Combining glimpsed auditory features and machine learning 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

Motivation

Human listeners can follow a desired speaker even in complex auditory scenes. This is due to the ability to segregate the incoming mixture of sounds into streams of information that can be associated with objects in the scene [1]. On one hand stream formation is a result of grouping the simultaneously incoming sound components by the similarities in their general properties like harmonic structure, time onset or spatial direction [3]. A second mechanism is related to the ability to form and maintain the streams of information sequentially over time.
Interaction of those mechanisms is especially important for speech perception in challenging acoustic conditions with many simultaneously active voices.
Previous studies suggest that in such conditions, the auditory system predominantly uses sparse, speaker-related bits of robust information - *auditory glimpses*, which provide reliable cues for simultaneous grouping([4]-[9]). Following a speaker requires integration of this sparse information over time. Sequential grouping is known to be supported by perceptual distance between the components of the individual streams [2], but recent studies also explore the role of attention and statistical inference in the stream formation ([14]-[19]).

This study presents a computational model of attentive tracking of voices, which takes these aspects into account and thus contributes to understanding the speech perception in complex auditory scenes. The novelty of this approach lies in combining sparse glimpsed features ([4, 6]) with Bayesian sequential estimation ([11]-[13]).

Modeling framework

In particular, we model an auditory scene consisiting of two simultaneously active voices [14]. The scene analysis task - attentively following one of these voices - is represented as tracking the *high-level* parameters defining the voice such as fundamental and formant frequencies.

The proposed modeling framework includes three main stages (see Figure 1): *a) Signal generation* stage, where the binaural signal containing mixture of two competing voices is generated, *b) Feature extraction* stage, which transforms the waveform into glimpsed periodicity features and *c) State estimation* stage, where the glimpsed fatures are decomposed and forwarded to two parallel particle filters, which sequentially estimate the parameters of the voices using *d) prior knowledge* on dynamical changes of voice *high-level* parameters and their statistical relation to the glimpsed periodicity features. The stages of the modeling framework will be discussed in the next sections.

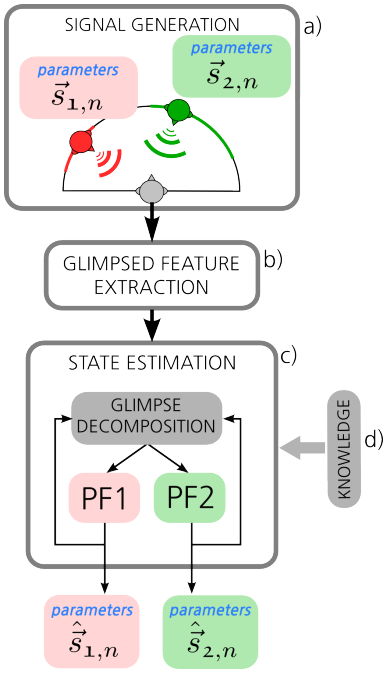

Figure 1: Modeling framework with three main stages: *Signal generation*, *Feature extraction*, *State estimation*.

Signal generation

We follow the concept from a psychoacoustic study [14], which investigated the human ability to attentively track one of two competing voices. We use two synthetic voices with time varying *high-level* parameters represented as *hidden state trajectories* (see Figure 2):

$$\mathcal{T}_{\vec{s}_1} = \{\vec{s}_{1,0}, \vec{s}_{1,1}, ..., \vec{s}_{1,N}\}$$
$$\mathcal{T}_{\vec{s}_2} = \{\vec{s}_{2,0}, \vec{s}_{2,1}, ..., \vec{s}_{2,N}\},$$

which define the state of the system in each time instance:

$$\vec{s}_{1,n} = \begin{pmatrix} F0_{1,n} \\ F1_{1,n} \\ F2_{1,n} \\ \alpha_{1,n} \end{pmatrix} \vec{s}_{2,n} = \begin{pmatrix} F0_{2,n} \\ F1_{2,n} \\ F2_{2,n} \\ \alpha_{2,n} \end{pmatrix}.$$

The voices change their parameters - fundamental frequency $F0$, first two formants $F1$, $F2$, and direction of arrival α - over time, but there are no constant dissimilarities between them, that could support the stream formation (e.g. voice timbre).

We generate each state trajectory as a random walk (see Figure 2) that evolves according to a predefined *state transition probablity density function (PDF)* (see Section *State estimation*).
Based on state trajectories, we generate binaural acoustic signals. We use the Klatt formant synthesiser [26] for generating signals with varying fundamental frequency and formants, and TASCAR [24] to auralise the time varying direction of arrival. We simulate a human listener, whose ears receive the mixture of sounds originating from both voices. The binaural signal is the input to the feature extraction stage.

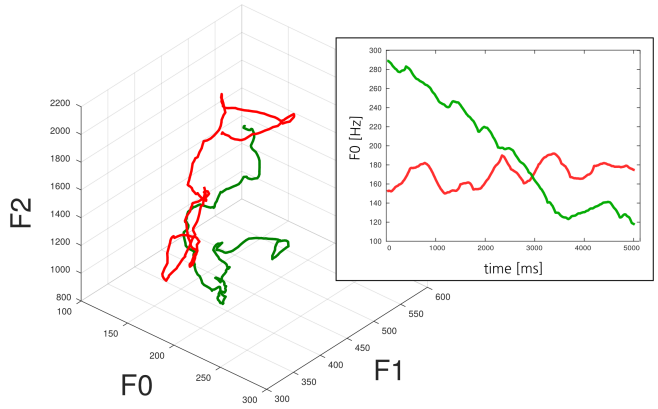

Figure 2: Example of a multidimensional hidden state trajectory. The left plot presents the first three dimensions (fundamental frequency $F0$ and the first two formants $F1$, $F2$) of two state trajectories. All dimension of the state trajectory evolve freely according to the *state transition PDF* (random walk). The right plot presents one dimension ($F0$) of the same trajectories over time.

Glimpsed feature extraction

The feature extraction stage transforms the input waveform into a feature space that carries more meaning in terms of the scene analysis task - following one of two competing voices.
Several studies show that for solving such tasks, the auditory system uses mainly the sparse, salient, speaker-related pieces of undistrupted information ([10, 8, 9, 6]), which we refer to as *auditory glimpses*. Mentioned studies usually compare the speech intelligibility for signals resynthesized from partial information against the speech intelligibility of the unprocessed mixture. For selecting the speaker-related information they use the perfect knowledge about the signal and masker energies.
In our model, we use the approach developed in [5, 4, 6], which is capable of blindly extracting speech glimpses from the mixture of signals. The method focuses on detecting the periodicity, as one dimension, in which the salience of speech is reflected ([7]). Salient periodicity in the signal is thus treated as robust speaker-related information that originates from a single source, which we refer to as *periodicity glimpses*.
The glimpsed feature extraction is depicted in Figure 3. Each channel of a binaural input signal is passed through the auditory preprocessing stage, which, after gammatone band-pass filtering, half-wave rectification, 5th-order lowpass-filtering at 770 Hz and 40 Hz highpass-filtering, outputs 23 time signals. Next, the periodicity analysis, which yields a normalized synchrogram ([25]), is performed on the time signals in each frequency band. Synchrograms represent the proportion of the periodic energy of a signal with fundamental period P in relation to the overall energy, for every time instance n , and a given range of P. Values in the normalized synchrogram, that exceed a certain threshold (e.g. > 0.9) indicate the dominant fundamental period $P0$ and are considered salient periodicity glimpses.

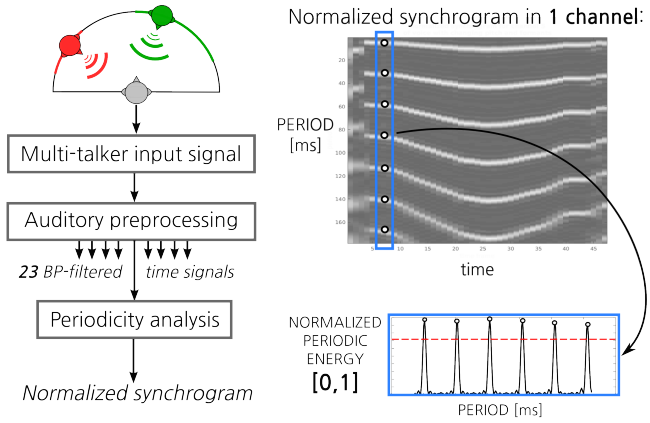

Figure 3: Extraction of periodicity glimpses.

In general, glimpses are found at the multiples of $P0$ corresponding to one of the harmonics of the fundamental frequency (see Figure 4 a),b)). Which harmonic of $F0$ will be mirrored in the glimpses depends both on the formants $F1$ and $F2$ and on the frequency channel. Glimpses extracted from a mixture of two vowels are a combination of glimpses from the individual sounds - in each channel they will exclusively originate from either of two voices (see Figure 4 c)). It is consistent with the assumption, that a glimpse always represents information related to only one voice.

We extract the periodicity glimpses from a binaural mixture signal every 20 ms and use them as an observation vector $\vec{o}_n$, which is an input to the state estimation stage at time n (Figure 5 a)).

State estimation

We model attentive tracking as sequential inference of *high-level* parameters represented in the hidden state vectors $\vec{s}_1$ and $\vec{s}_2$. In this study, we track a single dimension - $F0$ - of the multidimensional hidden state.
Numerous studies mention Bayesian estimation as a plausible model of the inference in the human brain, both in a general view on cognition ([18]-[20]), as well as in the context of auditory perception ([21]-[23]).
We follow this idea in our modeling framework by using *particle filters* [13]- a sampling solution to the sequential Bayesian estimation problem, which has already been successfully used in the context of speech tracking ([11, 12]).

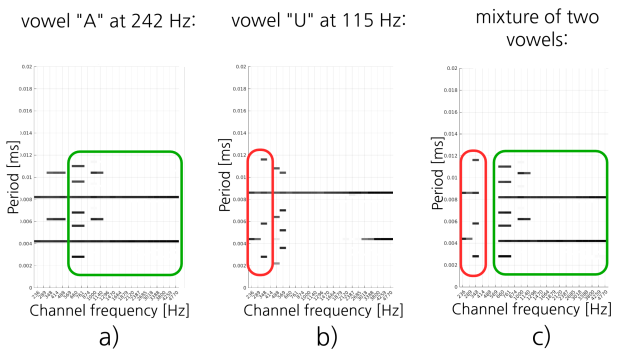

Figure 4: Examples of glimpsed periodicity patterns. All plots present glimpsed periodicity features across all 23 frequency channels. The first two plots from the left show the glimpses extracted individually for two different vowels, at different $F0$s.The plot on the right side shows the glimpses extracted from a mixture of those two vowels - glimpses in different channels originate either from the first or from the second vowel.

Tracking in the Bayesian framework requires *prior knowledge* (Figure 1 d)) in the form of the following statistical models:

1. *initial state PDF* $p(\vec{s}_0)$, used to initialize the hypotheses in the particle filter.
2. *state transition PDF* $p(\vec{s}_n|\vec{s}_{n-1})$, mentioned already in the *State generation* section, which describes the temporal evolution of the parameters identifying the voices.
3. *observation statistics PDF* $p(\vec{o}_n|\vec{s}_n)$, which describes the mapping between state and observation.

We use the same single-source statistical models for both voices. $p(\vec{s}_0)$ is defined as a uniform distribution in the range of possible values of voice parameters (for example $F0 \in [100, 300]$ Hz). $p(\vec{s}_n|\vec{s}_{n-1})$ is implemented as a probabilistic gaussian motion model [28]. $p(\vec{o}_n|\vec{s}_n)$ is a model inspired by [27] - we analytically compute the expected periodicity values for a given fundamental frequency and its harmonics (up to the 11th harmonic) and generate a glimpse probability density with Gaussian modes around those values.

For tracking $F0$ of two competing voices we use two particle filters - $PF1$ and $PF2$ - in parallel (see Figure 1c)). The state estimation stage begins with the *glimpse decomposition* (Figure 5b)) - the incoming glimpsed observation $\vec{o}_n$ is split into observations $\vec{o}_{1,n}$, $\vec{o}_{2,n}$ for the two competing voices. In particular, each frequency channel of glimpses is assigned to the voice that most likely generated it. The likelihood is computed by comparing $\vec{o}_n$ with the previous state estimate for both voices via the *observation statistics PDF* $p(\vec{o}_n|\vec{s}_n)$. $PF1$ receives observation $\vec{o}_{1,n}$ and $PF2$ $\vec{o}_{2,n}$, respectively.

A single particle filter is an iterative algorithm that alternates between the following processing steps (see Figure 5):

Initialization: A set of initial hypotheses about the state of the system - particles - is created by drawing values from the $p(\vec{s}_0)$ (Figure 5 c))

Prediction: New hypotheses are predicted based on previous hypotheses by drawing values from the $p(\vec{s}_n|\vec{s}_{n-1})$ (Figure 5 d)).

As a result of this step, the temporal evolution of the state is reflected in the hypotheses set.

Update: All current hypotheses are compared with the available observation $\vec{o}_{1,n}$ (selected in the glimpse decomposition step (Figure 5 f)), by evaluating $p(\vec{o}_n|\vec{s}_n)$ Figure 5e)).

Estimation: The hypothesis that gets the highest support is the estimated state$\hat{\vec{s}}_{1,n}$.(Figure 5 f)).

Resampling: Before the next iteration, the most likely hypotheses are duplicated and the unlikely hypotheses are eliminated from the hypotheses set. This way the focus of the particle filter is shifted to the region of interest (Figure 5 g)). The resampling step is done only if the particle filter has received reliable information in the current iteration. Otherwise, the search space will broaden with time (see Figure 6).

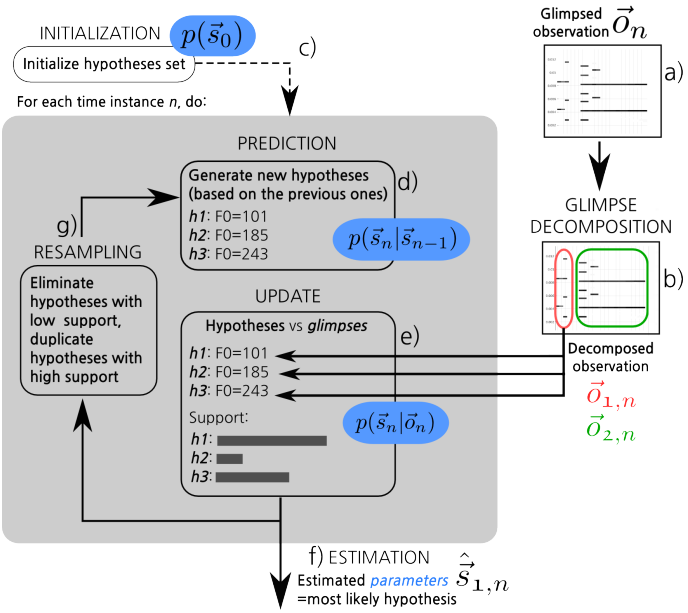

Figure 5: Processing steps in one of two parallel particle filters in the modeling framework.

After iterating over all time frames, the particle filters yield the estimated state trajectories

$$\mathcal{T}_{\hat{s}_1} = \{\hat{\vec{s}}_{1,0}, \hat{\vec{s}}_{1,1}, ..., \hat{\vec{s}}_{1,N}\}$$

$$\mathcal{T}_{\hat{s}_2} = \{\hat{\vec{s}}_{2,0}, \hat{\vec{s}}_{2,1}, ..., \hat{\vec{s}}_{2,N}\},$$

which can be compared with the hidden state trajectories in order to assess the tracking performance (see Figure 6).

Conclusions and outlook

We proposed a modeling framework for attentive tracking of voices, which combines periodicity glimpses and

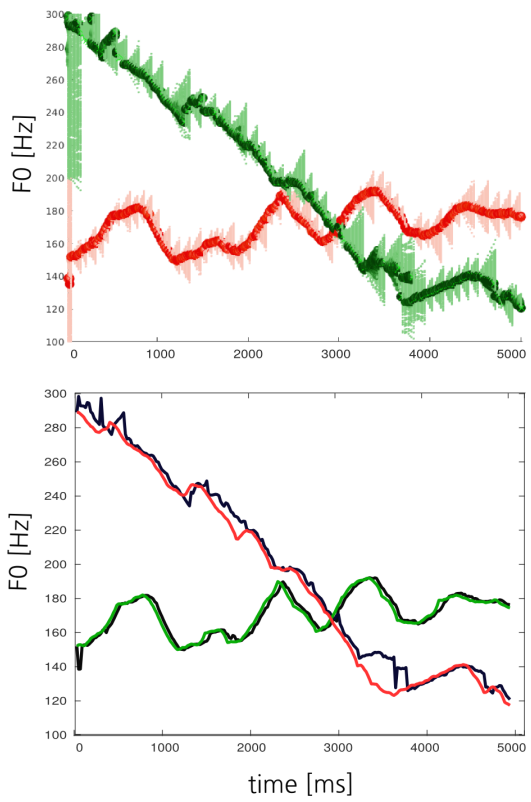

Figure 6: Example of $F0$ estimation. The upper plot shows how the hypotheses of two parallel particle filters evolve, while tracking $F0$ of two competing voices from periodicitity glimpses. The right plot shows the preliminary results - hidden $F0$ trajectories together with the estimated $F0$. The bottom plot shows only one dimension of the 4-dimensional hidden state trajectories (shown also in Figure 2) that were used to generate the voices contributing to the mixture)

particle filtering in a novel way. It is exemplarily shown (see Figure 6), that the framework is capable of precise estimation of the fundamental frequency of two synthetic competing voices from glimpsed periodicity features. We expect, that the robustness of the salience-based features will allow for voice tracking also in the case of real speech signals with various background noises. In the near future we plan to compare the model predictions with the human data collected in the psychoacoustic study on attentive tracking [14].

The modeling framework will be extended to track all *high-level* parameters included in the state vector $(F0, F1, F2, \alpha)$. This will require to include the information about the spectral energy distribution and interaural time and level differences in the feature space.

Another essential challenge lies in designing the *prior knowledge* probability models, in particular the *observation statistics PDF*, which has to be suitable for sparse data. We plan to further investigate both the analytical models and machine learning approaches.

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

References

[1] Bregman AS. Auditory scene analysis: The perceptual organization of sound. MIT press; 1994.

[2] van Noorden LS. Temporal coherence in the perception of tone sequences. PhD thesis, Eindhoven University of Technology. 1975.

[3] Darwin CJ. Listening to speech in the presence of other sounds. Philosophical Transactions of the Royal Society B: Biological Sciences. 2007 Sep 7;363(1493):1011-21.

[4] Josupeit A, Hohmann V. Modeling speech localization, talker identification, and word recognition in a multi-talker setting. The Journal of the Acoustical Society of America. 2017 Jul 6;142(1):35-54.

[5] Josupeit A, Kopco N, Hohmann V. Modeling of speech localization in a multi-talker mixture using periodicity and energy-based auditory features. The Journal of the Acoustical Society of America. 2016 May 20;139(5):2911-23.

[6] Josupeit A, Schoenmaker E, van de Par S, Hohmann V. Sparse periodicity-based auditory features explain human performance in a spatial multitalker auditory scene analysis task. European Journal of Neuroscience. 2018 Jun 1.

[7] Popham, Sara, et al. "Inharmonic speech reveals the role of harmonicity in the cocktail party problem." *Nature communications* 9.1 (2018): 2122.

[8] Schoenmaker E, van de Par S. Intelligibility for binaural speech with discarded low-SNR speech components. InPhysiology, psychoacoustics and cognition in normal and impaired hearing 2016 (pp. 73-81). Springer, Cham.

[9] Cooke M. A glimpsing model of speech perception in noise. The Journal of the Acoustical Society of America. 2006 Mar;119(3):1562-73.

[10] Best V, Mason CR, Swaminathan J, Roverud E, Kidd Jr G. 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. 2017 Jan 9;141(1):81-91.

[11] Nix J, Hohmann V. Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering. IEEE transactions on audio, speech, and language processing. 2007 Mar;15(3):995-1008.

[12] Spille C, Meyer BT, Dietz M, Hohmann V. Binaural scene analysis with multidimensional statistical filters. InThe technology of binaural listening 2013 (pp. 145-170). Springer, Berlin, Heidelberg.

[13] Arulampalam MS, Maskell S, Gordon N, Clapp T. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on signal processing. 2002 Feb;50(2):174-88.

[14] Woods KJ, McDermott JH. Attentive tracking of sound sources. Current Biology. 2015 Aug 31;25(17):2238-46.

[15] Shamma SA, Elhilali M, Micheyl C. Temporal coherence and attention in auditory scene analysis. Trends in neurosciences. 2011 Mar 1;34(3):114-23.

[16] Whiteley L, Sahani M. Attention in a Bayesian framework. Frontiers in human neuroscience. 2012 Jun 14;6:100.

[17] Carlyon RP, Cusack R, Foxton JM, Robertson IH. Effects of attention and unilateral neglect on auditory stream segregation. Journal of Experimental Psychology: Human Perception and Performance. 2001 Feb;27(1):115.

[18] Pouget A, Beck JM, Ma WJ, Latham PE. Probabilistic brains: knowns and unknowns. Nature neuroscience. 2013 Sep;16(9):1170.

[19] Chater N, Tenenbaum JB, Yuille A. Probabilistic models of cognition: Conceptual foundations.

[20] Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and brain sciences. 2013 Jun;36(3):181-204.

[21] Schröger E, Marzecova A, SanMiguel I. Attention and prediction in human audition: a lesson from cognitive psychophysiology. European Journal of Neuroscience. 2015 Mar;41(5):641-64.

[22] Heilbron M, Chait M. Great expectations: is there evidence for predictive coding in auditory cortex?. Neuroscience. 2017 Aug 4.

[23] Elhilali M. Bayesian inference in auditory scenes. In2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2013 Jul 3 (pp. 2792-2795). IEEE.

[24] Grimm G, Luberadzka J, Herzke T, Hohmann V. Toolbox for acoustic scene creation and rendering (TASCAR)-Render methods and research applications. InProceedings of the Linux Audio Conference 2015 (pp. 9-12).

[25] Hohmann V. Verfahren zur Extraktion periodischer Signalkomponenten und Vorrichtung hierzu.

[26] Klatt DH. Software for a cascade/parallel formant synthesizer. the Journal of the Acoustical Society of America. 1980 Mar;67(3):971-95.

[27] Schroeder MR. Period histogram and product spectrum: New methods for fundamental-frequency measurement. The Journal of the Acoustical Society of America. 1968 Apr;43(4):829-34.

[28] Bar-Shalom Y, Li XR, Kirubarajan T. Estimation For Kinematic Models in. Estimation with Applications to Tracking and Navigation. 2001:72-7.