Auditory scene analysis

Ville Huhtala
Aalto Universtiy
Master's Programme CCIS / AAT

ville.huhtala@aalto.fi

Abstract

Auditory scene analysis (ASA) is a model that explains how the (human) auditory system processes acoustic input. The main goal of auditory scene analysis is to describe how sounds are grouped/segregated. The process of auditory grouping/segregation happens both unconsciously and consciously. The unconscious part utilises sound features, such as pitch, location, and scale to form streams when as the conscious part utilises contextual and learned cues to further analyse the auditory streams. The aim of this paper is to give an overall understanding of ASA and its uses by going through the historical development of ASA.

1 Introduction

During everyday life we constantly locate different sound sources consciously or unconsciously. We hear people talking, cars moving, birds signing or wind blowing. All sound events produce their own sound waves that are then combined in our ears. Then, how is it possible that we are able to segregate these sounds and be conscious of several sound sources simultaneously? The book *Auditory Scene Analysis* written by Bregman in 1990 sets the theory of how complex sound sources can be grouped into so called "auditory streams" that are mental representations of our physical world.

In chapter 2 the basic concept of auditory streaming is explained. It is also presented where in the auditory system sound segregation occurs. In chapter 3 different phenomena are presented that show the complexity of sound segregation and two models are presented to explain the sound event creation process. In chapter 4 computational auditory scene analysis (CASA) is introduced. Lastly, a summary is given.

2 Auditory streaming

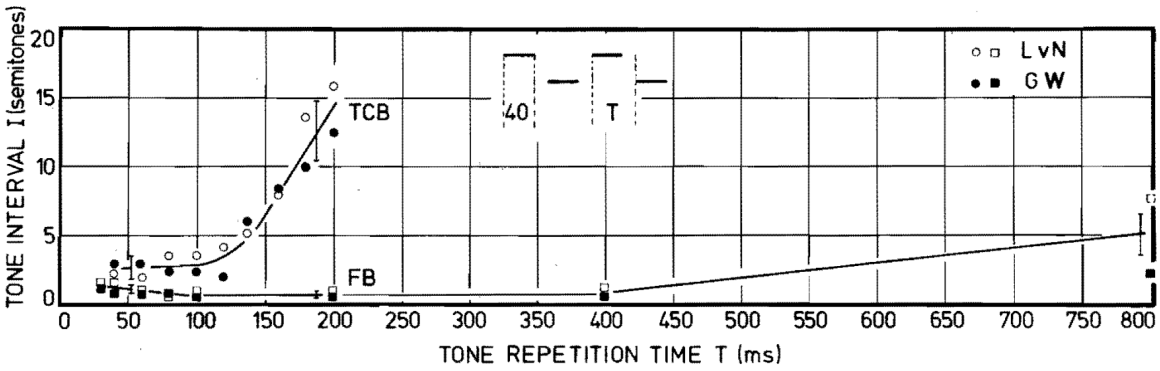
The concept of auditory streams is similar how our vision forms objects: Two 2D images of our surroundings are formed on our retinae when the light that is reflected off objects reaches our eyes. Then, our visual sense processes these images and forms separate descriptions of the individual objects. These objects have a shape, size, distance, coloring, and so on (Bregman, 1990). Bregman (1990) states that an auditory stream represents a single "happening", like fire burning, wind blowing, or person calling, and each happening have descriptions like "low", "high", "near", or "far" that are comparable to the descriptions of visual objects. Bregman (1990) argues that calling the mental representations of physical happenings as auditory streams instead of sounds is logical as happenings can incorporate more than one sound, and it is more convenient to come up with a new word that can be loaded with theoretical properties.

Auditory streams can be either individual sounds like a cough or a clap, or sounds consisting of several sounds that are connected in our mental representation of the physical happening like coughing or clapping. Sound sources contain unique frequency spectra, locations, and loudnesses which make it difficult to determine a universal function that takes in sounds and give auditory streams as an output. However, there have been several studies on how auditory streams are grouped or separated based on some feature of sound such as frequency, pitch, timbre, or loudness (Miller and Heise, 1950; Noorden, 1975, 1977; Darwin. and Bethell-Fox, 1977; Sussman, 2005). One of the earliest study dates back to 1950 to a study named "The trill threshold" where two alternating tones (100 ms) were played in succession to determine whether they are heard as grouped or as two separate signals. It was found out that when the frequency difference between the tones were about 15% or more, the two tones were separated into two streams up to frequencies of about 2000 Hz (Miller and Heise, 1950). This means that the frequency difference (i.e., the trill threshold) increases as a function of frequency as the ratio stays the same. It was later corrected that the segregation is enforced when the frequency difference is large enough and can be influenced by attention at smaller frequency differences (Noorden, 1975).

2.1 Temporal coherence boundary and fission boundary

Noorden (1975) made a distinction between stream segregation (also known as fission) of two tones A and B in a sequence of ABAB..., and the case where the observer can still hear the alternation ABAB... (also known as grouping). The largest frequency interval between the tones A and B where they can still be heard alternating is called the *temporal coherence boundary* (TCB) and the smallest frequency interval between the tones A and B where they can be heard as two separate streams A.A. and B.B. is called the *fission boundary* (FB). When the temporal coherence

Figure 1: Values for temporal coherence boundary and fission boundary by adjusting the frequency f_A. (Noorden, 1975)



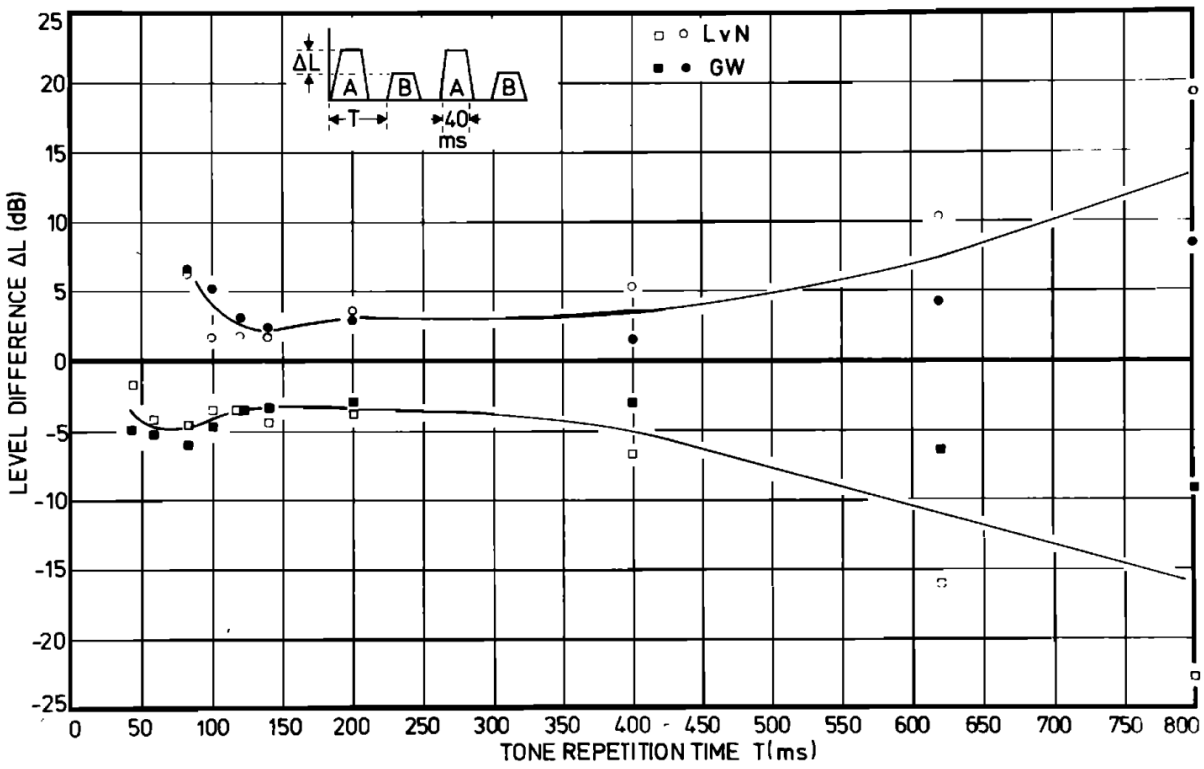
boundary depends on the tone rate, the fission boundary is approximately one semitone over a large range of tone rates (Noorden, 1975). Fig. 1. shows the difference between TCB and FB.

The study that Noorden (1975) conducted to determine the TCB and FB consisted of two pure tones A and B where f_A was adjustable and $f_B = 1$ kHz, tone level being $L_A = L_B = 35$ dB SL, and each tone lasting 40 ms with 5 ms trapezoidal envelope flanks. The tone repetition time T varied between 48 and 200 ms for TCB and 48 and 800 ms for FB (check Fig. 1. for exact times). In the experiment the observer had to adjust f_A to be as small as possible compared to f_B to determine the TCB and FB. However, only the case $f_A > f_B$ was measured with every point being the mean of 15 adjustments of f_A. The result of the experiment was that the observers (LvN and GW in Fig. 1.) were able to distinguish TCB and FB with relatively small spread. Each point represents the mean value of 15 adjustments. From Fig. 1. it can be seen that TCB depends heavily on the tone rate T when as FB can be divided into three ranges depending on T: short ($T < 100$ ms), medium (100 ms$< T <$ 400 ms), and long ($T > 400$ ms). In medium range the tone repetition time has no notable effect on FB frequency. Noorden (1975)

This study lead to the theory that FB could be related to the peripheral-frequency selectivity of the ear. The close relation between the trill threshold and the bandwidths of the critical bands also supported the theory (Noorden, 1977). Noorden (1977) conducted an another study to find out whether stream segregation takes place at such a peripheral level. In the study the amplitude of tone A was varied to find out whether it has an effect on fission boundary. If it has, it could be determined that stream segregation takes place at a higher level in the auditory system.

The study consisted of two pure tones A and B in a sequence ABAB... where the frequency was $f_A = f_B = 1$ kHz, tone level $L_B = 35$ dB SL with tone level L_A being adjustable, and tone duration 40 ms with 5 ms trapezoidal envelope flanks. The tone repetition time T varied between 43 and 800 ms with ten different values. In the study the observer had to adjust the level of tone A six times within the

Figure 2: Values for fission boundary determined by adjusting the amplitude L_A. (Noorden, 1977)



same T value. Three adjustments were made with $L_A > L_B$ and the other three with $L_A < L_B$. Every point is the mean of 15 adjustments. The observers were instructed to adjust the tone level of A to have as small difference to tone level B as possible (to hear tone B as a separate string). The level difference ΔL (dB) of the ten different repetition times are shown in Fig. 2. (Noorden, 1977)

When looking at Fig. 2., the fission boundary can be divided into the same ranges of T as in Fig. 1. This close similarity between frequency and amplitude adjustments indicate that stream segregation does not occur in the peripheral, frequency selective part of hearing, but rather on a more deeper level. (Noorden, 1977)

3 Auditory grouping

In the last chapter, it was showed that auditory streaming is a process that does not occur at the peripheral part of hearing. It was presented that in addition to frequency, the auditory system segregates sounds also based on amplitude changes. This raises the question which features of sound have an effect on grouping or segregating sounds. In this chapter features such as timbre, pitch and location are discussed on how they can affect grouping sounds together. Also, two different theories are presented to explain the presented phenomena.

3.1 Sound features as auditory cues

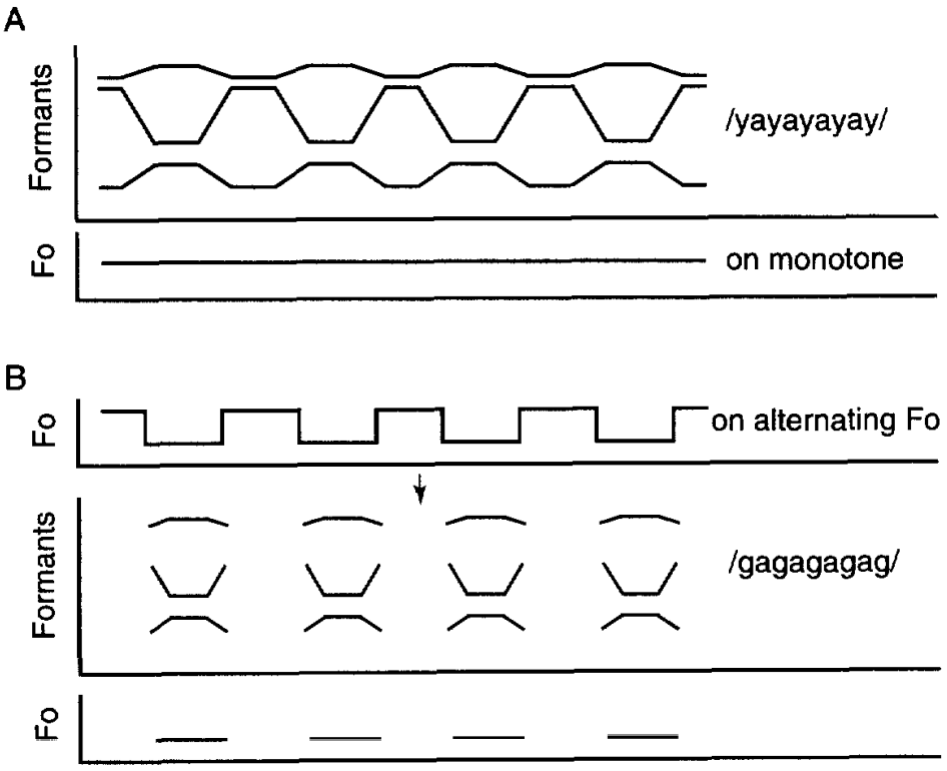
Timbre refers to the "color" of sound. It comprises the spectral energy distribution that changes over time. For example, the sound of a piano and singing on the same pitch are perceived as different sounds because their timbres are different. The Wessel illusion, demonstrates that timbre has an effect on streaming: when an ascending three-tone motif is played with two alternating notes with sufficiently different timbres, the motif streams into two slower motifs that are descending (Wessel, 1979). However, the motif needs to be played rapidly (motif lasting around 1 second) to create the illusion.

Another feature that has an effect on sound segregation is pitch. Pitch refers to the ability to discern sounds as "high" or "low". Discerning pitch is important when two or more speakers are talking. Fig. 3. shows how a change in pitch (with fundamental frequency Fo) of formants causes the auditory system to discern two different voices. It can be seen from the figure that in the first case (A), where three formant resonances are varied over time with a monotone fundamental frequency, only one voice is heard, and in the case where the fundamental frequency alternates (B), two voices that speaks with two different pitches are heard (Darwin. and Bethell-Fox, 1977). This indicates that fundamental frequency plays some kind of role in determining sound sources. Darwin (1997) states that pitch can be used to separate two simultaneous speakers in two ways: either by identifying formant peaks or group formants from the same vowel.

In addition to pitch and timbre, location is also used in segregating sounds. The dominant cue in localizing sounds is the interaural time difference (ITD) which indicates the time it takes for sound to arrive between the ears. Segregating simultaneous speech purely by ITD differences is not effective. Darwin (1997) gives an example that when four noise bands, that are similar to formants and form vowels based on how they are combined are played, the heard vowels are not influenced by noise bands that have the same ITD. On the other hand, Darwin (1997) states that ITD differences are heavily utilised in tracking sound sources across time.

These findings indicate that timbre, pitch and spatial location are mainly used in maintaining the presence of sound sources changing in time. In addition these

Figure 3: (A) When the formants whose resononces change in time, are excited by a monotone fundemental Fo, only one voice is heard. (B) When the formants are excited by an alternating fundamental, the speech breaks into two voices. Even though there are no consonants in the stimulus, the stream segregation creates a stop that is perceived as a consonant 'g'. (Darwin. and Bethell-Fox, 1977)



observations indicate that all features of sound create their own cues in the peripheral part of the auditory system that are then used to determine sound sources by more central mechanisms. (Darwin, 1997)

3.2 Sound segregation in the auditory system

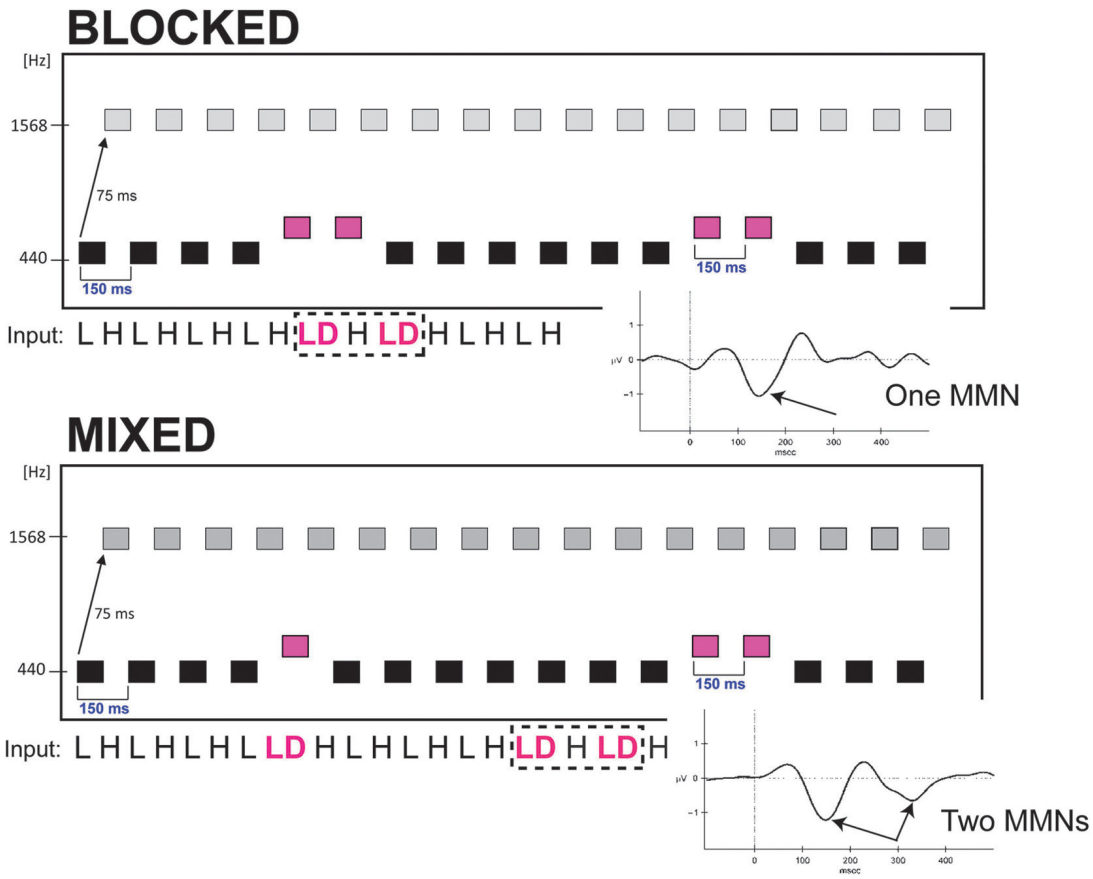
To explain the complexity of ASA, Bregman (1990) proposed two theories. The first theory, called the bottom-up process, occurs in the peripheral part of the auditory system and is thus independent of listeners' attention (Alain et al., 2001). In the bottom-up process the sound waves that arrive to the listener, usually consisting of sounds coming from several sound sources, are grouped into streams that have, for example, close physical similarity, temporal proximity, or continuity (Alain et al., 2001). In practice, this means that sounds are grouped into same streams if they have similar frequency content, intensity, or location. It has been shown that attention is not required to create the initial segregation into streams (Sussman, 2017). A study showed that when ears are presented with sounds containing several frequencies and

attention is not focused on the sound, sounds are still structured into streams in the auditory memory (Sussman, 2005).

The second theory that Bregman (1990) proposed is called the top-down process. In top-down process the streams that have been formed in the bottom-up process are subjected to a more detailed analysis in the central parts in the auditory system (Alain et al., 2001). This means that streams, that corresponds to happenings (also called sound events), are formed based on learned schemas. Schemas are based on previous experiences, learned knowledge, or the context of previously occurred sound events. Thus top-down process is useful when there is a lot of noise present (Alain et al., 2001). An example where attention helps forming sound events is when a person walks into a cocktail party: If the person does not pay attention, the overlapping frequencies of all the happenings, like a person talking and music playing, could result in the streams overlapping. However, if the person pays attention and uses their learned knowledge, they are able to distinguish identifiable sounds, like a person talking, glasses clinking, and music playing (Sussman, 2017). Thus it can be said that attention sharpens the stream segregation process. The study by Sussman (2005) confirmed that sound events are formed by first organising sounds into streams (bottom-up and unattended) and then the streams are formed into perceptual units where attention (top-down) can identify the changes within the streams.

To be able to confirm these theories, one must know how the brain procosses sounds that are either unattended or attended. One such thing is to use event-related brain potentials (ERPs) that give a direct quantifiable measure of brain activity (Sussman, 2017). The ERP component that is usually used in analysing auditory scenes is mismatch negativity (MMN). Sussman (2017) states that MMNs are useful because they are elicited whether or not attention is focused, and it shows how sounds are held in auditory memory among other reasons. Fig. 4. shows why sound segregation occurs before event formation. In both cases two streams are formed (check Fig. 1.) from an alternating series of tones lasting 75 ms. The low tone (L) and high tone (H) are streamed into two streams. On the blocked case, the low-tone stream consists of deviant tones (pink in figure) that are always in pairs of two. The mixed case consists of deviant tones that are either (randomly) one or two tones long. The MMN of the blocked case is only one drop even though there are two deviant tones. That is because the context based (top-down) process is focusing on the lower stream and thus processes the two deviant tones as one sound event (even if there is one high tone in between). There are no context based cues on the mixed case because the amount of deviant tones (one or two) is chosen randomly. Thus the mixed case elicites two MMNs. This proves that stream segregation occurs before contextual clues as the amount of MMNs are different in both cases. Fig. 5. shows how the bottom-up and top-down (attention driven) processes interact with each other. (Sussman, 2017)

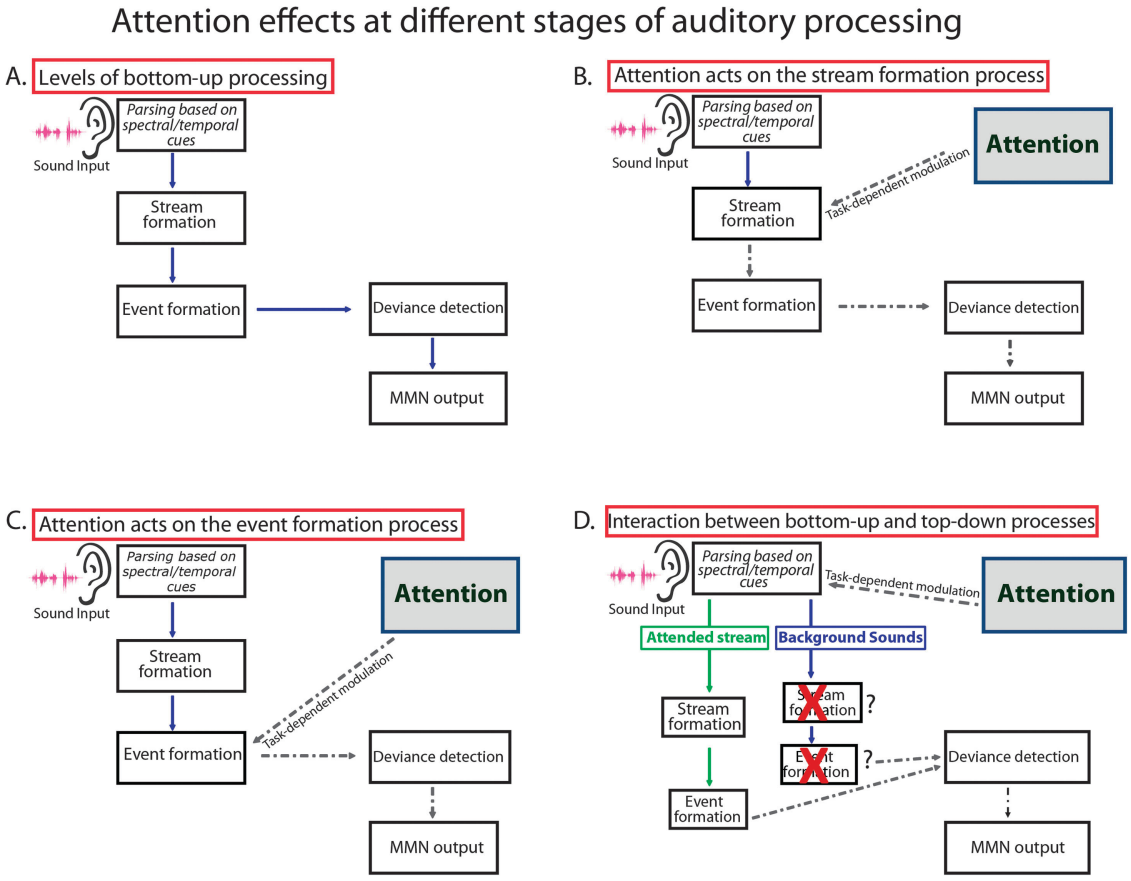
Figure 4: Two tones (high and low) forms two streams. On the blocked case only one MMN is elicited based on the contextual cues when as in the mixed case two MMNs are elicitied (when two deviant tones are played) as the amount of deviant tones is randomly chosen as one or two. This proves that segregation happens before applying context cues to form sound events. (Sussman, 2005)



4 Computational auditory scene analysis

Based on findings on how the auditory system processes sounds, computational models that separate and group complex acoustic input into auditory streams can be created. Computational models of ASA can have two different goals: they either try to process a signal, consisting of complex, realistic, and natural sounds, or try to simulate a specific neurophysiological experiment (such as fission and temporal coherence boundary) (Szabo et al., 2016). Regardless of the goal, Szabo et al. (2016) coins three different broad classes that categorize computational models based on their modeling principles. The three principles suggested by (Szabo et al., 2016) are Bayesian inference, neural processing, and temporal coherence. When explaining different models, it is important to know the basic terminology used to describe them. For simplicity, a sound event (for computational models) is a discrete, isolated sound that has a beginning and an end, The perceptual response that a sound

Figure 5: Model showing how attention (top-down) can effect event formation. In the first case (A), event formation is purely prudeced by the bottom-up process (e.g., unattended). In (B) and (C) cases attention is used to form sound events. In (B) case, attention helps forming the streams (i.e., identifying music in a cocktail party), and in (C) case, attention helps identifying sound events within the stream (i.e., a familiar melody within music). The (D) case shows that passive and active processes may interact for a given task and thus it may limit other processes. (Sussman, 2017)



event elicites within the auditory system is called a perceptual event. Based on the perceptual event, several competing streams are formed that are called proto-objects. The proto-object that "wins the race" and emerges to consciousness is called the perceptual object or object. (Szabo et al., 2016)

Bayesian inference models use predictive mechanisms to estimate the contents of the input. These models decompose the acoustic input into state vectors, that when added, estimate the acoustic input. The decomposition is affected from earlier decompositions. However, the prior prediction varies between models as some models use training, when others opt to use the current decomposition as the prior. In this model objects are created based on the most plausible (proto-object) probability derived using the Bayesian inference method. (Szabo et al., 2016)

In neural models objects are described by units which can be thought of as neurons or network of neurons that could be located in the brain (even though no model claims so). Thus the units' activation strength (i.e., competition), which is usually determined by inhibition, determines the output (object) of the model. The depth and size of the network depends on the model and the goal of the task. (Szabo et al., 2016)

Temporal coherence models forms objects based on the coherence of sound features (e.g, pitch and location) in time and thus eliminates the need for competition. In these models, the auditory scene is usually implemented by 'cortical representation' of the features. This means that after the feature extraction process, temporally similar features are grouped together to from streams. (Szabo et al., 2016)

All of the three models only capture parts of the complexity of ASA and are thus flawed. For example, the Bayesian inference model is abstract on defining the prior adjustment which makes it difficult to determine how they are implemented by the auditory system. Even though the neural models account for competition, they have trouble modeling the effects of prior knowledge. One disadvantage of temporal coherence models are that they do not account for higher-order coherences in sound (such as melodies in music). Future improvements would include the integration of different models utilizing their strengths. That is possible because the models can be thought as implementing different processes in the auditory system. (Szabo et al., 2016)

5 Summary and future development

Auditory scene analysis is a model that tries to explain how the auditory system perceives sound. The current model is divided into bottom-up and top-down processes. The processes are based and validated with physiological experiments. The bottom-up model is a process that happens unattended constantly. The bottom-up process uses spectral and temporal cues to group or segregate signals, when as the top-down process uses learned schemas to further process the groupings. The top-down process is controlled by attention that forms the final sound event that emerges in consciousness. The research on ASA is still progressing as acoustical signals are usually complex and the understanding how one can, for example, experience music is difficult to measure.

Computational auditory scene analysis tries to create mathematical models that describe how sounds can be extracted into sound events (i.e., happenings). CASA is closely related to signal separation but uses human hearing as the basis. CASA is still progressing as there have not been models that can perfectly separate arbitrary signals. However, there have been effective models that try to accomplish certain tasks, like tracking speech, or removing noise from signals.

References

Alain, C., Arnott, S. and Picton, T. (2001), 'Bottom-up and top-down influences on auditory scene analysis: evidence from event-related brain potentials', *Journal of Experimental Psychology Human Perception & Performance* **5**(25), 1072–1089.

Bregman, A. (1990), *Auditory Scene Analysis: The Perceptual Organization of Sound*, The MIT Press, Cambridge, Massachusetts.

Darwin, C. (1997), 'Auditory grouping', *Trends in Cognitive Sciences* **1**(9), 327–333.

Darwin., C. and Bethell-Fox, C. (1977), 'Pitch continuity and speech source attribution', *Journal of Experimental Psychology: Human Perception and Performance* **3**(4), 665–672.

Miller, G. and Heise, G. (1950), 'The trill threshold', *Acoustical Society of America* **22**(5), 637–638.

Noorden, L. (1975), *Temporal coherence in the perception of tone sequences*, Netherlands.

Noorden, L. (1977), 'Minimum differences of level and frequency for perceptual fission of tone sequences abab', *Acoustical Society of America* **61**(4), 637–638.

Sussman, E. (2005), 'Integration and segregation in auditory scene analysis', *The Journal of the Acoustical Society of America* **117**(3), 1285–1298.

Sussman, E. (2017), 'Contribution of stimulus-driven (passive) processing to scene analysis', *Journal of Speech, Language, and Hearing Research* **60**(10), 2989–3000.

Szabo, B., Denham, S. and Winkler, I. (2016), 'Computational models of auditory scene analysis: A review', *Front. Neurosci.* **15**.

Wessel, D. (1979), 'Timbre space as a musical control structure', *Mus. J.* **3**, 45–52.