Modelling Primitive Streaming of Simple Tone Sequences Through Factorisation of Modulation Pattern Tensors

Tom Barker[1]*, Hugo Van hamme*[2]*, Tuomas Virtanen*[1]

[1]Tampere University of Technology, Finland, [2]KU Leuven, Belgium

Thomas.Barker@tut.fi, Hugo.Vanhamme@esat.kuleuven.be, Tuomas.Virtanen@tut.fi

Abstract

We present a novel method for determining how the perceptual organisation of simple alternating tone sequences is likely to occur in human listeners. By training a tensor model representation using features which incorporate both low-frequency modulation rate and phase, a set of components is learned. Test patterns are modelled using these learned components, and the sum of component activations is used to predict either an 'integrated' or 'segregated' auditory stream percept. We find that for the basic streaming paradigm tested, our proposed model and method is able to correctly predict either segregation or integration in the majority of cases.

Index Terms: auditory modelling, stream segregation, tensor factorisation

1. Introduction

Blind computational speech source separation is a difficult and unsolved research problem. Although fairly effective techniques for separating speech from background noise exist [1, 2], these generally require an explicit model of either speech or noise, and performance of blind methods [3], is not yet close to that of the human auditory system. It therefore makes sense to look to the human auditory system for inspiration in terms of features and mechanisms which can be used to produce successful sound-source separation algorithms.

The human auditory system is extremely good at grouping sensory inputs which occur in the environment into perceptual objects which relate to their true source. The formation of these auditory 'objects' [4], a process commonly referred to as auditory scene analysis (ASA) can take place at one (or more) of a number levels within the auditory system; indeed it is proposed in [5] that 'primitive' processes in ASA occur automatically, and are data-driven, whilst 'schema-based' processes are more reliant upon the listener's attention and cognitive input, hence rely on higher level cognitive functions. By modelling the primitive processes, we can gain insight into how various features of a signal affect the mechanisms of perceptual grouping by the auditory system, although many behaviours are already fairly well documented [5, 6], the mechanisms behind such groupings are still subject to much investigation.

In van Noorden's classic experiments [7], it was shown that for an alternating sequence of two pure tones, A and B, the formation of the percept of either a single stream or two separate streams depends upon tone repetition time and frequency difference, Δf (see Figure 1). We train a model with data generated from such

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 290000 and Academy of Finland grant number 258708

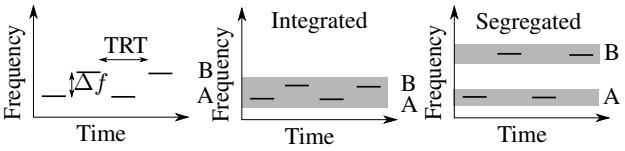

Figure 1: An example of possible perceptual organisation for pure-tone sequences. Depending on both the frequency separation, Δf, and tone-repetition-time (TRT), the percept of either a single continuous sequence of alternating tones (integration, middle axis), or two separate sequences of tones (segregation, right axis) is produced.

results, and investigate the use of such a signal representation to estimate the percept to be generated by novel stimuli.

To date, multiple attempts have been made to model auditory system stream formation behaviours. It is proposed and modelled in [8] that temporal coherence between neuron responses from the primary auditory cortex is one of the key features in stream formation, and strength of either an integrated or segregated percept is estimated via an eigenvalue ratio of the result of temporal coherence analysis. This work is extended in [9], where a biologically inspired auditory front-end, [10] is used to first pre-process the stimulus, whilst using the coherence analysis proposed in [8] as a back-end. Earlier models such as in [11] rely on frequency separation between tones presented, and grouping is based on activations of strongly overlapping auditory 'channels', but do not account for the integration of tones with common onset and offset (simultaneous grouping).

We propose the use of a novel feature representation, and supervised tensor factorisation method to learn features from training data. Supervised factorisation methods have been used to produce good audio source separation [1], but have not taken account of underlying representations within the auditory system. Our approach inherently considers the phase of components within a sequence, as in [8, 9], but instead makes a prediction as to an 'integrated' or 'segregated' percept based on the number and strength of learned component activations rather than eigenvalue ratios.

2. Model Description

The proposed model represents tone sequences as a sum of components which are learned through factorisation of a number of training examples. Each training example is modelled over 3 dimensions, with different auditory model-based frequency bands being represented in terms of modulation frequency and phase (see Section 3). A 4th dimension holds the weights of the components for each example and during testing, these weights are used to make a prediction as to whether a sequence of tones will be

perceived as integrated or segregated. Components in the first 3 dimensions are first learned during training, and then remain fixed for the test phase, whilst a new set of weights is learned for each example under test.

Lower weights suggest the stimulus requires fewer components to accurately approximate it, and would produce an integrated percept, whilst higher weights are more complex and produce a segregated percept. The model's capacity to make accurate predictions on novel data demonstrates the ability of the proposed data representation to capture relevant features for the formation of auditory streams.

3. Audio Representation

Audio is represented in terms of both modulation frequencies and phases across distinct auditory channels. A similar representation has been used in previous work, where phase independent modulation patterns have been used to group components originating from the same source in both unsupervised and semi-supervised manners [3, 12]. Here, the incorporation of phase is important since it allows the capture of co-occurrence or discrepancies between temporal onset times, which will allow useful extension to the model for more complex tone sequences. Phase information is encoded by quantising the phase component of the discrete Fourier transform (DFT) of modulation envelopes.

To represent audio, the following method is used: A training or test example, is normalised based on its RMS power, filtered with a constant-phase gammatone filterbank [13], with $R = 20$-channels, then halfwave-rectified and lowpass filtered (LPF) with a single-pole filter with 30Hz cutoff to produce a modulation envelope (ME) representation for each filterbank sub-band. This is similar to the front-end processing in many auditory models [14]. Since high frequencies are removed by the LPF operation, downsampling to 60Hz is performed to reduce data complexity. Each auditory channel is processed with a sliding window DFT; due to the repetitive nature of the stimuli considered in this paper, each DFT output frame should contain similar magnitude information which varies only in phase. Different phase shifts of the same DFT result are therefore considered equivalent within the model framework. It should be noted that the representation could accomodate temporally dynamic stimuli by considering multiple frames of audio input across the 4th dimension. A single time frame of sufficient length is able to characterise a periodic signal, and so each training example is represented by such. We use a single window of 0.5 seconds (30 samples) for each example. Phase components of the DFT result exist over a continuous range of $0 - 2\pi$. This range is quantised to 30 bins, and the magnitude of each DFT frequency bin is assigned to the appropriate quantised phase bin. The processing described is represented in Figure 2.

The phase-quantised representation of such a complex number used in our model, $\mathcal{T}$, is produced by considering the phase Z^{ϕ} in each auditory channel r, DFT bin q, and for each training example h is quantised such that:

$$\mathcal{T}_{r,q,n,h} = \begin{cases} |Z|_{r,q,h} & \text{if } (\frac{2\pi(n-1)}{V}) \leq Z^{\phi}_{r,q,h} < (\frac{2\pi n}{V}) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where V is the number of quantisation bins.

4. Rotated Tensor Factorisation

Training data is factorised in order to identify redundant patterns, which form the generalised model, the components of which are used for the representation of new data. The aim is to learn a set of components which best approximate the features which are commonly occurring in stimuli from which the tensor model is generated, with the aim that they will also fit new data well.

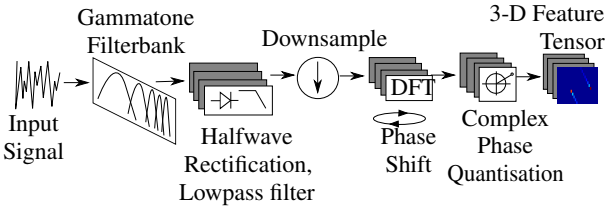

Figure 2: Schematic overview of the process to produce a 3-Dimensional Tensor representation of a single audio example. The phase rotation operation is applied prior to quantisation in each case.

In conventional shifted-factorisation approaches [15], shifts are applied to one or more dimensions and to one or more of the factor matrices in an attempt to better fit the data. We want to make no distinction between the same (periodic) signal sampled at different time points or shifts. The magnitude of the DFT of such a signal would be identical regardless of sample time, but phase would not. We therefore treat phase shifted versions of each training example h as identical, and modify the phase accordingly to minimise the Kullback-Leibler (KL) divergence between the training data tensor $\mathcal{T}$ and its approximation. A phase shift equivalent to a single sample in the (downsampled) time domain delay is applied by multiplication of the complex DFT result with a complex exponential, similarly to as in [16], which removes the need for an exhaustive search over all integer shifts of the tensor. Furthermore, a simple shift in the tensor domain would produce an equal phase shift for each frequency bin, which would be equivalent to a different time shift for each modulation frequency present in the signal, and thus not accurately represent time-shifted samples of the same stimulus.

A phase shifted version of complex data Z in the DFT domain, Z^{τ}, with a shift of τ time-domain samples, can be produced by multiplication with a complex exponential prior to the quantisation operation,

$$Z^{\tau}_{r,q,h} = Z_{r,q,h} e^{j\frac{2\pi q \tau}{N}} \tag{2}$$

where r is the filterbank channel, q is the DFT result sample index, h is the training or test example and N is the DFT length.

4.1. Tensor Model Updates

The 4-dimensional training data tensor representation is modelled by the summation of components learned through tensor factorisation. The training tensor $\mathcal{T}$, of dimensions $R \times Q \times N \times H$ (number of auditory channels $\times$ DFT bins $\times$ phase quantisation bins $\times$ training examples) is the sum of K components, each of which is a product of the factors contained in matrices $\mathbf{C}$ (size $R \times K$), $\mathbf{M}$ (size $Q \times K$), $\mathbf{P}$ (size $N \times K$) and $\mathbf{D}$ (size $H \times K$). Each of the K columns of the factor matrices describes a component of the tensor decomposition. The model for $\mathcal{T}$ is denoted $\hat{\mathcal{T}}$ such that:

$$\mathcal{T}_{r,q,n,h} \approx \hat{\mathcal{T}}_{r,q,n,h} = \sum_{k=1}^{K} \mathbf{C}_{r,k}\mathbf{M}_{q,k}\mathbf{P}_{n,k}\mathbf{D}_{h,k} \tag{3}$$

4.2. Update Equations

The components in matrices $\mathbf{C}$, $\mathbf{M}$, $\mathbf{P}$ and $\mathbf{D}$ are learned through update equations which minimise the KL-divergence via an alternating least squares approach, whilst enforcing sparsity over the $\mathbf{D}$ dimension. Model parameters are learned by minimising the Kullback-Leibler (KL) divergence D between $\mathcal{T}$ and $\hat{\mathcal{T}}$:

$$D(\mathcal{T}||\hat{\mathcal{T}}) = \sum_{r,q,n,h} \mathcal{T}_{r,q,n,h} \log \frac{\mathcal{T}_{r,q,n,h}}{\hat{\mathcal{T}}_{r,q,n,h}} - \mathcal{T}_{r,q,n,h} + \hat{\mathcal{T}}_{r,q,n,h} \tag{4}$$

whilst imposing a sparsity penalty on the matrix $\mathbf{D}$. The row-normalised L1-norm for $\mathbf{D}$ is used as a sparsity cost and is defined as:

$$\sum_{i=1}^{K} \frac{\sum_{j=1}^{H} \mathbf{D}_{i,j}}{\sqrt{\sum_{j'=1}^{H} \mathbf{D}_{i,j'}^2}} \tag{5}$$

where we also vary the weights of this for each row with the vector $\lambda = [0, 1...H]$ which holds the sparsity coefficient for each row in $\mathbf{D}$.

The multi-dimensional estimation problem is reduced to a set of matrix factorisation problems through matricisation and solving over each mode of the tensor $\mathcal{T}$ successively. Following each set of updates over all dimensions, $\mathcal{T}$ is rotated over all possible integer time domain sample shifts τ to again minimise KL-divergence. The mode-n matricised (unfolded) version of a tensor $\mathcal{T}$ is denoted as $\mathbf{T}_{(n)}$ as in [17]. We define $\mathcal{Q}$ as $\mathcal{T}/\hat{\mathcal{T}}$ and its matricised representation as $\mathbf{Q}_{(n)}$ and $\mathbf{1}_{(n)}$ is a matrix of 1s with the same dimensionality as $\mathbf{Q}_{(n)}$. The Khatri-Rao product (see also [17]) is denoted by the $\odot$ operator, and element-wise multiplication by $\otimes$.

The update equation which minimises KL-divergence between $\mathcal{T}$ and $\hat{\mathcal{T}}$ for $\mathbf{C}$ is:

$$\mathbf{C} \leftarrow \mathbf{C} \otimes \frac{\mathbf{Q}_{(1)}[\mathbf{M} \odot \mathbf{P} \odot \mathbf{D}]^T}{\mathbf{1}_{(1)}[\mathbf{M} \odot \mathbf{P} \odot \mathbf{D}]^T}. \tag{6}$$

Similarly, $\mathbf{M}$ is updated via:

$$\mathbf{M} \leftarrow \mathbf{M} \otimes \frac{\mathbf{Q}_{(2)}[\mathbf{C} \odot \mathbf{P} \odot \mathbf{D}]^T}{\mathbf{1}_{(2)}[\mathbf{C} \odot \mathbf{P} \odot \mathbf{D}]^T} \tag{7}$$

and $\mathbf{P}$ by:

$$\mathbf{P} \leftarrow \mathbf{P} \otimes \frac{\mathbf{Q}_{(3)}[\mathbf{C} \odot \mathbf{M} \odot \mathbf{D}]^T}{\mathbf{1}_{(3)}[\mathbf{C} \odot \mathbf{M} \odot \mathbf{D}]^T} \tag{8}$$

whereas the update rule for $\mathbf{D}$ also minimises the cost function defined in Equation 5 and thus becomes:

$$\mathbf{D} \leftarrow \mathbf{D} \otimes \left(\frac{\mathbf{Q}_{(4)}[\mathbf{C} \odot \mathbf{M} \odot \mathbf{P}]^T + \nabla c_s^-(\mathbf{D})}{\mathbf{1}_{(4)}[\mathbf{C} \odot \mathbf{M} \odot \mathbf{P}]^T + \nabla c_s^+(\mathbf{D})} \right) \tag{9}$$

where

$$\nabla c_s^-(\mathbf{D}_{h,t}) = \lambda_h \frac{d_{h,t}\sqrt{K}\sum_{i=1}^{K} d_{h,i}}{(\sum_{i'=1}^{K} d_{h,i'}^2)^{3/2}} \tag{10}$$

and

$$\nabla c_s^+(\mathbf{D}_{h,t}) = \lambda_h \frac{1}{\sqrt{\frac{1}{K}\sum_{i=1}^{K} d_{h,i}^2}}. \tag{11}$$

Each component k in $\mathbf{C}$, $\mathbf{M}$ and $\mathbf{P}$ is also L1 normalised after each application of the corresponding update equation.

4.3. Training the Model

A number of training examples are generated, based on audio data which is regarded to be perceived as unambiguously integrated or segregated. Each example is formed as per the process in Figure 2, and the H examples are used to form the 4-dimensional tensor which is factorised, as in Figure 3. The factorisation process is performed by initialising $\mathbf{C}$, $\mathbf{M}$, $\mathbf{P}$ and $\mathbf{D}$ with random non-negative values and then repeating the following until convergence (negligible change in the overall cost function):

1. Update $\mathbf{C}$, $\mathbf{M}$, $\mathbf{P}$, $\mathbf{D}$ in turn, using the rules defined in Equations (6 - 9)

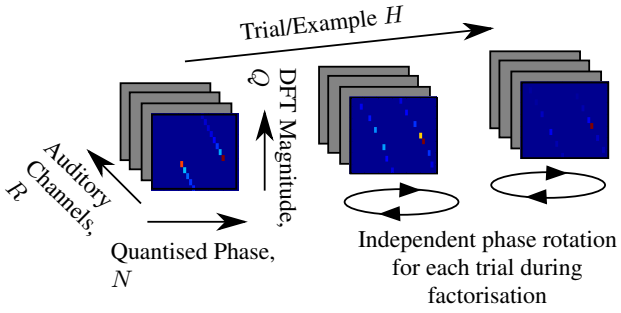

Figure 3: The 4 dimensions of the training tensor which is decomposed through non-negative PARAFAC factorisation.

2. For each h, rotate $\mathcal{T}$ through all possible integer phase shifts.
3. Set each h in $\mathcal{T}$ to the rotation which minimises KL divergence defined in Equation 4 for the next round of updates.

4.4. Using training data on novel stimuli

Following training of the model, new examples can be modelled using the learned components contained in matrices $\mathbf{C}$, $\mathbf{M}$ and $\mathbf{P}$. These components are used to model the test examples, and depending upon the weights of activations, an estimation as to the likelihood of either an integrated or segregated percept is produced. We simplify the problem domain by considering only integrated or segregated percepts, and do not attempt to model or classify ambiguous percepts.

Each test example is formed into a tensor, $\mathcal{X}$, of dimensions $R \times Q \times N \times 1$. For each test example, a vector, $\mathbf{D}'$ of dimension $1 \times K$ is initialised with positive random values, and used in the approximation $\hat{\mathcal{X}}$ by minimising the KL-divergence and sparsity constraint:

$$\mathcal{X}_{r,q,n} \approx \hat{\mathcal{X}}_{r,q,n} = \sum_{k=1}^{K} \mathbf{C}_{r,k}\mathbf{M}_{q,k}\mathbf{P}_{n,k}\mathbf{D}'_k \tag{12}$$

Since the integrated/segregated percept is inherently modelled by the rank of the decomposition, the number of active components is generalised by the sum of the activations. For each trial under test, the sum of activation energies is compared to a threshold. Values above the threshold should be stimuli which favour the segregated percept, whilst those below it are more likely to be perceived as integrated. The threshold T is taken as the mean activation sum, μ of matrix $\mathbf{D}$:

$$T = \mu \sum_{k=1}^{K} \mathbf{D}_{h,k} \tag{13}$$

although other methods for determining the appropriate threshold could be employed. The key point is that enough separation exists between the distribution of activations sums for the integrated and segregated examples.

5. Experimental Evaluation

It has been experimentally determined that for A-B patterns (Figure 1) of tones of alternating pitch, the sequential organisation of the tones forms one of three percepts: integration, segregation, or an ambiguous percept where neither of the two alternate states dominate, and switching between percepts occurs [7, 18]. The frequency difference between tone pairs, Δf, and the time between successive onsets of the same tone, the tone repetition time

Tone A Frequency	250	500	750	1000	1250	1500	1750	2000	2500	3000	Average
Tone Duration (ms)					**Integrated Percept**						
20	94	78	78	94	100	100	89	78	83	94	88,8
30	89	89	89	83	83	83	83	72	78	78	82,7
40	100	89	89	89	89	89	83	83	78	78	86,7
50	94	83	83	88	94	89	78	78	67	83	83,7
60	94	94	94	89	100	94	89	89	89	94	92,6
Mean	**94,2**	**86,6**	**86,6**	**88,6**	**93,2**	**91**	**84,4**	**80**	**79**	**85,4**	**86,9**
					Segregated Percept						
20	78	83	83	83	83	83	89	72	83	78	81,5
30	72	83	89	89	83	83	89	89	89	89	85,5
40	78	94	84	84	84	84	89	89	89	83	85,8
50	67	83	72	83	78	78	83	83	83	72	78,2
60	50	78	61	72	61	66	78	78	72	61	67,7
Mean	**69**	**84,2**	**77,8**	**82,2**	**77,8**	**78,8**	**85,6**	**82,2**	**83,2**	**76,6**	**79,74**

Table 1: Average classification accuracy over 36 different trials at various tone-A frequencies and tone durations.

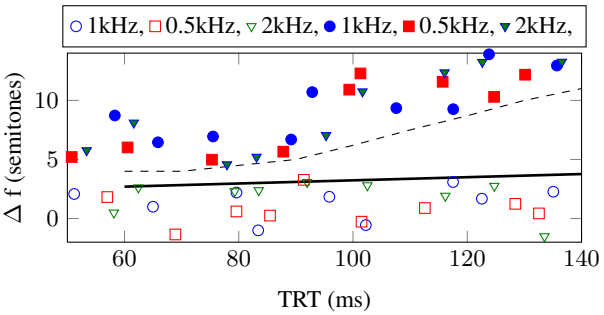

Figure 4: Data used to train the model. Solid filled data points were used as the 'segregated' training data, whilst hollow formed the 'integrated'. Each data point shows the tone-A frequency, TRT and Δf of a particular training example. The dotted line is the temporal coherence boundary, (TCB) whilst the solid line forms the fission boundary (FB).

(TRT), affect the segregation of the stimulus. Van Noorden determined thresholds as a function of Δf and TRT which describe the points at which either segregation or integration is consistently experienced [7]. The temporal coherence boundary (TCB) is the boundary above which the A and B tones split into two separate streams. The fission boundary divides the region below which a single stream is always perceived. Such boundaries are shown as a function of Δf and TRT alongside training data in Figure 4.

We train the model using tone sequences which would produce either an integrated or segregated percept, as per van Noorden's experiments, then use the results of the training to classify new tone sequences, distinct from the training data.

5.1. Training Data

The model is trained by presenting it with 27 examples of 'segregated' tone sequences and 27 'integrated'. The sparsity weights for integrated sequence examples are set to 0.8 and for segregated to 0.2, to encourage a lower number of components to be learned in the model of integrated percepts. A sequence of alternating high and low tones is generated, with differing TRTs and Δf for each example. TRT and Δf values are randomly chosen, with the constraint that they lie either above the TCB for the segregated percept, or below the FB for the integrated percept. Tone duration in all training examples is 40 ms, with a 5 ms raised cosine onset and offset. 9 examples with A-tone frequencies of 500 Hz, 1 kHz and 2 kHz were used for each training percept, and the training data parameters are shown in Figure 4.

5.2. Modelling Stream Segregation as a function of Tone Repetition Time and Frequency Separation

In this set of experiments, new test examples were generated for each tone distinct from the training set following the A-B tone paradigm used in training. Additionally, the length of the tones was varied across trials, from 20-60ms at 10ms increments. Frequency of the A tone was varied from 250Hz to 3000Hz, and TRT and Δf randomly selected to lie either above the TCB or below the FB. A sparsity coefficient of 0.5 (midway between the integrated and segregated sparsity coefficients used in training) was used on all test examples.

5.3. Results

The average classification accuracy over 18 trials for each tone duration and tone-A frequency are shown in Table 1. The experimental evaluation demonstrated that in the majority of cases, a correct prediction could be produced by the model as to either an integrated or segregated percept. On average, a correct integrated prediction was produced in 86.9% of test cases and segregated predictions had lower success, at 79.7% correct. For segregated predictions, highest accuracy was achieved for lower tone durations, with average prediction performance falling as duration increased. Interestingly, for the integrated data, the converse was true, with highest accuracy achieved for 60ms tone durations. There does not appear to be a strong trend in terms of tone-A frequency on prediction accuracy.

6. Conclusions

A feature representation and model was proposed which allowed the prediction of either an integrated or segregated auditory percept for a sequence of tones. It was shown that with a fairly low number of training examples, the proposed model can classify new audio examples of similar form, but distinct from training data. This suggests that audio representation in terms of components approximating modulation frequency and phase, learned through factorisation, is a viable approach for auditory stream segregation modelling.

The model and tests covered in this paper could be extended to cover temporal variation in patterns and percepts. Sparsity weight, and thus, the liklihood of a particular percept, could be updated on a frame-by-frame basis. By taking into account the probability of a different percept occuring based on the current percept state, similar to as in [19], a more perceptually realistic behavioural simulation could be achieved. Such an extension would be useful in working towards incorporating such a signal representation into useful real-world sound-source separation algorithms.

7. References

[1] T. Virtanen, J. Gemmeke, and B. Raj, "Active-set newton algorithm for overcomplete non-negative representations of audio," *IEEE Trans. Audio, Speech and Language Processing*, 2013.

[2] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-Time Speech Separation by Semi-supervised Nonnegative Matrix Factorization," in *Latent Variable Analysis and Signal Separation*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7191, pp. 322–329.

[3] T. Barker and T. Virtanen, "Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation," in *INTERSPEECH*, 2013, pp. 827–831.

[4] T. D. Griffiths and J. D. Warren, "What is an auditory object?" *Nature Reviews Neuroscience*, vol. 5, pp. 887–892, 2004.

[5] A. S. Bregman, *Auditory Scene Analysis*. MIT Press, 1990.

[6] B. C. J. Moore and H. E. Gockel, "Properties of auditory stream formation," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1591, pp. 919–931, 2012.

[7] L. van Noorden, "Temporal coherence in the perception of tone sequences," Ph.D. dissertation, Institute of Perception Research, Eindhoven, the Netherlands, 1975.

[8] M. Elhilali, L. Ma, C. Micheyl, A. J. Oxenham, and S. A. Shamma, "Temporal coherence in the perceptual organization and cortical representation of auditory scenes," *Neuron*, vol. 61, no. 2, pp. 317 – 329, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0896627308010532

[9] S. Christiansen, M. Jepsen, and T. Dau, "Effects of tonotopicity, adaptation, modulation tuning, and temporal coherence in primitive auditory stream segregation," *Journal of the Acoustical Society of America*, vol. 135, no. 1, p. 323333, 2014.

[10] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. ii. spectral and temporal integration," *Acoustical Society of America*, vol. 102, pp. 2906–2919, 1997.

[11] S. L. McCabe and M. J. Denham, "A model of auditory streaming," *The Journal of the Acoustical Society of America*, vol. 101, pp. 1611–1621, 1997.

[12] T. Barker and T. Virtanen, "Semi-supervised non-negative tensor factorisation of modulation spectrograms for monaural speech separation," in *In proc. the 2014 International Joint Conference on Neural Networks*, 2014.

[13] R. D. Patterson, K. Robinson, J. Holdsworth, D. Mckeown, C. Zhang, and M. Allerhand, "Complex Sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, L. Demany, and K. Honer, Eds. Pergamon, Oxford: Pergamon, 1992, pp. 429–443.

[14] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Acoustics, Speech, and Signal Processing, 1997 IEEE International Conference on, ICASSP-97.*, vol. 3, apr 1997, pp. 1647 –1650 vol.3.

[15] S. Hong and R. A. Harshman, "Shifted factor analysis - part iii: N-way generalization and application," *Journal of Chemometrics*, vol. 17, pp. 389–399, 2003.

[16] M. Mørup, L. K. Hansen, S. M. Arnfred, L. Lim, and K. H. Madsen", "Shift invariant multilinear decomposition of neuroimaging data," *NeuroImage*, vol. 42, pp. 1439–1450, 2008.

[17] A. Cichocki, R. Zdunek, A. Phan, and S.-I. Amari, *Nonnegative Matrix and Tensor Factorizations*. Wiley, 2009.

[18] S. Denham, T. M. Bőhm, A. Bendixen, O. Szalrdy, Z. Kocsis, R. Mill, and I. Winkler, "Stable individual characteristics in the perception of multiple embedded patterns in multistable auditory stimuli," *Frontiers in Neuroscience*, vol. 8, no. 25, 2014.

[19] R. Mill, T. M. Bőhm, A. Bendixen, I. Winkler, and S. L. Denham, "Modelling the emergence and dynamics of perceptual organisation in auditory streaming," *PLoS Computational Biology*, vol. 9, 2013.