A BIOLOGICALLY-INSPIRED APPROACH TO THE COCKTAIL PARTY PROBLEM

Mounya Elhilali and Shihab Shamma

Institute for Systems Research, University of Maryland, College Park
Email: {mounya,sas}@isr.umd.edu

ABSTRACT

Though seemingly effortless, our auditory system engages in complex processes and transformations which enable us to segregate speech and other sounds in cocktail party settings. This paper presents a computational approach to modelling monaural auditory scene analysis, where we attempt to account for perceptual and neuronal findings of receptive field selectivity and adaptation in the auditory cortex. The model introduces a biologically-inspired scheme of dynamic segregation of auditory streams, based on unsupervised clustering and the statistical theory of Kalman prediction. Our method demonstrates its ability to emulate known percepts reported by human subjects in auditory streaming and sound organization tests, and yields successful results in segregating speech from concurrent speaker and music interferences.

1. INTRODUCTION

The ability of our brain to identify and follow conversations in the midst of the most severe distortions is a testament to the ingenuity of the auditory system's design as a decoder. We are equipped with an amazing computational tool that is both competent and quite reliable in perceiving sounds and robustly segregating speakers in cluttered and noisy environments. The perceptual capabilities of the auditory system rely on various cognitive principles allowing us to attend to certain aspects of the acoustic information in the auditory scene.

From a computational perspective, development of stream segregation systems is invaluable for numerous engineering applications such as hearing prostheses, automatic speech recognition, and object tracking in sensor networks. Many biologically inspired approaches, including neural networks, have been developed in the last two decades to perform intelligent processing of complex sound mixtures [1]. Despite their valuable contributions, a major shortfall of most existing algorithms is the absence of information integration strategies to consolidate features extracted from the acoustic signal with contextual facts from the environment; hence hindering their applicability to general tasks [1]. On the other end of the spectrum, numerous studies have attempted to solve the problem of sound separation from a strictly engineering perspective (e.g. blind source separation, BBS). Systems built in this spirit are, however, limited by their own mathematical construction or model-based approaches. Hence, stream segregation continues to be a challenge for strictly statistical techniques like BBS systems, particularly in the monaural case, in absence of multi-sensor data.

In this paper, we propose a model that is largely inspired by the perceptual and neuronal findings of auditory stream segregation. The proposed system, described in section 3 presents a computational approach to monaural stream segregation based on unsupervised clustering techniques and the statistical theory of Kalman prediction. This approach yields a robust computational scheme for speaker separation, as validated by the results in section 4.

2. PERCEPTUAL PRINCIPLES FOR AUDITORY SCENE ANALYSIS

Our physiological knowledge of neural properties, particularly in the primary auditory cortex (A1), indicates that cortical neurons exhibit elaborate selectivities to spectral shape, symmetry and dynamics of sound [2]. This intricate mapping of acoustic waveforms into a multidimensional space, along with the known plastic and adaptive nature of cortical responses, suggest a role of the cortical circuitry in representing sounds in terms of auditory objects. Additionally, the time scales of cortical processing appears to be tightly linked to the temporal dynamics of stream formation and auditory grouping. The cortical time constants correspond well to the dynamics of the vocal tract in speech, the rates of musical melody and timbre, as well as the buildup and preservation of streaming [2].

Psychoacoustically, numerous studies have attempted to reveal the acoustic cues used by the auditory system to determine whether sound components are to be fused together into a single perceptual stream or segregate into separate streams. Various studies have identified frequency separation, harmonicity, onset/offset synchrony, AM and FM modulations, sound timbre and spatial location as the most prominent candidates used as grouping cues in auditory streaming [3]. It is however becoming more evident that any sufficiently salient perceptual difference along *any* auditory dimension may lead to stream segregation.

This research is supported by AFOSR and a CRCNS NIH grant.

Based on both psychophysical and psychacoustical evidence, we formalize our focal hypothesis that streaming of a complex sound from a cluttered acoustic environment can be quantitatively predicted based on its segregation in a higher *multidimensional* cortical representation that explicitly includes features related to spectral shape and temporal dynamics of the signal.

3. ALGORITHM

In this work, we propose an algorithm that investigates the stated hypothesis by mapping sound waveforms into a multi-dimensional cortical representation. The segregation of sounds into multiple objects follows the principles dictated by perceptual grouping cues, as described in section 3.2. The integration of these features is then performed using a Kalman-based estimation discussed in section 3.3.

3.1. Peripheral Auditory Processing

Sound signals undergo a series of transformations in the early auditory system, and are converted from a one-dimensional pressure time waveform to a two-dimensional pattern of neural activity distributed along the tonotopic (roughly logarithmic frequency) axis. This two-dimensional pattern, called auditory spectrogram, represents an enhanced and noise-robust estimate of the Fourier spectrogram. We simulate this transformation through a series of stages involving a filter-bank decomposition and sharpening of the temporal and spectral features as described in details in [4] and illustrated in Fig. 1.

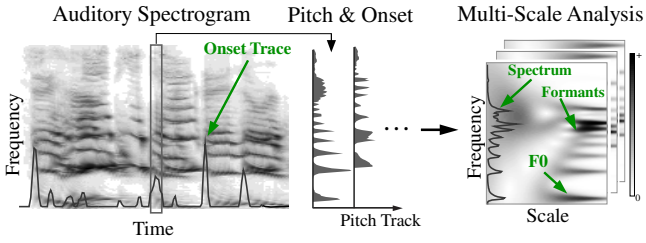

Fig. 1. Peripheral processing and primitive cues extraction.

3.2. Simultaneous Segregation

3.2.1. Primitive Cues

The auditory spectrogram reveals the time-frequency patterns in the auditory scene, hence setting the ground for a spectral analysis of the various acoustic primitive features. This stage focuses on analyzing sound at every cross section in time, and extracts two important cues (when available):

(1) Pitch extraction, which identifies harmonic structures at every temporal cross-section. This stage does not track any frequency trajectories, but works on identifying which frequency channels stand in harmonic relationship to each other at every time instant. Our pitch extraction algorithm is based on a template matching model, similar to that proposed by Goldstein [5]. The model compares incoming spectra against an array of harmonic patterns at different fundamental frequencies (F0), and builds a distribution of pitches based on the matching to the different templates. A threshold is then set to choose the values of F0 with the largest evidence weight at every instant in time;

(2) Onset Estimation: Along with pitch estimates, onset synchrony is a very effective and robust cue for segregating acoustic components. We employ a simple derivation via temporal differentiation to boost the detection of transient energy in the signal. We then proceed to a spectral integration across frequency bands, followed by an energy threshold. Synchronous frequency channels that are activated together emerge as onset spectral segments and tend to coincide with a common sound source.

3.2.2. Multi-scale Analysis

Relying on the premise that spectral shape is an effective physical correlate of the percept of timbre, we perform a multi-scale analysis on each extracted spectral pattern (from pitch and onset estimates). Inspired from findings of cortical spectral analysis, we employ a multi-scale model based on wavelet decomposition [4]. The local and global spectral shapes in the acoustic patterns are captured via a bank of spectral modulation filters tuned at different scales (spanning the range $1/8 - 8$ cycles/octave). This spectral decomposition offers an insight into the timbre components of each acoustic features extracted from the input, and accentuates key sound attributes used to segregate different streams (e.g., timbre features such as formant bandwidths and overall spectral tilts).

3.3. Sequential Integration

While perceptual cues of harmonicity, onset synchrony and timbre can readily yield clean instantaneous "looks" of each speaker or sound source in a mixture, the challenging phase of the analysis is to organize these features together to yield perceptually meaningful streams. Here, we propose to adhere to an approach that does not rely on any dictionaries of linguistic knowledge or databases of familiar sounds, but rather to take advantage of the statistical regularities of sound patterns emanating from a common source. This part of the algorithm proceeds in three steps:

3.3.1. Forward projection

In this stage, we explore the correspondence between slow cortical dynamics and streaming time constants. We model each stream (or sound source) as a bank of modulation selective filters, tuned to temporal rates ranging between 2-40 Hz (Fig. 2). The behavior of these rate filters is governed by their internal dynamics, where each cortical unit integrates sound inputs according to its own time constant.

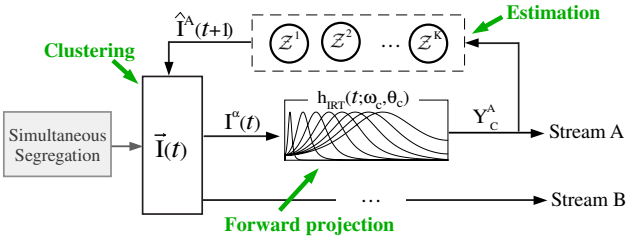

Fig. 2. Diagram of sequential segregation.

Computationally, our current implementation of the algorithm segregates sound mixtures into two streams, A and B (e.g. foreground vs. background). The input going to stream A or B is decided by the clustering module which we review in section 3.3.3. Based on this clustering, once sound feature I^{α} is identified as belonging to stream A for example, it gets processed through the bank of rate filters, where each modulation selective unit yields an output Y given by:

$$Y_c(t, f, \Omega, \omega_c, \theta_c) = I^{\alpha}(t, f, \Omega) *_t h_{IRT}(t; \omega_c, \theta_c) \qquad (1)$$

where t stands for time, f for frequency, Ω for spectral scale (described in section 3.2.2), ω for temporal rate and θ for phase. The cortical integration is captured by the a temporal gamma function $h_t(t) = t^3 e^{-4t} cos(2\pi t)$. This mother wavelet is scaled and shifted at different rates $\{\omega_c\}$, yielding $h_t(t; \omega)$. By sinusoidally interpolating the symmetric seed function $h_t(.)$ and its Hilbert transform, we obtain a directionally selective filter $h_{IRT}(t; \omega_c, \theta_c)$ determined by modulation parameter ω_c and characteristic phase θ_c [4].

Effectively, we implement the array of FIR cortical filters $\{h_{IRT}(.)\}$ as IIR filters whose dynamics are defined by coefficients $\{a_i\}$ and $\{b_i\}$ that capture their tuning and bandwidths [6]. This implementation allows us to project the sound features one instant at a time, while maintaining a memory of recently integrated sound features from previous inputs and outputs in state (or latent) variables $\mathcal{Z}$. These memory elements are introduced by converting the IIR difference equation into state-space form [6], and correspond to delay registers in a direct form II implementation of the IIR equation.

3.3.2. Estimation (feedback) stage

The next stage consists of an estimation problem, where each stream aims to predict its expected input at time $t+1$, based on the accumulated statistical information in that stream up to time t. We define the optimization function for this estimation as a maximization of the model's posterior probability given the recently integrated inputs. This cost function is given in Eq. 2, and expanded using Bayes rule.

$$\begin{aligned} \mathcal{J} &= max P(\vec{\mathcal{Z}}|\mathbf{I}) \\ &= min \textstyle\sum_i \left[-\log P(\mathbf{I}|\mathcal{Z}^i) - \log P(\mathcal{Z}^i) \right] \end{aligned} \qquad (2)$$

The solution of Eq. 2 yields an optimal estimate of the latent variables $\vec{\mathcal{Z}}(t+1)$ and hence can estimate the expected inputs $\hat{I}(t+1)$ for streams A and B. In order to solve Eq. 2, we use the state space form defined in the forward stage but flipping the roles of the input and output vectors (since our goal is to predict the input). The output-input relationship is now defined as:

$$\begin{aligned} \hat{\mathbf{I}}(t) &= \mathbf{A}\mathcal{Z}(t) + \eta(t) \\ \mathcal{Z}(t+1) &= \mathbf{B}\mathcal{Z}(t) + \mathbf{C}Y(t) + \nu(t) \end{aligned} \qquad (3)$$

where the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are derived from the "inverse" rate filter coefficients $\{a_i, b_i\}$ following a canonical state-space derivation described in [6]. $\eta(t)$ and $\nu(t)$ are two noise terms introduced to allow for variability in the estimation. Assuming a Gaussian distribution for the noise elements η and ν, we can replace the probability functions in Eq. 2 with the normal distribution. Hence, we can derive the optimal solution $\hat{\mathcal{Z}}(t+1)$ which is in fact a Kalman estimator [6]. This Kalman estimate $\hat{\mathcal{Z}}$ directly yields the predicted input $\hat{I}(t+1)$ per rate filter per stream (Eq. 3), which we then sum across rate filters to obtain one estimated input prediction $\hat{I}^A(t+1)$ and $\hat{I}^B(t+1)$ for streams A and B, respectively.

3.3.3. Clustering stage

The outcome of the estimation stage is a predicted pattern $\hat{I}^{A,B}(t+1)$ for streams A and B (Fig. 2). The next stage reconciles the predicted input from each stream with the actual incoming inputs extracted from the sound mixture at every time instant. We use a mean-square error (MSE) criterion to cluster the input patterns into belonging to streams A or B. Based on this grouping, we decide which input patterns I^{α} and I^{β} get projected to streams A and B respectively, and hence follow the forward loop described in section 3.3.1.

4. SIMULATION RESULTS

The model was tested on several classic stream segregation conditions to demonstrate its ability to emulate known percepts as reported by human subjects [3]. Figure 4 illustrates examples of the model's results.

Alternating Tone Sequences: In the first row of Fig. 3, we show the results of the classic alternating ABA tone sequence, with a presentation rate of 4Hz for the A-tone and B-tone [3]. The middle and right panels of this first row show the sum of outputs of the different rate filters in each cluster. They reflect what would be considered the "perceived auditory streams".

Crossing Trajectories: The theory of "crossing trajectories" examines the segregation of rising and falling tone sequences that "cross" at a certain point in time. Listeners always report hearing a bouncing pattern when the sound elements are individual tones, and it is very hard for subjects to follow an entire rising or falling sequence. Such effect is also exhibited by the model in the second row of Fig. 3.

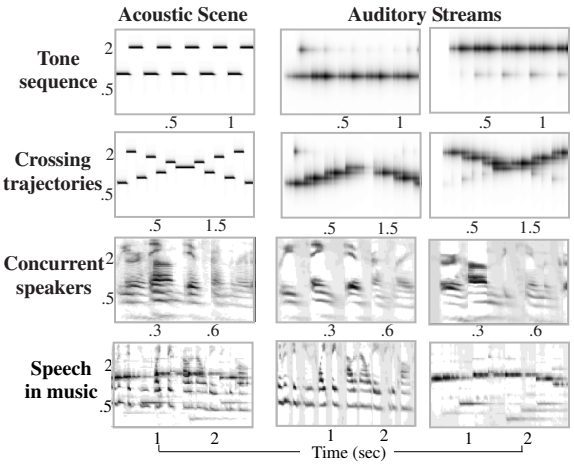

Fig. 3. Stream segregation results.

Speech with Interference: The third and fourth rows of Fig. 3 show the model's results in separating a speech utterance from an interfering second voice (two male voices), or a masking musical melody (a female voice and a flute). The model is successful in segregating between the two interfering streams, primarily using their timbre differences as well as pitch mismatches.

We ran a more through test of speaker segregation using the TIMIT database. These tests consisted of mixtures of pairs of utterances from: male-female, male-male and female-female speakers. We also simulated mixtures of female speakers with male voices whose sentences have been modified so that the pitch matches the female range, but without altering the spectral ratios of his formant energies. Each one of these tests was performed on 50 different pairs of different male and female speakers and utterances from the TIMIT database, where sentences range between 2 and 4 seconds long. The success of the speaker segregation was quantified by correlating the output of the cortical model for each cluster (or stream) with the original "clean" sentence, so as to assess the similarity between the two. The 50 simulations yield average correlations shown in the second column of Table 1.

In order to assess the performance of the model independent of sources of error introduced by pitch estimation and onset detection algorithms (of which there are many in the literature), we repeated the same tests but this time using pairs of clean sentences from two different speakers. Each sentence was analyzed separately through the pre-processing stages in order to map the sound into a multi-scale representation. The features extracted from each are then presented as a combined array of sound patterns, with no reference to which speaker they belong to. These unlabelled patterns are then clustered using the adaptive learning model as before. In the absence of any labelling of their sources, the only evidence that could integrate the patterns of the same speaker in this case was the regularity inherent in the pattern features. The resulting correlation coefficients improved as shown in the third column of Table 1. These results are indicative of the potential power of this computational model.

	Mixture	Original
Male-Female	0.82 ± 0.04	1
Male-Male	0.89 ± 0.05	0.98 ± 0.02
Female-Female	0.92 ± 0.04	0.95 ± 0.05
Female-Modified male	0.85 ± 0.05	1

Table 1. Results of speaker segregation.

5. CONCLUSIONS AND FUTURE DIRECTIONS

The question of how the acoustic scene is parsed by the auditory system into auditory objects and streams is one of the most fundamental in perceptual science. In this work, we presented a model which operates by reconciling acoustic evidence from the input signal (accentuating attributes such as timbre features, e.g. formant bandwidths and overall spectral tilts), and expectations of an internal representation of the different perceptual streams in the environment.

We are currently working on various extensions of the model (e.g. binaural scheme), as well as more thorough testing with variations of speaker identity, gender and accents, under different interference loudness levels. An important future direction is the re-synthesis of the clustered streams. The computational complexity of the model requires intricate processes to revert the cortical representation of the segregated streams into sound waveforms. On-going work on the model is tackling this problem.

Given the promising results yielded by the current model, it is an invaluable tool for exploring the neural basis of auditory scene analysis. This work would lead to important applications in speech recognition front-ends as well as hearing prosthesis systems, which are at the forefront of speech technologies that could immensely profit from our growing understanding of auditory perception in the brain.

6. REFERENCES

[1] M. Cooke and D. P. W. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 35, pp. 141–177, 2001.

[2] S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay, Eds., *Speech processing in the auditory system*, Springer handbook of auditory research. Springer-Verlag, New York, 2004.

[3] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT Press, 1990.

[4] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE transactions on speech and audio processing*, vol. 3, pp. 382–395, 1995.

[5] J. L. Goldstein, "An optimum processor theory for the central formation of the pitch of complex tones," *Journal of the Acoustical Society of America*, vol. 54, pp. 1496–1516, 1973.

[6] D. S. G. Pollock, *A handbook of time-series analysis, signal processing and dynamics*, Academic Press, 1999.