Buenos Aires – 5 to 9 September, 2016
Acoustics for the 21st Century…

PROCEEDINGS of the 22nd International Congress on Acoustics

The Technology of Binaural Listening & Understanding: Paper ICA2016-718

Contributions of binaural processing to segregating and selecting speech in a complex sound mixture

Barbara Shinn-Cunningham[(a)]

[(a)] Boston University, United States, shinn@bu.edu

Abstract

Intuitively, we all believe that binaural processing plays a critical role in communication, especially at the venerable “cocktail party.” Indeed, if you attend a poster session at a large conference (like the ICA), close your eyes, plug one ear, and try to follow a scientific discussion, you will experience the importance of having two ears. Here we will discuss how binaural processing contributes to two key aspects of understanding speech in crowded settings: focusing attention on whichever source is important in the sound mixture (selection), and separating that source from other sources in the mixture (segregation). Behavioral data show that binaural cues help with source selection, or focusing of auditory attention. When it comes to sound segregation, the contributions of binaural hearing depend on the time scale that one considers. For segregating one speech syllable from a sound mixture, the data suggest that binaural cues are not very salient; they are overridden by other cues such as common onset and offset. However, when connecting together syllables into a continuous stream of speech, spatial cues play a much stronger role. Understanding the role of binaural hearing, and the time scales on which spatial cues matter, perceptually, can guide how binaural cues are used in hearing devices, such as hearing aids, to improve speech understanding in everyday settings.

Keywords: psychoacoustics, binaural hearing, attention, segregation

Contributions of binaural processing to segregating and selecting speech in a complex sound mixture

1 Introduction

Intuitively, when you think about listening to a talker in a crowded setting, you no doubt imagine that knowing *where* the sound source you want to hear is located helps you to understand what a person is saying. Indeed, perceptual research shows that spatial hearing helps in such settings [1, 2]. But the way in which spatial hearing helps is specific, and may not aid perception in the ways that you may imagine. This paper discusses two specific aspects of how listeners use spatial cues: source selection, and source segregation.

2 Selective attention is necessary in crowded settings

When listening for a particular talker in a crowded setting, you must block out other sources, and focus *selective attention* on the talker of interest [3]. It may be that you know the timbre or the pitch of the talker's voice. It may be that you know their location. Such sound *features* allow you to filter out the competing sources (to some degree) and analyse the content of the talker you are trying to follow.

Source: (Shinn-Cunningham, circa 2004)

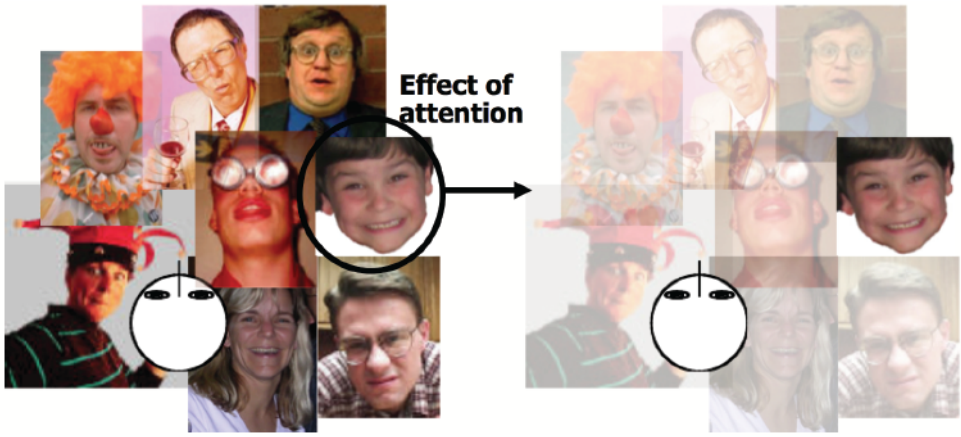
Figure 1: Illustration of the problem of selective attention. Try to look for the author's son Nick (left) in a crowded scene. Finding him (right) requires searching through the scene and focusing on each face until you find him. Importantly, you cannot easily process the whole scene at once to search for him, but must interrogate each individual image to see if it is the correct one. This search can be speeded if you focus for an attribute (yellow shirt) that is unique to Nick.

The challenge of listening in a crowded setting is illustrated by visual analogy in Figure 1. If you are trying to find my son Nick in the scene on the right, it is difficult to do. The challenge of finding Nick in the scene arises not because his face is obscured in any way. Instead, the problem is one of attentional focus. Each face must be examined, one by one, to analyze whether or not it is the target. This visual analogy demonstrates the problem of selective attention, which allows you to analyze one perceptual *object* in a scene. If you know that Nick is wearing glasses, or has on a yellow shirt, you can direct your search of the scene to focus on people with the desired property or feature, and find him more quickly. For instance, looking in the scene for a red target no doubt allows you to quickly focus on the top right clown, who wears a floppy red hat.

Similar mechanisms arise in auditory scenes. There are multiple sources vying for your attention. You can focus on one at any given time and process the content of their speech. Knowing a feature that defines the desired voice, such as the talker's timbre, pitch, or location, is important for allowing you to focus attention on him or her.

3 Spatial hearing is helpful for selecting sources in a scene

The effect of selectively knowing where to listen is to suppress the representation of other competing sounds coming from other directions and "pass through" the information from the desired direction [4, 5]. This is illustrated by visual analogy in Figure 2.



Source: (Shinn-Cunningham, circa 2004)

Figure 2: Illustration of the effect of spatial attention. If attention is directed to a particular location in the scene (left, the author's son Will), the effect on the neural representation is to suppress all other sources in the scene and to "pass through" the source at the attended location.

Binaural cues are powerful cues for guiding attention, providing a feature that you can focus on to select a talker from a scene [6]. This kind of spatial attention focus seems to engage spatio-attentional networks in the brain that have typically been associated with visual processing; increasing evidence suggests that these same networks, which encompass regions in the back,

parietal parts of the brain as well as areas in the frontal (decision-making) regions of the brain, are used for directing spatial auditory attention, as well. For instance, functional magnetic imaging results (which are very precise about where brain activity is originating, but have very sluggish temporal precision) show that the same regions of the brain are engaged by visual and auditory spatial attention [7, 8]. Analysis from electro-encephalography, which reveals brain activity with great temporal precision (but relatively poor precision for where the measured activity is) also suggest that parietal "maps" of external space are engaged during spatial attention tasks [9]. Knowing were to listen is helpful in crowded settings because it helps to filter out unwanted sound sources [10].

4 Spatial hearing helps connect a talker through time

In order to listen to a message, you must successfully focus on a speaker through time. Continuity of various features is important for helping you to segregate one sound source from another—which is a critical step if you are to suppress the unwanted sound sources in a mixture.

There are a number of features that help a listener stream together a voice through time, such as continuity of pitch, timbre, and location. Indeed, many of the features that can be used to focus attention in a volitional manner also are important for helping to maintain attention to an ongoing voice. This might suggest that continuity of features is important because a listener keeps focusing on the feature (such as location) that is continuous—that is, it could be that selecting a talker with a particular feature and maintaining top-down selectional focus on that feature is why continuity of that feature is important. However, a number of studies suggest that the continuity of these features *automatically* supports maintenance of attention on the sound with that continuous feature.

For instance, when listeners are asked to report a sequence of digits that they know may or may not have the same spatial location, they are far better at the task when the spatial location is fixed than when it jumps around [11, 12]. This benefit of spatial continuity is present even when listeners are given large time gaps between digits, and provided with a visual cue about where the upcoming digit will come from well before it starts. Such work demonstrates that spatial continuity is important for helping you to keep attention on a stream.

Other features also provide such continuity advantages. For instance, fixing the identity of the talker of target speech tokens over time improves a listener's ability to report a sequence of digits [12]. Moreover, this benefit of talker continuity is present even if the repetition is random and unpredictable, and is not very likely to occur [13]. Continuity of features also influences perception even if the feature that is continuous is one that a listener knows that they should ignore [14] or if it is irrelevant, such as from a visual input [15]. These results strongly argue that continuity of features automatically helps listeners attend to ongoing sound sources.

5 Spatial hearing does not drive "local" segregation

Bregman [16] noted several "local" sound features that cause sound to be grouped together, perceptually, which he called "integration of simultaneous components" (see reviews by [17, 18]). The rule of spectro-temporal proximity says that sounds that are close together and continuous in time and/or in frequency tend to be perceived as coming from the same source. Sounds that turn on and/or off together also tend to group together, even when they are far separated in frequency and "close together" only in time; more generally, sounds that have correlated fluctuations in amplitude modulation tend to group into the same local perceptual object.

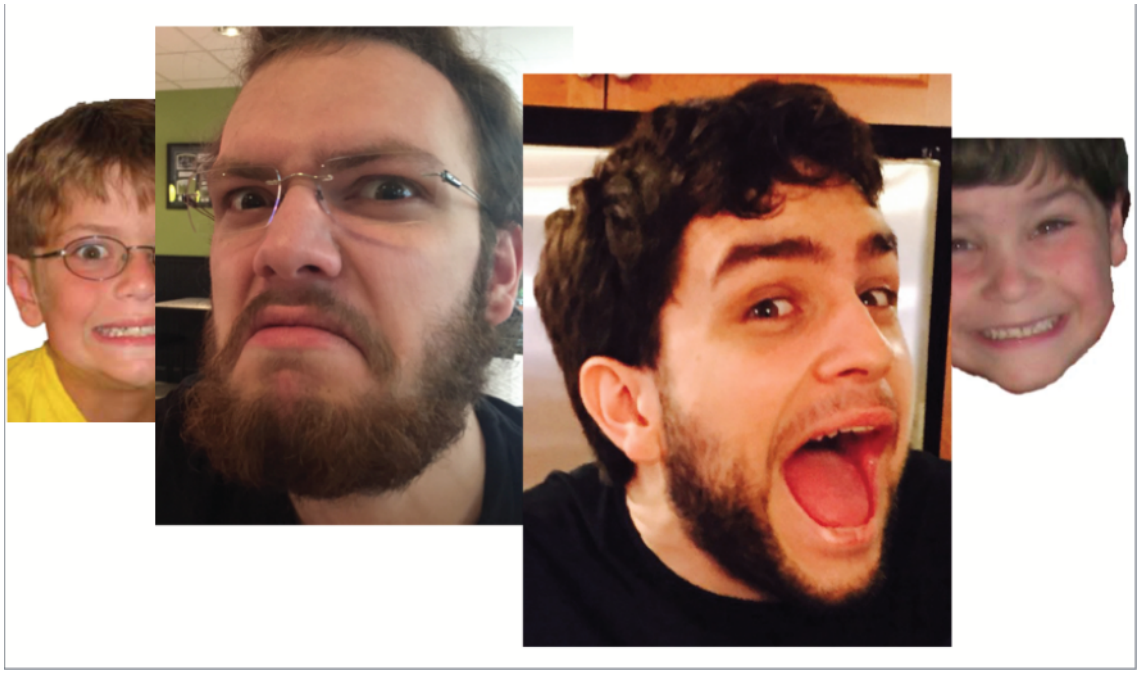
Syllables in everyday spoken English have onset/offset envelopes whose fluctuations fall in the below 10 Hz range, with durations typically between 100 – 450 ms [19], where common onsets and common amplitude and frequency modulations cause the syllabic energy to be perceptually bound together. Often, people assume that the spatial cues of concurrent sounds will impact auditory grouping strongly at the syllabic level; they intuitively suspect that if, at a given time, specific frequencies have binaural cues that all indicate the same location, that those frequencies will be bound together into a syllable, and that frequencies whose spatial cues suggest a different source direction will not be bound with that syllable. However, instantaneous spatial cues actually are relatively weak cues for grouping at the syllabic level. For instance, sound elements that turn on and off together tend to fuse together even if they have spatial cues that are inconsistent with one another [20]; conversely, spatial cues influence local grouping only weakly, with effects that may only be observable when other spectro-temporal cues are ambiguous (e.g., [21, 22]).

This counter-intuitive result may reflect the fact that spatial cues are derived cues, requiring a comparison of the inputs to the two ears, whereas amplitude and harmonic cues are inherent in the peripheral representation of sounds. The modest influence of spatial cues on object formation may also reflect the fact that in the real world, spatial cues are quite unreliable due to effects of reverberation as well as interference from other sound sources [23, 24]. While such effects can distort interaural time and level differences quite significantly, their effects on amplitude modulation or harmonic structure are less pronounced; in line with this, moderate reverberant energy often degrades spatial cues significantly without interfering with perception of other sound properties, such as speech meaning [25, 26]. Although spatial cues have relatively weak effects on grouping at the syllabic level, when target and masker sources are at distinct locations, spatial cues can provide a strong basis for grouping of sequences of syllables into perceptual streams and for disentangling multiple interleaved sequences of sounds [14].

6 Conclusions

Binaural hearing is important in allowing listeners to focus and maintain attention to an ongoing stream of sound. However, focusing and maintaining attention can only be effective when the stream is perceptually and neutrally segregated from other sounds in a mixture. Although binaural cues play a large role in our perception of the world, they do not drive sound

segregation. Nonetheless, one major way in which binaural hearing supports communication in everyday settings is through helping listeners direct attention in noisy settings.



Source: (Shinn-Cunningham, circa 2016)

Figure 3: An update on the author's children.

Acknowledgments

This work was funded in part by grants from the National Institute on Deafness and Other Communication Disorders, the National Science Foundation the Oticon Foundation, and the Coulter Foundation. I would like to offer thanks to my sons for lending their expressive faces to my presentations for so many years (see Figure 3 for an illustration of the duration of their steadfast, good-natured indulgence).

References

[1] Kidd, G., Jr., T.L. Arbogast, C.R. Mason, and F.J. Gallun, The advantage of knowing where to listen. J Acoust Soc Am, 2005. 118(6): p. 3804-15.

[2] Brungart, D.S., B.D. Simpson, and R.L. Freyman, Precedence-based speech segregation in a virtual auditory environment. Journal of the Acoustical Society of America, 2005. 118(5): p. 3241-51.

[3] Shinn-Cunningham, B.G., Object-based auditory and visual attention. Trends Cogn Sci, 2008. 12(5): p. 182-6.

[4] Giuliano, R.J., C.M. Karns, H.J. Neville, and S.A. Hillyard, Early Auditory Evoked Potential Is Modulated by Selective Attention and Related to Individual Differences in Visual Working Memory Capacity. Journal of Cognitive Neuroscience, 2014. 26(12): p. 2682-2690.

[5] Choi, I., S. Rajaram, L.A. Varghese, and B.G. Shinn-Cunningham, Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography. Front Hum Neurosci, 2013. 7: p. 115.

[6] Ihlefeld, A. and B. Shinn-Cunningham, Spatial release from energetic and informational masking in a selective speech identification task. J Acoust Soc Am, 2008. 123(6): p. 4369-79.

[7] Michalka, S.W., M.L. Rosen, L. Kong, B.G. Shinn-Cunningham, and D.C. Somers, Auditory Spatial Coding Flexibly Recruits Anterior, but Not Posterior, Visuotopic Parietal Cortex. Cereb Cortex, 2016. 26(3): p. 1302-8.

[8] Michalka, S.W., L. Kong, M.L. Rosen, B.G. Shinn-Cunningham, and D.C. Somers, Short-Term Memory for Space and Time Flexibly Recruit Complementary Sensory-Biased Frontal Lobe Attention Networks. Neuron, 2015. 87(4): p. 882-92.

[9] Banerjee, S., A.C. Snyder, S. Molholm, and J.J. Foxe, Oscillatory alpha-band mechanisms and the deployment of spatial attention to anticipated auditory and visual target locations: supramodal or sensory-specific control mechanisms? Journal of Neuroscience, 2011. 31(27): p. 9923-32.

[10] Marrone, N., C.R. Mason, and G. Kidd, Tuning in the spatial dimension: evidence from a masked speech identification task. Journal of the Acoustical Society of America, 2008. 124(2): p. 1146-58.

[11] Best, V., B.G. Shinn-Cunningham, E.J. Ozmeral, and N. Kopco, Exploring the benefit of auditory spatial continuity. J Acoust Soc Am, 2010. 127(6): p. EL258-64.

[12] Best, V., E.J. Ozmeral, N. Kopco, and B.G. Shinn-Cunningham, Object continuity enhances selective auditory attention. Proc Natl Acad Sci U S A, 2008. 105(35): p. 13174-8.

[13] Bressler, S., S. Masud, H. Bharadwaj, and B. Shinn-Cunningham, Bottom-up influences of voice continuity in focusing selective auditory attention. Psychological Research, 2014. 78(3): p. 349-60.

[14] Maddox, R.K. and B.G. Shinn-Cunningham, Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention. J Assoc Res Otolaryngol, 2012. 13(1): p. 119-29.

[15] Maddox, R.K., H. Atilgan, J.K. Bizley, and A.K. Lee, Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. Elife, 2015. 4.

[16] Bregman, A.S., Auditory Scene Analysis: The Perceptual Organization of Sound. 1990, Cambridge, MA: MIT Press.

[17] Carlyon, R.P., How the brain separates sounds. Trends Cogn Sci, 2004. 8(10): p. 465-71.

[18] Griffiths, T.D. and J.D. Warren, What is an auditory object? Nat Rev Neurosci, 2004. 5(11): p. 887-92.

[19] Greenberg, S., H. Carvey, L. Hitchcock, and S. Chang, Temporal properties of spontaneous speech—a syllable-centric perspective. Journal of Phonetics, 2003. 31(3–4): p. 465-485.

[20] Darwin, C.J. and R.W. Hukin, Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. J Acoust Soc Am, 1997. 102(4): p. 2316-24.

[21] Shinn-Cunningham, B.G., A.K.C. Lee, and A.J. Oxenham, A sound element gets lost in perceptual competition. Proceedings of the National Academy of Sciences of the United States of America, 2007. 104(29): p. 12223-12227.

[22] Schwartz, A., J.H. McDermott, and B. Shinn-Cunningham, Spatial cues alone produce inaccurate sound segregation: the effect of interaural time differences. J Acoust Soc Am, 2012. 132(1): p. 357-68.

[23] Ihlefeld, A. and B.G. Shinn-Cunningham, Effect of source spectrum on sound localization in an everyday reverberant room. J Acoust Soc Am, 2011. 130(1): p. 324-33.

[24] Palomaki, K.J., G.J. Brown, and D.L. Wang, A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. Speech Communication, 2004. 43(4): p. 361-378.

[25] Culling, J.F., Q. Summerfield, and D.H. Marshall, Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels. Speech Communication, 1994. 14: p. 71-95.

[26] Ruggles, D., H. Bharadwaj, and B.G. Shinn-Cunningham, Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. Proc Natl Acad Sci U S A, 2011. 108(37): p. 15516-21.