Detecting Spatial Displacement in Musical Scenes

Georgios Marentakis
Institute for Electronic Music and Acoustics, Universtitat für Musik und Darstellende Kunst Infelldgasse 10/3, Graz, A-1010 Austria,marentakis@iem.at

Cathryn Griffiths
Department of Psychology, McGill University, 1205 Dr. Penfield Ave., Montreal, Quebec, Canada, H3A 1B1, cathryn.griffiths@mail.mcgill.ca

Stephen McAdams
CIRMMT, Schulich School of Music, McGill University, 555 Sherbrooke St. West, Montreal, Quebec, Canada, H3A 1E3, smc@music.mcgill.ca

Summary

We investigated the detection of sound displacement in a four-voice musical piece in which each voice originated from a different position in space. Displacement was easiest to detect when the voice was played in isolation. Performance dropped in the presence of spatially stable distracting voices when participants were cued to the target voice (selective attention). When participants were not cued to a target voice (divided attention), performance was significantly worse than with selective attention when only one voice moved at a time and marginally worse when all voices moved simultaneously in the same direction. Overall, participants were much less sensitive to spatial displacement than has been observed in studies with white-noise bursts in isolation. Performance improved under conditions of divided attention if a constant pitch was played by each voice. These latter results indicate a task effect in which the increased cognitive load induced by melody processing and by interactions between pitch and spatial-location perception influences the ability to detect changes in the spatial position of sounds.

PACS no. xx.xx.Nn, xx.xx.Nn

1. Introduction

There is an emerging practice among composers in contemporary music to lay out music so that voices occupy space in a static or dynamic way by taking advantage of advancements in spatial audio technology. Spatially distributed music has been composed since Giovanni Gabrieli in the 16th century and more recently by Henry Brant and Charles Ives who created simultaneous spatial layers of music. Edgar Varèse, Karlheinz Stockhausen, Pierre Boulez, Iannis Xenakis, and Roger Reynolds have also created geometrical manipulations and simultaneity of sounds in space [1, 2].

Listening to music requires listeners to attend to multiple voices in a piece. Theories of attention such as the Filter theory [3, 4] deal with how it is possible to separate voices in the mixture by means of attentional processes. It is known that differences in timbre, frequency extent, time structure and spatial location of sounds provide cues that assist in selectively attending to sounds in a mixture. The exact stage of processing at which these cues become available and the extent to which they become available for unattended objects is debatable. Recent theories have proposed that the listening process is assisted by the formation of schemas [5]. It has also been proposed that schemas function dynamically based on internal clocks or time hierarchies. In music, it has been confirmed that listeners can divide their attention more easily among voices that are well segregated (in terms of pitch and timbral differences). It is also known that structural properties such as key relatedness, metric position and harmonic structure further enhance the capacity to divide attention among voices [4]

It has been proposed that the spatial release from masking that occurs when sounds are spatially separated improves both selective and, to a smaller extent, divided attention to sound [6, 7], although attending to multiple spatial locations simultaneously results in

---

a greater processing cost that increases with spatial separation [8]. The auditory spotlight model [8] postulates that auditory attention can be directed to a spatial locus and in this way accounts for the improvement encountered when attending to specific regions of space relative to unattended areas. There is debate, however, as to how the model deals with attention divided among different spatial locations.

2. Presentation of the Experiments

In our experiments, we allocate each voice in a four-voice musical piece to one of four distinct positions in space and measure spatial displacement detection performance for each combination of voice and position under conditions designed to manipulate attentional load. We measure the Minimum Audible Angle (MAA), i.e., the angular displacement of a sound that yields threshold detection performance [9]. We estimate MAAs at a certain location in space with a two-alternative forced-choice (2AFC) paradigm in which an estimator for the angular displacement that yields 75% correct performance is obtained based on the listeners' identification scores for variable angular intervals using the method of constant stimuli [10].

In Experiment 1, participants were asked to detect displacements in the spatial position of each voice under four conditions: 1) with displacement in a single voice played in isolation (Selective Attention without distractors – SA_{isol}); 2) with displacement in a target voice in the presence of the other three spatially fixed voices (Selective Attention with distractors – SA_{dist}); 3) with displacement in any one of the voices in scenes comprising all four voices (Divided Attention with uncoordinated change – DA_{unco}); and 4) with displacement in any of the voices when all four voices move at the same time, in the same direction and by the same amount (Divided Attention with coordinated change – DA_{coord}). Spatial displacements of variable magnitude were presented in each condition. There were two participant groups: musicians and nonmusicians. We hypothesized that change-detection performance would: 1) increase with the amount of spatial displacement in all conditions; 2) decrease as the attentional demand increased; 3) be higher when all four voices moved simultaneously (DA_{coord}) compared to when only one voice at a time moved (DA_{unco}), due to information accumulation; and 4) show superior performance in musicians compared to nonmusicians, the former having had more experience attending selectively to individual voices in complex musical scenes.

In Experiment 2, a task involving DA_{unco} was employed, and two stimulus manipulations were included: 1) because performance was different for the front and back voices in Experiment 1, the nominal locations of the corresponding instruments were reversed to determine whether it was the timbres or the spatial positions that were responsible for this effect, and 2) because a false-alarm analysis in Experiment 1 suggested interference due to pitch changes in each voice, all pitch variation in each voice was removed, and the rhythms of the voices were played at different pitches to determine what role pitch variation played.

3. Experiment 1

3.1. Method

3.1.1. Participants

After giving informed consent, 16 nonmusicians (11 female; mean age = 21 years, SD = 4.0 years) and 16 musicians (7 female; mean age = 23, SD = 5.2 years) participated in the study and were paid for their services. Nonmusicians did not currently play a musical instrument and had less than 2 years training with a musical instrument during childhood. All of the musicians currently played at least one musical instrument on a regular basis. All participants possessed normal hearing within the range 125-8000 Hz as determined by a standard audiogram prior to the start of the experiment.

3.1.2. Stimuli

We employed a 17th-century four-voice chanson by Claudin de Sermisy ("Pour n'avoir onc faulse chose promise", duration: 2 min 4 s), originally composed for a choir. After reviewing a number of scores, we decided on this piece because it contained a similar level of melodic and rhythmic variation in all voices. Four synthetic instruments were used to render the voices: flute at 45° azimuth, clarinet at −45°, English horn at 135°, and French horn at −135°. Instruments were chosen to have easily discriminable timbres so that the task would not be too difficult. The score was coded as a multi-channel MIDI file. MIDI events were embedded in the score to signal a spatial change or a catch trial at predetermined locations. MIDI messages were synthesized into the appropriate instrument sounds using Synful Orchestra (Synful LLC, Woodland Hills, CA). The sounds were subsequently spatialized using Vector-Based Amplitude Panning software [11][12].

3.1.3. Apparatus

A MacMini computer (Apple Computer, Cupertino, CA) running Max/MSP software (Cycling '74, San Francisco, CA) controlled the experiment. Participants were seated on a chair at the center of a circular array of 24 Genelec 8020A loudspeakers (Genelec, Iisalmi, Finland) with a radius of approximately 2 m. Participants indicated when they perceived a spatial change by pressing a button on the keyboard. The experiment took place in an acoustically treated room with dimensions 7.2 m (l) × 5.8 m (w) × 2.4 m (h). The reverberation time, RT_{60}, was estimated using the sine-sweep method. RT_{60} was 1.40, 0.70, 0.34, 0.32, 0.20, 0.18, 0.16, 0.15, 0.13 seconds at 63, 125, 250, 500, 1k, 2k, 4k, 8k, 16k Hz, respectively. The levels of each voice were set to approximately 54.8 dB SPL as measured with a Bruel & Kjaer 2250-D sound-level meter positioned at the center of the loudspeaker array.

3.1.4. Procedure

Each presentation of the piece contained 12 target events at which spatial change might occur. On a given presentation of the piece, half would have spatial change and the other half wouldn't (catch trials). The two classes were marked A and B in the score. Spatial change events were embedded in the piece at locations at which: 1) all four voices attacked the notes simultaneously, 2) no rest preceded any of the voices, and 3) a preceding change was at least 1.5 s earlier (corresponding to at least 1.5 measures in the music). The locations of the spatial change events and the catch trials were identical for each voice and remained constant throughout all experimental conditions. On reception of a spatial change MIDI event, voices were displaced around their nominal spatial origin. In order to have 12 change and catch trials in a given condition, the piece was looped twice for the selective-attention conditions and eight times for the divided-attention conditions. The location of the catch and change trials was reversed each time the piece was repeated to avoid learning effects (e.g., A=change, B=catch in one presentation, then A=catch, B=change in the next one). In the DA_{unco} condition, the voice for which a spatial change would be performed was pre-selected semi-randomly out of the four possible voices, so that no voice's location would change more than twice in succession. In each repetition, a different voice was displaced at a given point of the piece to avoid any learning effects.

There was a response-time window of at least 2 s after each target event (change or catch) within which participant responses were recorded. For change trials, the detection of a spatial change within this window was scored as a hit. No response within this time window was scored as a miss. A response recorded within 2 s succeeding a catch trial was recorded as a false alarm and no response during this period was registered as a correct rejection.

Angular displacements within each condition appeared sequentially in blocks of 12 repetitions combined with 12 catch trials. The angular displacements used were: 6° and 14° for SA_{isol}; 14°, 30°, 45° and 60° for SA_{dist}; 45° and 80° for DA_{unco}; and 14°, 30°, 45°, 60° for DA_{coord}. Angular displacements were decided based on results of pilot experiments with a different set of participants and were tested one at a time in sequential blocks.

There were two experimental sessions on different days, each lasting 1.25 hours. The SA_{dist} and half of the SA_{isol} (angular displacements of 14° and 30°) conditions were completed in the first experimental session; the remaining were completed in the second session. The order of conditions and the sequential presentation of instrumental voices within each condition were counterbalanced across participants within each session. Participants were given one practice run at the beginning of the first experimental session with a voice chosen at random for each participant.

Table I. Experiment 1. Within-condition effects of displacement (D) and voice (V) on d' scores. The effect of musicianship and its interactions were only significant for one condition and are thus not listed here. In DA_{coord}, there is no Voice factor as all voices move simultaneously. SA_{isol} = Selective attention with no distractor sounds, SA_{dist} = Selective attention with distractors, DA_{unco} = Divided attention with uncoordinated spatial displacement and DA_{coord} = Divided attention with coordinated displacement.

Cond	IV	DV(d')
SA_{isol}	D	F(1,30)=133.8, $p < 0.001$
	V	F(3,90)=1.8, $p = 0.151$
	V × D	F(3,90)=4.7, $p < 0.001$
SA_{dist}	D	F(3,90)=93.1, $p < 0.001$
DA_{unco}	D	F(1,30)=18.7, $p < 0.001$
	V	F(3,90)=13.2, $p < 0.001$
	V × D	F(3,90)=5.3, $p = 0.002$
DA_{coord}	D	F(3,90) = 9.8, $p < 0.001$

3.2. Results

In all conditions, there was a significant main effect of Displacement: sensitivity increased with angular displacement ($p < 0.001$). In SA_{isol} and DA_{unco} the amount of increase was different for some of the voices as revealed by the significant interactions between Voice and Displacement ($p < 0.001$ in SA_{isol} and $p < 0.05$ in DA_{unco}).

Voice affected sensitivity only in the SA_{dist} and DA_{unco} conditions ($p < 0.001$). Post-hoc t-tests with Bonferroni-Holm correction showed that sensitivity for the English (EH) and French (FH) horns was significantly higher than that of the clarinet (Cl) and flute (Fl) in both the SA_{dist} and DA_{unco} conditions. Sensitivity was not significantly different between EH and Fl and between Fl and Cl voice pairs. A significant main effect of Musicianship was only observed in the DA_{coord} condition ($F(1,30) = 4.4, p < 0.05$): musicians were significantly better than nonmusicians at detecting scene rotation.

Now consider comparisons between conditions. Sensitivity decreased significantly ($p < 0.001$) from SA_{isol} to SA_{dist} and from SA_{dist} to DA_{unco}. There was a significant increase from DA_{unco} to DA_{coord}. When comparing SA_{dist} to DA_{unco}, no effect of condition was observed for musicians, but nonmusicians' sensitivity dropped significantly ($p = 0.005$) from SA_{dist} ($M = 0.82$, $SE = 0.06$), to DA_{coord} ($M = 0.51$, $SE = 0.12$).

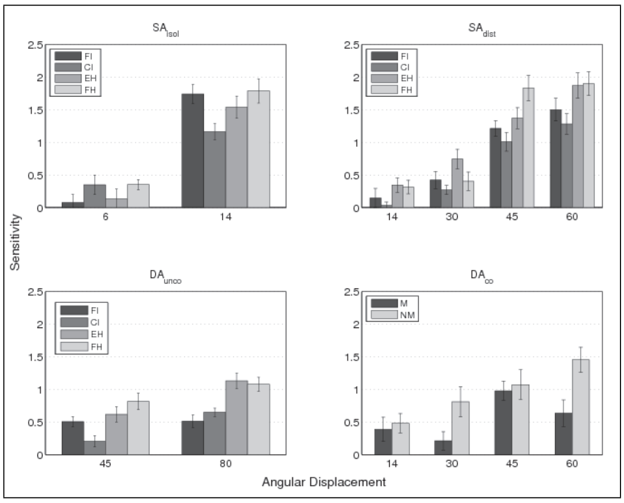

Figure 1. Experiment 1. Mean sensitivity (d') vs. angular displacement (°) in each experimental condition averaged across nonmusicians and musicians. SA_{isol} = Selective attention with no distractor sounds, SA_{dist} = Selective attention with distractors, DA_{unco} = Divided attention with uncoordinated change and DA_{coord} = Divided attention with coordinated change.

4. Experiment 2

4.1. Method

4.1.1. Participants

Sixteen nonmusicians (14 female; Mean age = 24, SD = 7.5) were paid to participate in the study. Five had limited previous experience with a musical instrument in early childhood (flute, clarinet, piano or guitar), and only one had previous experience with experiments related to sound localization. None of them had participated in Experiment 1.

4.1.2. Stimuli

Two different musical pieces were used. Piece A was the same as in Experiment 1. Piece B was created by removing all pitch variation from piece A, maintaining the original rhythms. Different pitches were assigned to each voice to avoid having them fuse perceptually into a single auditory image. The pitch G4 (a fundamental frequency of 391 Hz) was assigned to the synthetic flute, B4 (493 Hz) to the clarinet, E3 (164 Hz) to the English horn and G3 (195 Hz) to the French horn. Spatial change events were embedded at identical locations in the musical piece as in Experiment 1. Angular displacements of 45° and 80° were used. The spatial configuration of the musical scene was flipped with respect to that of Experiment 1, with French horn originating at 45°, English horn at −45°, flute at 135° and clarinet at −135°.

4.1.3. Apparatus & Materials

The apparatus and setup were identical to those used in Experiment 1.

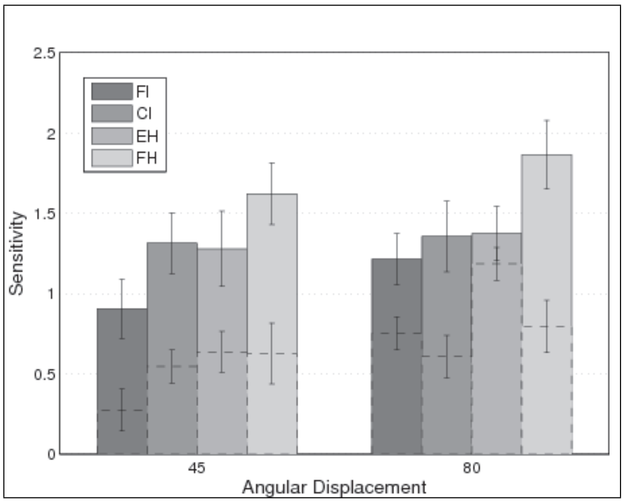

Figure 2. Experiment 2. Mean sensitivity and standard error for each voice in the pitch-change (dashed line) and constant-pitch (solid line) conditions as a function of angular displacement and voice (Fl=flute, Cl=clarinet, EH=English horn, FH=French horn)

Table II. The results of the statistical analyses of data from Experiment 2. Pitch condition comparison: a three-way PC(2) × V(4) × D(2) ANOVA on individuals' d' scores.

IV	d'
PC	$F(1,15) = 20.5, p < 0.001$
V	$F(3,45) = 7.23, p < 0.001$
D	$F(1,15) = 3.93, p = 0.066$
PC × V	$F(3,45) = 2.87, p = 0.047$
PC × D	$F(1,15) = 1.42, p = 0.253$
V × D	$F(3,45) = 2.15, p = 0.107$
PC × D × V	$F(3,45) = 0.51, p = 0.675$

4.1.4. Procedure

There was one experimental session lasting 1 hour within which two experimental conditions were completed in counterbalanced order. These were the same as DA_{unco} in Experiment 1 however they involved: a) pitch changes (Piece A, pitch-change condition) and b) no pitch changes (Piece B, constant-pitch condition). Participants were given trial runs to ensure that they understood the task prior to beginning the experiment.

4.2. Results

Comparing the constant-pitch and pitch-change conditions (factor PC), sensitivity to angular displacement decreased significantly ($p < 0.001$) from d'= 1.28 (SE = 0.08) with constant pitch to d'= 0.65 (SE = 0.13) with pitch change. Response criterion values were significantly higher ($p < 0.001$) in the constant-pitch condition, $M_{Change} = 0.59$, ($SE = 0.08$) vs. $M_{Constant} = 0.83$, ($SE = 0.107$). False-alarm rates were lower in the constant-pitch condition, $M_{Change} = 0.28$, ($SE = 0.04$) vs. $M_{Constant} = 0.22$, ($SE = 0.04$). However, the variance was not

Table III. Cross experiment comparison: Experiment(2) × V(4) × D(2) mixed ANOVA with Experiment as a between-subjects factor on d' for the DA_{coord} conditions of the two experiments.

IV	d'
Exp	F(1,30) = 0.103, $p = 0.751$
V	F(3,90) = 13.2, $p < 0.001$
D	F(1,30) = 17.45, $p < 0.001$
V × D	F(3,90) = 2.93, $p = 0.038$
V × D × Exp	F(1,30) = 4.83, $p = 0.004$

homogeneous for the two displacements involved. A two-way PC(2) × D(2) ANOVA was conducted on the log_{10} transform of the false-alarm rates to account for nonhomogeneity of variance. The reduction in false-alarm rates in the constant-pitch condition was significant ($p = 0.027$).

As in Experiment 1, there was a significant main effect of Voice ($p < 0.001$) with post-hoc comparisons (Bonferroni-Holm correction) showing a similar trend: no differences between flute and clarinet, French horn yielding significantly higher sensitivities than flute and clarinet, but English horn sensitivity being higher than that of flute, but not of clarinet. The latter is likely due to the fact that sensitivity to spatial change for English horn improved less compared to the rest of the voices in the constant-pitch condition, a fact that also explains the significant ($p = 0.047$) interaction between pitch condition and voice. Results are summarized in Figure 2 and in Table II, which present the output of a three-way PC(2) × V(4) × D(2) ANOVA on individuals' d' scores.

Data from the pitch-change condition of Experiment 2 and nonmusicians' DA_{unco} data from Experiment 1 were compared to see if any of the voices yielded different sensitivities in the two experiments. The null effect of experiment, the persisting significant main effect of Voice, and the fact that English and French horns yield higher sensitivity values than did flute and clarinet in Bonferroni-Holm corrected t-tests argue for an effect of the voices' timbres rather than their locations (see Table III). It is worth noting that when examining the pitch-change condition of Experiment 2 alone, a significant effect of voice was replicated, F(3,45)=7.23, $p < 0.001$, and in post-hoc tests French horn yielded superior performance compared to both flute and clarinet, and English horn yielded superior performance compared to clarinet.

5. Discussion & Conclusions

Change detection performance increased with angular displacement. The rate of increase varied as a function of the difficulty of parsing the auditory scene. As a consequence, the same increment in angular displacement did not result in the same improvement in detection performance when participants attended, for example, to a single voice compared to when they divided their attention among four concurrent voices.

Sensitivity to spatial displacement decreased as the attentional load increased and sensitivity was higher in the selective-attention vs. the divided-attention tasks. Dividing attention among the four spatial locations decreased the ability of listeners to detect spatial displacement of the sounds, consistently with the results of studies in which listeners attended to the semantic content of sounds [6, 7]. Interestingly, increasing scene numerosity decreased detection performance even though participants attended selectively to one voice, likely due to the cumulative effect of distracter interference in sound localization [13, 14] and the increased cognitive load induced by complex musical stimuli.

The hypothesis predicting superior change detection performance with coordinated change vs. uncoordinated change under conditions of divided attention was confirmed for musicians but not for nonmusicians. We attribute this finding to the increased ability of musicians to parse complex musical scenes, which enabled them to make use of the coordinated change cues. Hypothesis 4 predicting higher performance for musicians was rejected outside the aforementioned context, because detection ability was similar across groups in all but the scene-rotation condition. The ability to parse a musical scene, which may be enhanced by musical training, does not appear to improve much a listener's ability to attend to spatial manipulations of the musical content.

Our hypothesis that performance would improve when pitch changes were eliminated from the musical piece was confirmed. Overall, participants were far more sensitive to spatial change in the musical piece devoid of pitch changes than in the piece with pitch changes. As argued earlier, this could be due to the increased cognitive load of dealing with melodic and spatial variation simultaneously. Higher response-criterion values and lower false-alarm rates were observed in the absence of melodic variation, indicating a lesser inclination to indicate spatial changes. This difference did not, however, result in lower hit rates as in Experiment 1. The interaction between pitch and spatial change is therefore confirmed by these findings.

The results pertaining to origin location and timbre of the instruments are suggestive. When combining data from both experiments or examining data from Experiment 1, spatial changes for both French and English horns were easier to identify compared to flute and clarinet. However, when examining the data from the piece with normal pitch variation in Experiment 2, we replicated the effect of voice for French horn, but not that of English horn. We believe that the inconsistency for English horn is due to an experimental artifact, in particular the unreasonably low sensitivity scores obtained for this voice in the 80° pitch-change condition in Experiment 2. Consequently, we argue that our results point to an effect of instrumental timbre on listeners' ability to identify changes in its spatial position. Such an effect could be attributed to an influence of timbre either on the localization of sounds or on the perceptual salience each instrument had when all were played together.

Our results confirm the hypothesis that attentional load affects the detection of spatial displacement in a way similar to that found in studies in other domains of auditory attention. Furthermore, the results suggest that timbre is likely to influence perception of spatial displacement in an ensemble. The interaction between spatial and melodic complexity that was observed is likely to influence music perception research, since such a hypothesis has not been explicitly tested here. Similarly, the impact of dividing attention across multiple locations in space could potentially interact with music perception. As the focus of this study was on the perception of spatial displacement, we can only postulate that divided attention in music [15] could be degraded by the presence of voices at multiple spatial positions. Although the effect of spatial separation has been shown to be beneficial when listeners attend selectively to one voice in the piece [16], the magnitude of the degradation imposed by spatial separation remains to be quantified.

The results are relevant for composers seeking to manipulate spatial location in their musical pieces. Interference from other instrumental voices present plays a crucial role. Attentional processes are also important; listeners are more sensitive to spatial manipulations inside their spatial focus of attention and for salient instrumental voices than they are to spatial manipulations of instrumental voices that are less prominent and outside the focus of attention. A balance between melodic complexity and spatial complexity needs to be achieved. Composers need to take into account that dynamically manipulating spatial configurations may not always effectively translate into enhanced perceptual experience.

Acknowledgement

We would like to thank Michel Vallières for helping us with the selection of the musical piece. Albert Bregman provided helpful discussions on this research project. This work was supported by a Canada Research Chair and a grant from the Canadian Natural Sciences and Engineering Research Council (NSERC) to Stephen McAdams and an NSERC grant to Albert Bregman.

References

[1] R. Reynolds. *Form and Method: Composing Music (The Rothschild Essays)*. Routledge, New York, NY, 2002.

[2] M. A. Harley. *Space and Spatialization in Contemporary Music*. PhD thesis, Schulich School of Music, Montreal, 1994.

[3] D. E. Broadbent. The role of auditory localization in attention and memory span. *Journal of Experimental Psychology*, 47:191–196, 1954.

[4] M. R. Jones and W. Yee. *Thinking in sound: The cognitive psychology of human audition*, chapter Attending to Auditory Events: The role of temporal organisation, pages 69–112. Oxford University Press, 2001.

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

[6] A. Ihlefeld and B. G. Shinn-Cunningham. Spatial release from energetic and informatinal masking in a selective speech identification task. *The Journal of the Acoustical Society of America*, 123(6):4369–4379, 2008.

[7] A. Ihlefeld and B. G. Shinn-Cunningham. Spatial release from energetic and informatinal masking in a divided speech identification task. *The Journal of the Acoustical Society of America*, 123(6):4380–4392, 2008.

[8] V. Best, F. J. Gallun, A. Ihlefeld, and B. G. Shinn-Cunningham. The influence of spatial separation on divided listening. *Journal of the Acoustical Society of America*, 120:1506–1516, 2006.

[9] A. W. Mills. On the minimum audible angle. *Journal of the Acoustical Society of America*, 30(4):237–246, 1958.

[10] W. M. Hartmann and B. Rakerd. On the minimum audible angle – a decision theory approach. *Journal of the Acoustical Society of America*, 85(5):2031–2041, 1989.

[11] V. Pulkki. *Spatial Sound Generation and Perception by Amplitude Panning Techniques*. PhD thesis, Helsinki University of Technology, 2001.

[12] E. Lindemann. Music synthesis with reconstructive phrase modeling. *Signal Processing Magazine, IEEE*, 24(2):80–91, 2007.

[13] Michael D. Good and Robert H. Gilkey. Sound localization in noise: The effect of signal-to-noise ratio. *The Journal of the Acoustical Society of America*, 99(2):1108–1117, 1996.

[14] Norbert Kopco, Virginia Best, and Barbara G. Shinn-Cunningham. Sound localization with a preceding distractor. *The Journal of the Acoustical Society of America*, 121(1):420–432, 2007.

[15] E. Bigand, S. Forêt, and S. McAdams. Divided attention in music. *International Journal of Psychilogy*, 35(6):270–278, 2000.

[16] C. J. Darwin and R. W. Hukin. Auditory objects of attention: The role of interaural time differences. *Journal of Experimental Psychology: Human Perception and Performance*, 25:617–629, 1999.