On the influence of interaural differences on onset detection in auditory object formation

Othmar Schimmel
Eindhoven University of Technology, P.O. Box 513 / Building IPO 1.26, 5600 MD Eindhoven, The Netherlands, e-mail: o.v.schimmel@tm.tue.nl

Armin Kohlrausch
Eindhoven University of Technology, P.O. Box 513 / Building IPO 1.33, 5600 MD Eindhoven, The Netherlands, and Philips Research Laboratories, Prof. Holstlaan 4 / Building WO 02, 5656 AA Eindhoven, The Netherlands, e-mail: armin.kohlrausch@philips.com

The research reported here explored a possible influence of interaural differences of a broadband target sound of several different durations on detection of its onset in a diotic broadband masker sound. For two positions in the lateral plane, five target/marker duration combinations and various signal-to-noise ratios, the subject’s ability to accurately align a lateralized Gaussian noise target’s onset to the meter of a regular series of marker pulses in a Gaussian noise masker was investigated. It was found that decreasing the signal-to-noise ratio between target and masker yields a coherent degradation in onset positioning accuracy, used as a quantitative measure for detection of the onset, for all the lateralization conditions and target/marker duration combinations applied.

1 Introduction

Perceptual stream segregation, or auditory object formation, is defined as the human listener’s ability to group signal components, and consequentially separate discrete sources from a complex mixture of sounds. Grouping of signal components is established in the human hearing system by a number of cues that adhere to Gestalt-like principles underlying this auditory organization, like onset/offset synchrony, harmonicity (the ratio between the frequencies of the constituent signal components), common modulation, and spatial localization [1]. These so-called primitive organization principles are assumed to be innate to the human hearing system, and therefore subject independent.

Of particular interest to initiate the grouping of signal components are rapid increases in intensity, if they have a level above detection threshold and occur simultaneously in several decomposed spectral parts of an acoustic mixture, the onsets. Such synchronous onsets are likely to come from the same physical source, and are combined to form a structure that corresponds to a perceptual stream or auditory object. Once a stream is formed on the basis of a detected onset, it is further monitored on frequency content and modulations. Harmonicity, frequency modulation, amplitude modulation, and directional information over time then continue to support the structure, until the energy in the concerning spectral parts decreases below perception threshold and the auditory object is supposed to have disappeared.

The onset of a sound is commonly defined as the beginning of a discrete event in an acoustical signal, providing information about when exactly something is happening. The percept of an onset is caused by a noticeable change in intensity, pitch or timbre of a sound [2]. In studying the effects of rate and asynchrony of tone onsets on the discriminability of onset order, Bregman et al. [3, 4] found that the larger the onset asynchrony and the more sudden the onset, the stronger the segregation from concurrent sounds in a mixture. Theoretically an onset can be detected at the moment a certain level of energy, or change in energy over frequency parts, is reached to make it stand out from an acoustical complex, and a new sound can actually be observed.

Less obvious is the role of spatial localization in forming a new auditory object. Interaural differences in time and intensity enable localization of sound sources, and assist the human auditory system in segregating sounds from and within complex acoustic scenes, due to a spatial release from masking, commonly referred to as the binaural masking level difference [5, 6]. Differences in location of concurrent sound sources provide the physical cues to support efficient listening in so-called cocktail-party conditions [7, 8]. Since directional information is part of every sound when listening binaurally, and it is already present at the onset of a sound, it is not a priori clear what its contribution to grouping is. Do the directional cues in interaural disparities support grouping of signal components and thus make it easier to detect the onset, or is grouping of signal components by onset detection required for deriving localization information?

An established effect of onset on sound source localization is the precedence effect, i.e. the observation that localization of a sound in a reverberant environment is mainly determined by the interaural cues of the first-arriving direct sound to the neglect of later-arriving reflections [9, 10]. Zurek [11] observed that interaural sensitivity to changes in both time and intensity follows a non-monotonic course after the abrupt onset of a sound, and from his experiments the precedence effect can be
seen as a result of a temporary degradation of interaural sensitivity, between approximately 0.5 and 10 ms after an immediate onset.

In assessing a model of auditory object formation, Woods and Colburn [12] assumed object formation to take place before allocation of perceptual features like direction of a sound source. This implies object formation and localization to be independent processes. Drennan et al. [13] found, in their research on the role of spatial location on perceptual segregation, auditory object formation to be depending on interaural differences in time and intensity. Although binaural cues may not be required to segregate a perceptual stream when monaural cues are strong enough, they may contribute concurrently to segregation, and a parallel process can not be excluded.

In general, onsets are shown to have an influence on both auditory object formation and spatial localization. The research presented here was performed as continuation of our previous research [14] that explored the influence of interaural differences on forming a new auditory object, i.e. detecting its onset. The question was addressed whether the availability of different directional cues for two sounds improves their onset detection, and thereby affects the ability to segregate one sound from another, compared to a situation without any difference in directional information in the two sounds. The focus of the current research in the same paradigm was on exploring the possible effects of duration of a lateralized sound on the ability to detect its onset.

2 Experiment

2.1 Method

To determine a sound’s onset detection, an experimental procedure was adapted from speech perception research, based on perceptual center location [15, 16]. The perceptual center procedure comprises judging the temporal position of a target sound, and/or moving a target sound along the time axis of an acoustical stimulus, relative to isochronous reference sounds in the stimulus, in order to find the desired temporal alignment, or perceptual center. These P-centers are related to the acoustic makeup of the stimulus, and in speech they are determined both by vowel onset and duration of preceding consonants. Since perceptually regular lists have regular P-centers, the assumption in the adaptation here is that perceptual isochrony can be achieved with isochronous onsets of the stimulus components.

The stimulus structure used in this experiment consisted of a regular series of five marker pulses of which the middle marker pulse was omitted, against a continuous masker (see Figure 1 for a graphical representation of a stimulus). The subject’s task was to temporally align the spatialized target (in white) in the masker (in light gray), randomly positioned between the second and the third marker pulse at the start of the procedure, to the meter of these clearly audible marker pulses (in dark gray). The time difference (Δt) between the adjusted temporal position of the onset of the target, from the bisector of the time between the two flanking marker pulse onsets (the dotted line at 1000 ms) was recorded, and served as a quantitative measure for the positioning accuracy. From these timing data, the positioning accuracy and bias of differently spatialized targets could be obtained. A monotonic relationship between positioning accuracy and the underlying dimension of onset detection is assumed.

In accordance with Marcus [15], the step size for timing adjustments in this experiment was set to 5 ms, which corresponds to about the minimum sensitivity for perceptual center changes found. The subject controlled the experiment from the computer keyboard. After an initial playback of the stimulus with a randomly positioned target, the target position could be adjusted by pressing f (for ‘forward’) or b (for ‘backward’) any number of times the target was to be shifted by the step size. Confirmation by pressing Enter would start the playback of the stimulus with the target at the new temporal position. This procedure could be repeated until the subject was satisfied with the target’s fit to the meter of the marker pulses. The final time instant of the target sound on the time line of the stimulus was then recorded as the perceived moment of occurrence.

The experiment was performed in an acoustically isolated listening room at the Philips Research Laboratories, using a computer running MATLAB software to digitally generate the stimuli, and automate the experiment and data collection. The stimuli were converted to the analog domain by a Marantz CDA-94 two-channel 16-bit digital-to-analog converter at a sampling rate of 44.1 kHz, and presented to the subjects over Beyer Dynamic DT990 Pro headphones. A within-subject experimental design with counterbalanced complete randomization was applied. Each condition was presented twice to each subject, and the results were averaged across subjects.

---

**Figure 1:** Example of an acoustic stimulus, in which the noise target (white bar) is randomly positioned at about +200 ms from the temporal center at 1000 ms.
2.2 Stimuli

Three types of sound were used: a continuous noise masker with a level of 70 dB SPL, a number of 5-ms or 350-ms marker pulses with a level of 76 dB SPL, and a lateralized noise target of different durations, at a number of different levels. All types were noises with independent Gaussian probability distributions and had immediate onsets and offsets, to enable detection at the actual start of the stimulus’ constituents, and to enable their unambiguous fixation on the stimulus’ time line. The applied target durations were 5 ms, 50 ms or 350 ms. The 50-ms target/50-ms marker combination was established earlier, under otherwise equal experimental conditions [14], and these data were pooled with the results of the current experiment.

The noise masker and the marker pulses were identical at both ears, to establish lateralization in the center. The noise target was presented with different interaural parameters to yield lateralization at two different positions, one in the center (C) and one to the far right (R). This latter case was achieved by usage and controlled manipulation of the interaural time difference (ITD) or interaural level difference (ILD). The value of the interaural time difference in the ITD [R] condition was set at 29 samples at 44.1 kHz, to approximate a 660-µs lag at the left ear. The value of the interaural level difference in the ILD [R] condition was set at 18 dB: target level +9 dB at the right ear, and target level -9 dB at the left ear.

Five different sound pressure levels were applied to the noise target: four, from 0 to +3 dB relative to the previously established detection threshold of the noise target, measured per lateralization condition per target duration per subject (cf. [14]), and one with the same level as the noise masker level, 70 dB. By choosing target levels relative to the corresponding threshold, the sensation levels of the target are equal in all different conditions and data can be pooled across subjects. The measurements at a target level equal to the noise masker level, measured both with and without noise masker, served as control conditions for assessing the temporal positioning accuracy at high signal-to-noise ratios.

2.3 Subjects

Five male subjects participated voluntarily in the experiment. Their ages ranged from 26 to 38 years, with an average age of 30.8 years. All subjects had previous experience in listening tests, and reported normal hearing.

Training for this particular experiment was thought not to be required, since subjects were allowed to reiterate the procedure for each condition as long as necessary to reach a high confidence level about the result, before continuing with the next condition.

2.4 Results

Figure 2 gives an overview of the results per lateralization condition. Indicated on the abscissae is the target level with decreasing signal-to-noise ratio from left to right. To the left are displayed the two control conditions at a target level equal to the 70 dB SPL noise masker [without noise masker, indicated as target level 70 dB SPL (−)], and with noise masker, indicated as target level 70 dB SPL (+)], to the right are the four target levels relative to the detection threshold in the noise masker for that particular condition (target levels 0 to 3 dB SL). The ordinates show the time deviation from the temporal center. The five lines connecting the different symbols represent the mean results of each target/marker duration combination. Negative values signify onsets positioned earlier than the temporal center, positive values signify onsets positioned later than the temporal center. Please note that the target level 70 dB SPL [+] is missing for the 5/5 target/marker combination, since this target level is below detection threshold of the 5-ms noise target in the 70-dB noise masker.

With regard to these mean time deviations from the temporal center, the following is shown in Figure 2. A possible bias in the time deviations of all five target/marker duration combinations in the monaural [C] condition (upper panel) is constrained to a range of about ±30 ms, as already established in the control condition without the noise masker, i.e. at an ‘infinite’ signal-to-noise ratio [70 dB SPL (−)]. The means of the four target/marker combinations with the 5 and 50-ms marker durations in general stay close to the temporal center with a decrease in signal-to-noise ratio, although mainly at the negative side, at all target levels that include the noise masker, except at the lowest target level (0 dB SL). The means of the 50/350 target/marker combination are predominantly at the positive side, and have larger time deviations than the other target/marker combinations.

In both binaural [R] conditions (middle and lower panel), the results in the control condition without the noise masker and the upper limits are very similar to the [C] condition. However, the four target/marker combinations with the 5 and 50-ms marker durations exhibit a considerable shift in mean toward the negative at decreasing signal-to-noise ratios. For example, the mean time deviation in positioning a 350-ms target with 5-ms markers at 2 dB above detection threshold is more than twice (ILD [R]) or four times (ITD [R]) the magnitude of the time deviation at the control condition without the noise masker. Again, the means of the 50/350 target/marker combination are mainly at the positive side, and have time deviations within the same range as in the monaural [C] condition. In general, the mean time deviations of the various target/marker duration combinations are seen to vary more between the different target levels in the ILD [R] condition than in the ITD [R] condition.
Analysis of variance on the pooled raw data reveals, next to the expected significant differences at the 5% level between the lateralization conditions ($F_{(2,810)} = 8.20$, $p = 0.000$, $R^2 = 0.02$) and target levels ($F_{(5,810)} = 4.00$, $p = 0.001$, $R^2 = 0.02$), the differences between the target/marker duration combinations to be significant and account for 9% of the variance in the data ($F_{(5,810)} = 22.39$, $p = 0.000$, $R^2 = 0.09$). Furthermore, the interactions of lateralization condition with target level and target/marker combination are also significant (respectively $F_{(10,810)} = 2.28$, $p = 0.012$, $R^2 = 0.02$ and $F_{(8,810)} = 2.30$, $p = 0.019$, $R^2 = 0.02$). Four out of five significant effects do not exceed a correlation ratio of 0.02, indicating only minor contributions to the variance in the data.

In a post-hoc analysis, the results in the [C] condition are found to be significantly different from those of both ITD [R] and ILD [R], which do not differ significantly from each other. Also not surprisingly, most target levels with low signal-to-noise ratio are found to be different at the 5% level from the two control conditions with high signal-to-noise ratio. With regard to the target/marker duration combinations, with a correlation ratio of 0.09 accounting for the largest part of the variance next to the error, the ones with short durations (5/5, 50/5 and 50/50) are not significantly different from each other. However, all three do differ significantly from the combinations with the longest durations (50/350 and 350/5), which in their turn also differ from each other.

Since the interest was mainly in the accuracy of positioning the target’s onset to the regular meter of marker pulse onsets, regardless of a possible bias, the standard deviations were used for further analysis. In analogy to the previous figure, Figure 3 gives an overview per lateralization condition of the standard deviations of the time deviations, for the five applied target/marker duration combinations.

As can be seen in all three panels of Figure 3, in the control condition without the noise masker [70 dB SPL (−)] the standard deviations of all target/marker duration combinations are generally the smallest. With introduction of the noise masker, the standard deviations of all target/marker duration combinations increase with a decrease in signal-to-noise ratio, showing similar reductions in accuracy of target’s onset positioning for all three lateralization conditions.

However, the precision in the ability to align the target’s onset of the 350-ms target condition (350/5) is seen to decline more rapidly with a decrease in signal-to-noise ratio in the ITD [R] condition, and reach the maximum standard deviations found in both binaural conditions. The other four target/marker combinations stay within the comparable ranges, exhibiting equivalent positioning accuracy of the target at low signal-to-noise ratios for all lateralization conditions.

**Figure 2:** Means of the time deviation from the temporal center, per lateralization condition, N=10.
3 Discussion

At the highest signal-to-noise ratio [70 dB SPL (−)], the means are consistently closer to zero for the target/marker combinations with short durations (5/5, 50/5 and 50/50) than for the ones with the longer durations (50/350 and 350/5), suggesting a shift in P-center for the applied 350-ms target and markers. For the long marker combination (50/350), the target is positioned on average later (about 27 ms), and for the long target combination (350/5), the target is positioned on average earlier (about 25 ms) than the temporal center. Occurring in the control condition with the highest signal-to-noise ratio, this leads to the conclusion that the P-center of the 350-ms target and marker is shifted by approximately the same amount away from the corresponding perceptual center of the 5-ms and 50-ms durations. Supported by the statistical analysis on the pooled raw data, the overall results of these two target/marker combinations are seen as different from each other and from the other three target/marker combinations. This change in P-center of the 350-ms sounds then accounts for the larger time deviations to either side of the temporal center found. It remains, however, a surprising observation that the P-center of these longer sounds is no longer associated with the sound’s onset, despite immediate onsets and an ‘infinite’ signal-to-noise ratio.

When comparing between the monaural and both binaural conditions, from high to low signal-to-noise ratios, the shift in means toward the negative for the four target/marker combinations with short marker durations (5/5, 50/5, 50/50 and 350/5) may reveal a change in the so-called onset effect (cf. [17]). In the case of perceptual ambiguity, due to the inability to perceive a clear onset cue at low signal-to-noise ratios, the binaural hearing system is to look for fine-structure cues in the ongoing part of the signal to establish or support the aggravated lateralization. The required processing of the binaural hearing system to resolve and integrate over time the interaural differences in the acoustical input may interfere with perceiving the exact onset of its constituent elements. In line with Drennan et al. [13], this leads to the conclusion that realization of a new auditory object under these circumstances is dependent on processing of binaural cues.

Target/marker combinations with short durations (5/5, 50/5 and 50/50) are found to yield similar accuracy in temporal positioning of the noise target’s onset, i.e. the 5/5 or 50/5 combinations do not lead to more precise onset detection than the 50/50 combination. Given the established sensitivity for P-center variation [15], this finding shows that, for durations of 50 ms, subjects indeed focus on the onset of both marker and target, supporting our previous conclusions with regard to the influence of interaural differences on onset detection of the 50-ms target and markers [14].
4 Conclusion

Different methods for lateralization of a target sound were applied with different though coherent results in influencing the onset detectability of a Gaussian noise target of several durations and levels in a continuous Gaussian noise masker. It was found that the ability to accurately align the onset of a noise target to a regular meter of marker pulses is influenced by the target’s directional information (time deviations are generally larger in the binaural conditions), level (time deviations increase with decreasing signal-to-noise ratio) and duration (time deviations are smaller for short durations). The obtained data show a decrease in positioning accuracy, but no significant bias, at low signal-to-noise ratios for the short target durations. The longer target duration yielded a shift in the perceived onset, indicating a more complex interaction of the target’s features on determining its onset. The significant differences between lateralization conditions, between target levels, and between target/marker duration combinations suggest binaural processing to unmask the noise target from the noise masker, due to interaural disparities and weak onset cues, to be responsible for deteriorated performance in detecting and temporally positioning the noise target’s onset at low signal-to-noise ratios. With regard to auditory object formation, this reduced accuracy indicates that for long sounds there is less advantage in binaural hearing for a more complex task such as precise onset detection than for simply detection.

References