowledge to integrate the information over time. Another obviously essential aspect that influences both the decomposition of the auditory glimpses into streams as well as holding on to the desired streams is attention ([8]).

Many recent studies propose Bayesian estimation as a computational framework that integrates the bottom-up and top-down processes including attention ([21, 16, 23, 18, 4, 5, 11, 9]). We present a computational model of attentive tracking of voices, which takes the above-discussed aspects into account and thus contributes to understanding the speech perception in complex auditory scenes. The novelty of this approach lies in integrating the Glimpsed Periodicity Features ([12]) and probabilistic knowledge models into a Bayesian sequential estimation framework. We also report the results of preliminary numerical simulations.

MODEL

We model an auditory scene consisting of two simultaneously active voices - foreground and background voice. Each voice is characterised by a *hidden state vector*, defining the most important properties (fundamental frequency $F0$, first two formant frequencies $F1$ and $F2$ and direction of arrival α) at a given time instance t:

$$\vec{s}_{fg}(t) = \begin{pmatrix} F0_{fg}(t) \\ F1_{fg}(t) \\ F2_{fg}(t) \\ \alpha_{fg}(t) \end{pmatrix}, \; \vec{s}_{bg}(t) = \begin{pmatrix} F0_{bg}(t) \\ F1_{bg}(t) \\ F2_{bg}(t) \\ \alpha_{bg}(t) \end{pmatrix}. \tag{1}$$

The scene analysis task - attentively following one voice in the presence of the second voice - is represented as simultaneous tracking of the *hidden states* of both voices $\vec{s}_{fg}(t)$ and $\vec{s}_{bg}(t)$.

At the input to the model, we have a time frame of the binaural signal containing the mixture of two competing voices. At the output, we receive the estimate of a temporal state of the foreground (attended) and background (unattended) voice - $\hat{\vec{s}}_{fg}(t)$ and $\hat{\vec{s}}_{bg}(t)$. Although we track both voices, the tracking of the attended voice is, in contrast to the tracking of unattended voice, initialized using an informative prior (see 2.4.1). In this study, we focus on tracking one dimension of the hidden state - fundamental frequency $F0$, which yields:

$$\hat{\vec{s}}_{fg}(t) = F0_{fg}(t), \; \hat{\vec{s}}_{bg}(t) = F0_{bg}(t). \tag{2}$$

The computational framework that we propose for modeling attentive tracking, is depicted in Figure 1. The *glimpsed feature extraction* stage (Fig. 1A.) simulates bottom-up processing of the auditory system - salient periodicity glimpses are extracted from the binaural input signal. The *glimpsed feature grouping* stage assigns the sparse glimpses into competing auditory streams (Fig. 1B.). Foreground and background glimpses enter the *state estimation* stage (Fig. 1C.), which simulates the inference in the auditory system. It consists of two parallel particle filters, which sequentially estimate the foreground and background *hidden state* in a Bayesian framework. It requires probabilistic models (PMs), that simulate the top-down *world knowledge* (Fig. 1D.) - perceptual and contextual knowledge that the human brain has learned throughout life as well as the temporal attention.

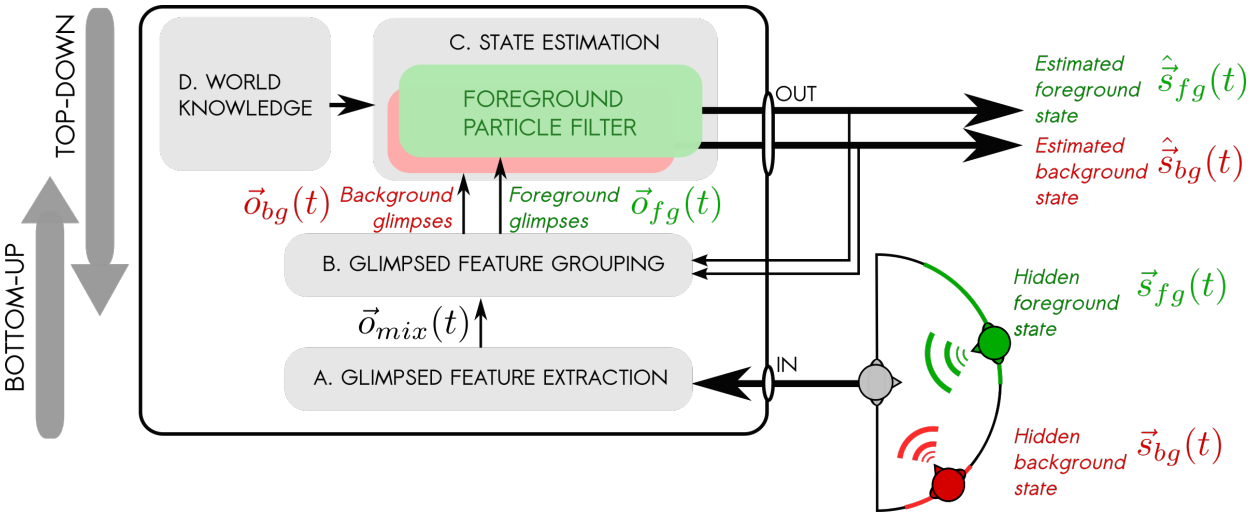

Figure 1. Schematic representation of the computational framework for modeling the attentive tracking.

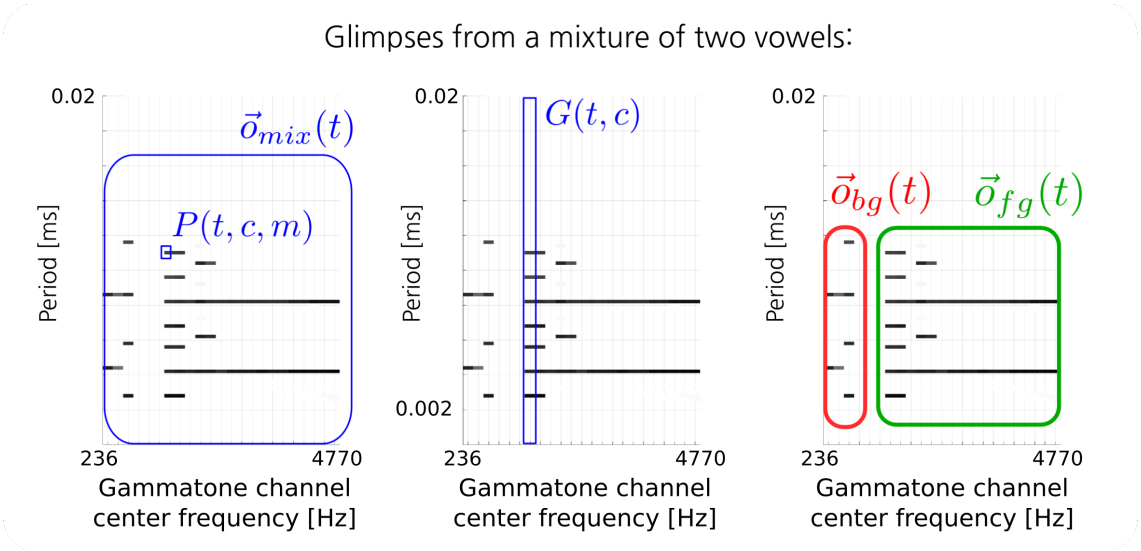

Figure 2. Example of glimpsed periodicity pattern. $\vec{o}_{mix}(t)$ is the total information that is available to the model at the input. It consists of glimpse multi-sets in 23 frequency channels $G(t,c)$, that contain individual glimpses $P(t,c,m)$, which indicate periods that are in the best alignment with the periodicity of the signal. Glimpse multi-sets $G(t,c)$ are assigned either to foreground or to background, yielding observations $\vec{o}_{fg}(t)$, and $\vec{o}_{bg}(t)$. All plots present the average glimpse pattern that results from accumulating glimpses extracted from a 5 s of a mixture of vowel *A* at 242 Hz and vowel *U* at 115 Hz.

2.1 Glimpsed feature extraction

The feature extraction stage simulates how the acoustic signals at the eardrums are transformed into feature space, used in the further processing of the auditory system. In our model, we use the approach developed in [12, 13, 14], which simulates the extraction of *auditory glimpses*, predominantly used in complex acoustic environments. The core idea of the method is to use periodicity as a footprint of speech in a sound mixture ([17]). The feature extraction stage, consisting of auditory pre-processing and periodicity analysis, reveals the dominant periods in different frequency bands, here considered as glimpses (See Figure 2). We define the output of the feature extraction at time instance t as an observation vector $\vec{o}_{mix}$:

$$\vec{o}_{mix}(t) = \{G(t,c), |c = 1,...,23\}, \tag{3}$$

where $G(t,c)$ is a glimpse multi-set in channel c, containing $M(t,c)$ salient periodicity values $P(t,c,m)$:

$$G(t,c) = \biguplus_{m=1}^{M(t,c)} P(t,c,m). \tag{4}$$

The fundamental property of the glimpsing approach is that it gets rid of all the noisy components and passes a 'data point' containing only clean bits of information, that are sparsely available in the signal. Number of glimpses in an observation vector $\vec{o}_{mix}(t)$ may thus vary, depending on the complexity of the input signal.

2.2 Glimpsed feature grouping

This stage of the model simulates how the bottom-up cues are assigned to established auditory streams. The observation vector $\vec{o}_{mix}(t)$ with the glimpses is, analogously to (2), decomposed into foreground observation vector $\vec{o}_{fg}(t)$ and background observation vector $\vec{o}_{bg}(t)$, i.e,. each glimpse multi-set $G(t,c)$ is assigned to either $\vec{o}_{fg}(t)$ or

$\vec{o}_{bg}(t)$.

$$\vec{o}_{fg}(t) = \{G(t,c)\ \forall\ c \mid p\big(G(t,c)|\hat{\vec{s}}_{fg}(t)\big) > p\big((G(t,c)|\hat{\vec{s}}_{bg}(t)\big)\}$$
$$\vec{o}_{bg}(t) = \{G(t,c)\ \forall\ c \mid p\big(G(t,c)|\hat{\vec{s}}_{fg}(t)\big) < p\big((G(t,c)|\hat{\vec{s}}_{bg}(t)\big)\} \tag{5}$$

The decision is made by evaluating the *observation statistics* at the previous state estimate and choosing the more likely stream for each multi-set.

2.3 State estimation

This stage of the model simulates how the auditory system integrates the bottom-up sensory input with the top-down knowledge in order to make sense of an auditory scene. Numerous studies mention Bayesian estimation as a plausible model of the inference in the human brain both in a general view on cognition ([18, 4, 5]), as well as in the context of auditory perception ([20, 11, 9]). According to the Bayesian view on the perception, the brain generates expectations using top-down knowledge and recursively compares them with the information 'at the output of' sensory organs.

We follow this line of research and use two parallel particle filters (PF) - one responsible for the foreground and one for the background. PFs sequentially calculate the estimates of the foreground and background state - $\hat{\vec{s}}_{fg}(t)$ and $\hat{\vec{s}}_{bg}(t)$, using PMs provided by the *World knowledge*. PMs and their counterparts in the auditory system are discussed in the following section.

2.4 World knowledge

Sensory information collected from the environment is often ambiguous. To interpret it correctly, the human brain most likely uses high-level cognitive information - here called the *world knowledge*. It includes conceptual knowledge, that is evolutionally hard-wired or learned by experience, and goal-oriented attention.

2.4.1 Attention prior

Attention can be seen as process that dictates the allocation of the limited neural resource ([23]). A common approach of modeling attention in a Bayesian framework is to treat it as prior information, which guides the expectation and reduces the number of possible explanations of the stimulus. In our modeling framework we differentiate between the attended and the unattended voice. The initial expectation is represented as:

$$p\big(\vec{s}_{fg}(0)\big) = \mathcal{N}\big(\vec{s}_{fg}(0); \Sigma = \mathrm{diag}(10 * \sigma)\big),\ \sigma = \begin{pmatrix} \sigma_{F0} = 0.5 \\ \sigma_{F1} = 1 \\ \sigma_{F2} = 5 \\ \sigma_{\alpha} = 1 \end{pmatrix}. \tag{6}$$
$$p\big(\vec{s}_{bg}(0)\big) = 1 - p\big(\vec{s}_{fg}(0)\big)$$

The initial hypotheses set is drawn from a normal distribution centered around the true initial state of the voice, whereas for the unattended voice the hypotheses are drawn from a distribution covering 'everything but the attended voice state'.

2.4.2 State transition

The evolution of the fundamental frequency, formants, and DOA in time is naturally limited due to physical constraints of the speech production process. Throughout life, human listeners internalise this knowledge and are able to use it to predict the incoming events. In our attentive tracking model, we simulate this knowledge by the *state transition PM*:

$$p\big(\vec{s}(t)|\vec{s}(t-1)\big) = \mathcal{N}\big(\tilde{\vec{s}}(t), \Sigma = \mathrm{diag}(\sigma)\big), \tag{7}$$

which is the model of temporal transitions of the state and is the same for attended and unattended voice ([1]). We use a normal distribution centered at the $\tilde{\vec{s}}(t)$, which is a state value extrapolated from the two previous time instances. We additionally make sure that the extrapolated value stays in the reasonable range for a given dimension (for example, $F0$ will never exceed 400 Hz).

2.4.3 Observation statistics

Humans can follow voices, even in the presence of complex background noises, that they have never heard before. It suggests that, in such complex acoustic conditions, the human brain is able to always relate the voice components in the noisy input to the abstract representation of a clean voice. In our modeling framework, it is simulated by the *observation statistics PM*: $p\big(\vec{o}(t)|\vec{s}(t)\big)$. This function provides the mapping between the observation space and the state space. To evaluate a probability of the incoming observation given a particular state value (in this case the one-dimensional state $\vec{s}(t) = F0(t)$) we first compute the probability of a single glimpse. We assume that auditory system can resolve up to 11 harmonics of $F0$, therefore our probabilistic model is defined as a mixture of 11 circular von-Mises distributions ($\mathscr{M}$):

$$p\big(P(t,c,m)|\vec{s}(t) = F0(t)\big) = \sum_{n=1}^{11} C_n \cdot \mathscr{M}\Big(\frac{P(t,c,m)}{\frac{1}{n*F0}} \cdot 2\pi, \kappa\Big), \tag{8}$$

where $P0_n = \frac{1}{n*F0}$ is the period corresponding to the $n-th$ harmonic and C_n is a normalising constant, which reduces the contribution of the probability associated with the higher harmonics of $F0$ (as the evidence collected based on them is more ambiguous). The probabilities of glimpses within one channel c are multiplied

$$p\big(G(t,c)|\vec{s}(t) = F0(t)\big) = \prod_{m=1}^{M(t,c)} p\big(P(t,c,m)|F0(t)\big) \tag{9}$$

and the support from different channels is accumulated with a sum, so that the final probability of a full observation (either foreground or background) given a state is defined as:

$$\begin{aligned} p\big(\vec{o}_{fg}(t)|\vec{s}(t) = F0(t)\big) &= \sum_c p\big(G(t,c)|F0(t)\big) \;\forall\; c \mid G(t,c) \in \vec{o}_{fg}(t) \\ p\big(\vec{o}_{bg}(t)|\vec{s}(t) = F0(t)\big) &= \sum_c p\big(G(t,c)|F0(t)\big) \;\forall\; c \mid G(t,c) \in \vec{o}_{bg}(t) \end{aligned} \tag{10}$$

EXPERIMENTAL EVALUATION

3.1 Stimuli generation

We follow the concept from a psychoacoustic study [24], which investigated the human ability to attentively track one of two competing voices. We use two synthetic voices with time-varying parameters represented as *hidden state trajectories*, which define the state of the system in each time instance (see Figure 3):

$$\mathscr{T}_{\vec{s}_{fg}} = \{\vec{s}_{fg}(t)|t = 0,...,T\} \text{ and } \mathscr{T}_{\vec{s}_{bg}} = \{\vec{s}_{bg}(t)|t = 0,...,T\}, \tag{11}$$

where $T = \frac{L}{0.02}$ and L is the length of the signal in seconds.

We generate each state trajectory as a random walk that evolves according to a predefined *state transition* PM. Based on state trajectories, we generate binaural acoustic signals. We use the Klatt formant synthesiser ([15]) for generating signals with varying fundamental frequency and formants, and TASCAR ([10]) to auralise the time-varying direction of arrival.

Each trial in our numerical experiment consists of two simultaneously active voices. We use the parallel particle

filters to track one dimension (fundamental frequency) of state vectors for each voice. One voice is always considered to be the foreground (attended) voice and one particle filter is initialized with the ground truth information about the initial state of the voice. In each trial we obtain estimated state trajectories:

$$\mathscr{T}_{\hat{s}_{fg}} = \{\hat{\vec{s}}_{fg}(t)|t = 0,...,T\} \text{ and } \mathscr{T}_{\hat{s}_{bg}} = \{\hat{\vec{s}}_{bg}(t)|t = 0,...,T\} \tag{12}$$

which we compare with the hidden state trajectories to evaluate the tracking performance.

3.2 Performance measures

We compare the hidden with the estimated state trajectories to evaluate tracking performance. We are specifically interested in how well our system can track the foreground (attended) voice in the presence of the second (background) voice. The following performance measures are used:

- RMSE - the root mean square error between the hidden $F0$ trajectory and estimated $F0$ trajectory.
- Hit rate ROC - We count a number of times one of the estimated trajectories is within a certain range r from the hidden trajectory of a target. For varying r, we compute the number of hits (when the estimate of foreground causes a hit) and the number of false alarms (when the estimate of the background causes a hit). We plot the ROC curve - hit rates against the false alarm rates. The area under this curve (AUC) quantifies detection performance.

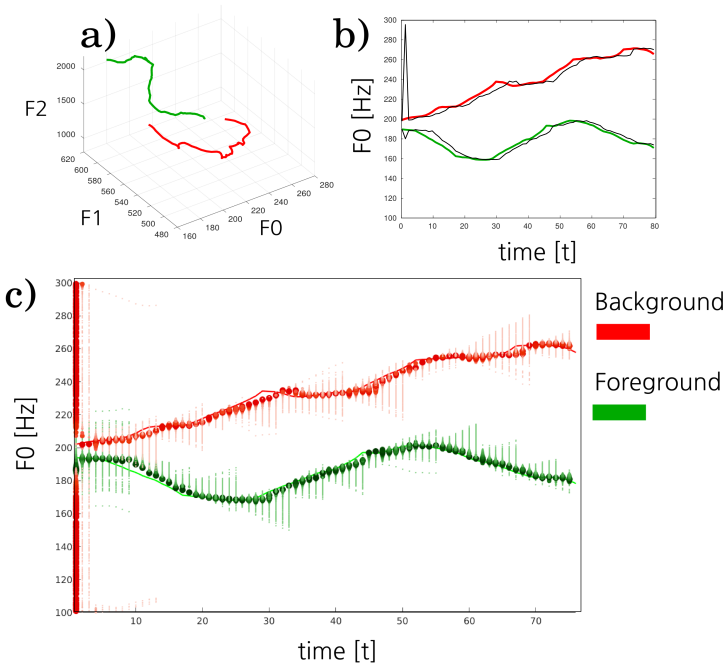

Figure 3. a) First three dimensions of a hidden state trajectory for foreground and background voice. b) Estimated state trajectory ($F0$) plotted together with the hidden $F0$ trajectory. c) Evolution of hypotheses in the particle filters while estimating foreground and background voice. Time axis of b) and c) is given in time instances t (sampling frequency FS_t=50 Hz)

RESULTS

We computed performance measures based on 100 experiment trials, each of which contains 1.5 s signal with competing voices that follow randomly generated parameter trajectories (different in each trial). We obtain the median RMSE of 3.9351 Hz with the interquartile range of 4.5617 Hz, which shows the ability of the auditory model to precisely track $F0$ of the target voice. The ROC curve (See Figure 4) also shows that the system estimates are good enough to use them for the detection of the attended voice in the presence of the second voice with a very high accuracy - the area under the curve (AUC) of 0.9793 and sensitivity index d' of 2.8850.

CONCLUSIONS

In the current study, we presented our computational model of attentive tracking in the human auditory system. The novelty of the model lies in revising the recently investigated aspects of the human auditory system and representing them in the computational world as a unique collection of methods. Our modeling framework consists of *glimpsed periodicity feature extraction* simulating sensory input, *glimpsed feature grouping* simulating grouping of sensory input into streams, and *recursive Bayesian estimation* simulating sequential inference in the auditory scene. The latter integrates *probabilistic models* simulating an abstract representation of a clean voice, as well as attention. This study serves as an overview of our modeling approach. The preliminary results prove the usability of the model for binary decision tasks. They will be a basis for designing a numerical simulation of the psychoacoustic study on attentive tracking of voices ([24]). In the near future, the model results will be compared with the psychoacoustic data. We also plan to further investigate the performance of the model for different noise types. Future work in this project will lead towards tracking multiple parameters of the voices, which will require expanding the feature and the state space and advancing the PMs, potentially with probabilistic machine learning approaches.

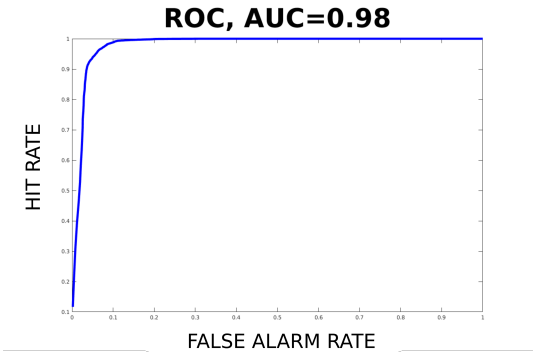

Figure 4. ROC curve of the target detection task.

ACKNOWLEDGEMENTS

Funded by the German Research Foundation DFG project number 352015383 - SFB 1330 and by the National Institutes of Health (NIH Grant1R01DC015429-01).

REFERENCES

[1] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan. *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.

[2] V. Best, C. R. Mason, J. Swaminathan, E. Roverud, and G. Kidd Jr. 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, 2017.

[3] A. Bregman. Auditory scene analysis: The perceptual organization of sound. cambridge, ma, us, 1990.

[4] N. Chater, J. B. Tenenbaum, and A. Yuille. Probabilistic models of cognition: Conceptual foundations, 2006.

[5] A. Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

[6] M. Cooke. A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3):1562–1573, 2006.

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

[8] N. Ding, X. Pan, C. Luo, N. Su, W. Zhang, and J. Zhang. Attention is required for knowledge-based sequential grouping: insights from the integration of syllables into words. *Journal of Neuroscience*, 38(5):1178–1188, 2018.

[9] M. Elhilali. Bayesian inference in auditory scenes. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2792–2795. IEEE, 2013.

[10] G. Grimm, J. Luberadzka, and V. Hohmann. Virtual acoustic environments for comprehensive evaluation of model-based hearing devices. *International journal of audiology*, 57(sup3):S112–S117, 2018.

[11] M. Heilbron and M. Chait. Great expectations: is there evidence for predictive coding in auditory cortex? *Neuroscience*, 2017.

[12] A. Josupeit and V. Hohmann. 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, 2017.

[13] A. Josupeit, N. Kopčo, and V. Hohmann. 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, 2016.

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

[15] D. H. Klatt. Software for a cascade/parallel formant synthesizer. *the Journal of the Acoustical Society of America*, 67(3):971–995, 1980.

[16] J. Nix and V. Hohmann. 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, 2007.

[17] S. Popham, D. Boebinger, D. P. Ellis, H. Kawahara, and J. H. McDermott. Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nature communications*, 9(1):2122, 2018.

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

[19] E. Schoenmaker and S. van de Par. Intelligibility for binaural speech with discarded low-snr speech components. In *Physiology, psychoacoustics and cognition in normal and impaired hearing*, pages 73–81. Springer, Cham, 2016.

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

[21] C. Spille, B. Meyer, M. Dietz, and V. Hohmann. Binaural scene analysis with multidimensional statistical filters. In *The technology of binaural listening*, pages 145–170. Springer, 2013.

[22] L. S. van Noorden. Temporal coherence in the perception of tone sequences. *PhD thesis, Eindhoven University of Technology*, 1975.

[23] L. Whiteley and M. Sahani. Attention in a bayesian framework. *Frontiers in human neuroscience*, 6:100, 2012.

[24] K. J. Woods and J. H. McDermott. Attentive tracking of sound sources. *Current Biology*, 25(17):2238–2246, 2015.