Original Article

Trends in Hearing
2016, Vol. 20: 1–9

DOI: 10.1177/2331216516664343
tia.sagepub.com

A Detection-Theoretic Analysis of Auditory Streaming and Its Relation to Auditory Masking

An-Chieh Chang[1], Robert Lutfi[1], Jungmee Lee[1], and Inseok Heo[2]

Abstract

Research on hearing has long been challenged with understanding our exceptional ability to *hear out* individual sounds in a mixture (the so-called cocktail party problem). Two general approaches to the problem have been taken using sequences of tones as stimuli. The first has focused on our tendency to hear sequences, sufficiently separated in frequency, split into separate cohesive streams (auditory streaming). The second has focused on our ability to detect a change in one sequence, ignoring all others (auditory masking). The two phenomena are clearly related, but that relation has never been evaluated analytically. This article offers a detection-theoretic analysis of the relation between multitone streaming and masking that underscores the expected similarities and differences between these phenomena and the predicted outcome of experiments in each case. The key to establishing this relation is the function linking performance to the information divergence of the tone sequences, *DKL* (a measure of the statistical separation of their parameters). A strong prediction is that streaming and masking of tones will be a common function of *DKL* provided that the statistical properties of sequences are symmetric. Results of experiments are reported supporting this prediction.

Keywords

hearing, auditory perception, auditory masking, auditory streaming

Date received: 1 April 2016; revised: 19 July 2016; accepted: 19 July 2016

Introduction

We take for granted our ability in noisy surroundings to focus our attention on those sounds that have interest for us, filtering out the rest. It is, however, a truly remarkable ability considering that the sound ultimately reaching our ears is a superposition of all sounds present at any one time in our environment. Understanding this ability has long been a challenge for hearing research known as the cocktail party problem—a reference to the everyday example of having to follow the conversation of a single speaker in a noisy crowd (cf. Cherry, 1953). Work on the problem now represents one of the most active areas of research in acoustics, informing related work on the problem of computational auditory scene analysis (Wang & Brown, 2006) and serving as the launching point for research on noise interference in individuals with hearing loss (Kidd, Mason, Richards, Gallun, & Durlach, 2008).

Historically, research on the cocktail party problem has followed two parallel lines of investigation born of two fundamentally different theoretical approaches. The first approach takes its inspiration from work in vision on the perception of complex scenes (Wertheimer, 1924/1938). Here, Gestalt principles of perceptual grouping describe how elements of the scene are organized in the perceptual formation of objects. These principles then serve as a heuristic to guide future research. In audition, the approach is fundamentally the same; the cocktail party problem is viewed as one in which Gestalt principles of grouping govern the perception of *auditory objects* making up an *auditory scene* (Bregman, 1990).

[1]Department of Communication Sciences and Disorders, University of Wisconsin–Madison, WI, USA
[2]Department of Electrical and Computer Engineering, University of Wisconsin–Madison, WI, USA

Corresponding author:
Robert Lutfi, Department of Communication Sciences and Disorders, University of Wisconsin–Madison, 1975 Willow Drive, Madison, WI 53706, USA.
Email: robert.lutfi@wisc.edu

The stimuli used in these studies are characteristically long, repetitive sequences of sounds (typically tones or tone complexes) that differ in some qualitative sound attribute, most often pitch or timbre. After some time, listening (typically around 5 s) the tones can be heard to split into separate cohesive streams—much like how one hears separate streams of ongoing conversations at a cocktail party. The interest in perceptual streaming stems from its identification with the formation of auditory objects (the separate streams in this case) comparable to the visual objects of a graphic scene. The general observation made in these studies is that streaming becomes less likely (grouping more likely) as the *perceptual similarity* between tone sequences increases (see Moore & Gockel, 2012 for a contemporary review).

The second line of research on the cocktail party problem has its underpinnings in the statistical theory of signal detection. It frames the problem as a classical signal-in-noise detection task. The heuristic, in this case, is the decision variable of an ideal observer, an observer who, based on what is known regarding the statistical properties of signal and noise, maximizes the likelihood of reporting correctly whether a signal is present in the noise. The focus is on masking studies wherein the nature of the interference produced by the noise can be compared with what may be expected based on the analysis of ideal observers. As in streaming studies, the stimuli of these masking studies are often tones, but unlike the long repetitive sequences of tones used in streaming studies, the tones are presented in brief collective bursts, typically a half-second or less, with the pattern of tones varying at random on each presentation or trial. The brevity and novelty of the tone patterns introduce an element of uncertainty regarding stimuli, closer to everyday listening. This challenges the listener's ability to detect as signal a change in a subset of the tones identified as targets. The focus of masking studies has been on factors that allow a listener to overcome such detrimental effects of *stimulus uncertainty* (see Kidd et al., 2008 for review).[1]

Because of the fundamental differences in stimuli and task, streaming and masking studies have long coexisted as separate lines of investigation, each offering a somewhat different take on the cocktail party problem. This situation has changed in recent years as authors have begun to show interest in applying principles of perceptual grouping, gleaned from streaming studies, to the design of masking studies and the interpretation of their results. The approach is perhaps best exemplified by the work of the Boston group of researchers (Durlach et al., 2003; Kidd, Mason, & Arbogast, 2002; Kidd, Mason, Deliwala, Woods, & Colburn, 1994; Kidd, Mason, & Richards, 2003). These authors report multitone masking studies wherein the patterns of target and masker tones are made qualitatively similar to one another so as to promote perceptual grouping of target and masker (failure of streaming). For example, in one such study (Durlach et al., 2003), the masker was a complex of tone glides with random starting frequencies, all sweeping upward in frequency at the same rate. The target was a tone glide that swept upward in frequency, identically to that of the masker, or downward in frequency over the same frequency range. Consistent with the Gestalt principle of grouping by common fate, significantly more masking was found when the target was swept upward in frequency identical to the masker (cf. Bregman, 1990, pp. 213–393). Similar effects on masking have been reported for other factors expected from streaming studies to promote perceptual grouping. These include grouping by common onsets (Durlach et al., 2003), common timbre (Bey & McAdams, 2003), similar spectral content (Hartman & Johnson, 1991; Micheyl & Oxenham, 2010; van Norden, 1975; Vliegen, Moore, & Oxenham, 1999), close spatial proximity (Durlach et al., 2003; Hartman & Johnson, 1991; Kidd et al., 1994), temporal coherence (Micheyl, Kreft, Shamma, & Oxenham, 2013), harmonicity (Micheyl, Kreft, et al., 2013; Oh & Lutfi, 2000; Vliegen et al., 1999), and frequency and amplitude comodulation (Dau, Ewert, & Oxenham, 2009; Kidd et al., 1994, 2002, 2003; Micheyl, Shamma, & Oxenham, 2013; Oxenham & Dau, 2001).

The findings of these more recent studies raise the important question as to how streaming and masking are to be interpreted in the broader context of the cocktail party problem. Given the many parallels evidenced between the two phenomena, one might begin by asking whether there is a compelling reason to distinguish between them. Might they be, at least in the context of relevant studies, complementary phenomena; streaming being synonymous with release from masking and masking being synonymous with perceptual grouping? Several of the authors mentioned earlier have hinted at the possibility in ruling out alternative explanations for the release of masking observed in their studies (cf. Kidd et al., 2003; Micheyl, Shamma, et al., 2013). And, in light of the supporting evidence cited thus far, the presumption that the phenomena are complementary continues to be a major force motivating contemporary theory and research on auditory masking and hearing loss (Moore, 2002). Still, it is difficult to conclude this or any other relation between streaming and masking at this point because the relation has so far only been half explored. While the Gestalt approach has proven successful in predicting the effects of target-masker similarity on masking, the effects of stimulus uncertainty on streaming have so far been virtually ignored (cf. Bendixen, Denham, Gyimesi, & Winkler, 2010; Bendixen et al., 2013; Micheyl, Shamma, et al., 2013; Szalardy, Bendixen, Bohm, Davies, & Denham, 2014).

This study attempts somewhat to remedy this situation. It uses detection theory to make predictions for the effects of stimulus uncertainty on streaming based on the premise that listeners approach the streaming task as would an ideal observer. By this view, streaming is simply the perceptual by-product of an auditory system that has evolved to maximize the likelihood that sounds emanating from separate sources will, in fact, be perceived as separate. The premise gives rise to a strong prediction regarding the relation between streaming and masking that is evaluated in the present study. The prediction is that stimulus uncertainty and similarity are conflated in masking and streaming, that a single function of the statistical separation of tone sequences describes the effects of both factors on both phenomena.

Ideal Observer Analysis

Our problem can be stated generally as follows: Given two sequences of sounds A and B varying along some acoustic dimension x, (a) what is the best strategy for deciding, based on the observed values of x, whether A and B belong to the same or separate sources (streaming experiment), and (b) under what conditions are these decisions expected to agree or disagree with those of an optimal decision strategy for detecting a change in B alone (masking experiment)?

Consider the first question. Here the task of deciding whether A and B belong to the same or separate sources can be framed as a statistical test of whether the A and B values of x represent samples drawn from the same or different population distributions. This manner of stating the problem has the desirable property of being quite general, but it is too general to allow the specification of a single ideal decision statistic; that statistic must depend on what the listener knows about the form and parameter values of the underlying distributions. Rather than consider special cases, we consider a *good* statistical test, not necessarily optimal in every case, that has broad application to such problems; this is the two-sample, Kolmogorov–Smirnoff (*KS*) test (Noether, 1978). The two-sample, *KS* test estimates the likelihood that two samples are drawn from different populations by providing a nonparametric measure of the difference between their cumulative distribution functions (cdfs). Let P and Q represent the cdfs of the sampled x for the A and B sequences. *KS* is then given as the supremum (maximum) of the absolute value of the difference between P and Q

$$KS(p,q) = \text{supremum}\left|P - Q\right| \tag{1}$$

where p and q are the associated probability densities (pdfs) for each sample.

KS has three properties of interest here. First, it is a dimensionless quantity; its predictions for streaming depend less on the acoustic properties of sequences than their statistics given by p and q. The acoustics are expected to affect streaming only to the extent that they cause interactions among tones peripherally, in the cochlea or auditory nerve, or in cases where streaming is based on the statistics of derived higher level features computed centrally (e.g., interaural time differences or fundamental frequency). Second, *KS* does not explicitly distinguish between the effects of stimulus similarity and uncertainty. Stimulus similarity could conceptually be associated with a difference in the central tendencies of p and q, and uncertainty with their spread, but there is little reason to make such as distinction as the two factors are entirely conflated in *KS*. Third, *KS* is symmetric, its value is the same whether p is considered with respect to q or q is considered with respect to p, $KS(p,q) = KS(q,p)$. It thus makes the strong prediction that the variance and higher moments of the pdfs for the A and B sequences can be entirely interchanged without any impact on the listener's judgments of streaming.

Now consider how these predictions compare with those for the ideal observer in masking experiments (Question 2). The task is to detect a change in B; hence, the values of A now represent an additive source of interference or noise whose effect must depend on the degree to which the pdfs of A and B overlap. Focusing exclusively on the degree of overlap of the pdfs (ignoring the change to be detected), the relevant statistic is Kullback–Leibler divergence (*DKL*), also known as information divergence, discrimination information, or relative entropy (cf. Kullback & Leibler, 1951). *DKL* is defined as the expected value of the log-likelihood ratio of x under p and q

$$DKL(p\|q) = \mathrm{E}\ln\left(\frac{p(x)}{q(x)}\right) \tag{2}$$

and, in fact, has been shown to be predictive of many of the results of multitone masking experiments (Lutfi, 1993; Lutfi & Doherty, 1994; Lutfi, Gilbertson, Chang, & Stamas, 2013; Oh & Lutfi, 1998, 1999). As an index of the statistical overlap of pdfs, *DKL* shares many of the same properties of *KS*. Like *KS*, it is a dimensionless quantity; it depends on the statistical properties of sequences not their acoustics. Like *KS*, and for the same reason, *DKL* conflates stimulus similarity and uncertainty. However, unlike *KS*, *DKL* is asymmetric, $DKL(p\|q) \neq DKL(q\|p)$; only in special cases will its value remain the same when the statistical properties of p and q are interchanged. For masking experiments, this means that results can differ depending on whether the A or the B sequence is identified as the target.

This article is one in a series intended to evaluate these three properties of streaming and masking predicted by the ideal observer analysis. The first property,

nondimensionality, has so far been evaluated for masking by Lutfi, Gilbertson, et al. (2013) and for streaming by Chang, Lutfi, and Lee (2015). Results of these studies do, as predicted, show for both phenomena a strong dependence on the statistical properties of stimuli, independent of their acoustic properties. The second property, the conflation of stimulus similarity and uncertainty, is evaluated in the present study. The prediction is that a single function of *DKL* (or *KS*) will describe the data in both cases for both phenomena. The third property, (a)symmetry is the one case where streaming and masking are expected to differ. This property will be evaluated in a future study.

Method

Stimuli

Figure 1 gives a schematic of the stimuli used in the present experiment. In keeping with past streaming studies, the stimuli were standard ABA-ABA tone sequences, where the A and B tones differed in frequency. The present sequences departed from the standard only inasmuch as a small level increment of 3 dB was added to every other B tone; a requirement for the masking experiment. The tones were 100 ms in duration and were gated on and off with 5-ms, cosine-squared ramps. A silent interval of 100 ms separated each ABA tone triplet. The programming language MATLAB (version r2015a) was used to synthesize tones played over an RME audio interface at 16-bit resolution and a 44100-Hz sampling rate. From the interface, the sounds were buffered through a Rolls RA62c headphone amplifier and then delivered diotically over Beyerdynamic DT 990 headphones to listeners seated individually in a double-wall, IAC sound-attenuated chamber. A loudness balancing procedure was used to calibrate tone level to be approximately 65 dB SPL at the eardrum, 68 dB SPL for every other B tone (Lutfi, Liu, & Stoelinga, 2008).

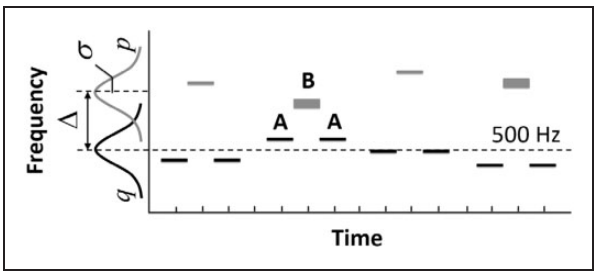

Figure 1. Schematic of stimulus configuration. The frequencies of the A (black) and B (gray) tones were drawn at random from equal variance normal distributions (*p* and *q*) separated in mean, with the mean of the A tones fixed at 500 Hz. The independent variables were the distribution parameters Δ and σ. The listener's task in different conditions was to report their confidence of hearing the B tones alternate in level indicated by width of gray rectangles (masking task) or hearing the A and B tones split into separate cohesive streams (streaming task).

The frequencies of the A and B tones were drawn at random on each presentation from equal variate normal distributions differing in mean. The distributions thus define the *p* and *q* values used in the computation of *KS* and *DKL* in this study. The only constraint on sampling was that the two A tones within each triplet have the same frequency. The A and B sequences were made perceptually more similar, as is commonly done in streaming studies, by reducing the mean frequency separation, Δ, between the sequences. They were made more uncertain, as is commonly done in masking studies, by increasing the range, in this case σ; over which the frequencies of tones varied at random. Nine conditions were tested in which three values of Δ (100, 600, and 1,500 cents, ref: 500 Hz) were combined with three values of σ (100, 200, or 600 cents, ref: 500 Hz), with the mean frequency of the A tones fixed at 500 Hz.

Procedure

On a given trial, the tone sequences were played continuously for 1 min (corresponding to 150 ABA triplets) with the values Δ and σ fixed. During this time, listeners rated continuously their level of confidence hearing the A and B sequences form separate streams (streaming experiment) or, in a separate condition, their level of confidence hearing the B tones alternate in level (masking experiment). The *signal* (separate sequences or alternating level) was always present during this time so as to be consistent with past streaming studies. The confidence ratings thus reflected both the listener's bias to report hearing the signal as well as their sensitivity to hearing the signal, a point we shall return to later. The confidence ratings were obtained by having the listener adjust, with a computer mouse, a video pointer on a sliding scale. The pointer began each trial in the middle of the scale and could be adjusted continuously from *very confident* to *not at all confident* labels assigned to each end of the scale. Pointer readings were recorded at a rate of 50 per second with the first 10 s rejected, yielding a total of 2,500 readings per trial. The first 10 s were rejected so not to allow buildup of streaming to influence the data. However, to test whether buildup may have extended over longer durations mean confidence ratings were computed for the first and second half of each trial (after the first 10 s). The difference is mean ratings was negligible, 0.04 for streaming and 0.01 for masking. The average of the 2,500 readings was therefore taken as a single estimate of confidence and five such estimates were obtained for

each task and combination of Δ and σ. The final estimate of confidence for each condition was taken to be the mean of the five estimates after using the method of Thompson Tau to reject outliers (Thompson, 1985). Of the total 630 trial estimates obtained for all conditions and all listeners in the study, 20 (3%) were identified as outliers.

Subjects

A total of nine listeners (two men), with an average age of 23.9 (range 19–47) years, were recruited online from University of Wisconsin–Madison campus. All listeners had normal pure-tone hearing thresholds at the audiometric frequencies from 250 to 8000 Hz (ANSI3.6–2004, 2004) and were paid at an hourly rate for their participation. Prior to data collection, the listeners were given practice trials to familiarize them with the masking and streaming tasks. For the masking task, they were given a block of trials in which the A tones were absent. For the streaming task, they were given a block of trials in which Δ was 1,500 cents and σ was 0. In both cases, mean confidence ratings were near the *very confident* end of the scale. All procedures involving human subject recruitment and participation were performed in compliance with the University of Wisconsin–Madison Institutional Review Board guidelines.

Results

Effect of Δ and σ

Figure 2 shows the confidence ratings averaged across listeners for the streaming (top panel) and masking (bottom panel) tasks as a function of Δ with σ as parameter. Two of the nine listeners were not included in the average, as these listeners reported for all conditions 100% confidence hearing the level of the B tones alternate and the A and B tones split into separate streams. The pattern of results was much the same for the remaining seven listeners; hence, only the average data are presented. A two-way analysis of variance of ranks, Δ by σ, was performed separately for the data from the streaming and masking tasks. There were significant main effects of Δ and σ for both tasks. As expected, confidence ratings for streaming increased with increasing Δ ($F(2,294) = 121$, $p < 10^{-16}$) and decreased with increasing σ ($F(2, 294) = 18.8$, $p < 10^{-7}$). Also, as expected, confidence ratings for detection of the alternating level increased with increasing Δ ($F(2, 299) = 67.1$, $p < 10^{-16}$) and decreased with increasing σ ($F(2, 299) = 56.5$, $p < 10^{-16}$). There was also a significant interaction between Δ and σ, with the effect of σ being greatest for the intermediate value of Δ for both streaming and masking tasks: $F(4, 294) = 4.739$, $p < .00102$ and $F(4, 299) = 6.26$, $p < 10^{-4}$, respectively.

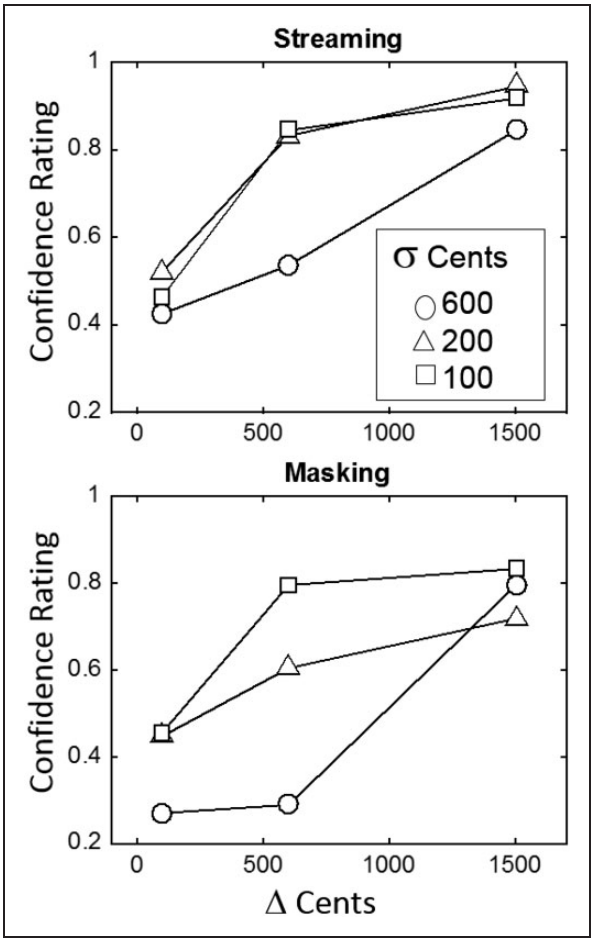

Figure 2. Confidence ratings averaged across seven listeners for the streaming (top panel) and masking (bottom panel) tasks plotted as a function of Δ. Different curves correspond to the different values of σ. Error bars not shown for clarity of presentation (see, instead, analysis of variance described in the Results section).

Effect of KS and DKL

The rough similarity in the effects of Δ and σ for streaming and masking, as seen in Figure 2, tends to support the hypothesis of a complementary relation between the two phenomena. However, a much stronger test is given by the prediction for the effects of *KS* and *DKL* as described in forgoing analysis of the Methods section. For the special case considered here, where the pdfs of A and B are equal variance normal (are symmetric), *KS* and *DKL* are monotonically related to one another; hence, we need only consider the predictions for *DKL*. (As previously noted, only the case for which p and q are asymmetric do *KS* and *DKL* make different predictions

for the streaming and masking tasks.) For normal pdfs given by p and q, the formula for DKL is

$$DKL(p\|q) = \frac{(\mu_A - \mu_B)^2 + \sigma_A^2 - \sigma_B^2}{2\sigma_B^2} + \ln\left(\frac{\sigma_B}{\sigma_A}\right) \quad (3)$$

For the present case, $\sigma_A = \sigma_B$, so that Equation 3 reduces to

$$DKL(p\|q) = \frac{1}{2}\left(\frac{\Delta}{\sigma}\right)^2 \quad (4)$$

The prediction then is that the confidence ratings for both streaming and masking tasks will be a common, monotonically increasing function of Δ/σ.

One approach to testing this prediction would be to fit the data separately with some expected function and then compare the parameters of these fits. The problem, however, with this approach is that it amounts to an attempt to accept the null hypothesis of no difference in parameters. The number of datum involved in each case (9) could very well not be enough to yield significant differences in parameter estimates where there are, in fact, real differences. A better approach, and the one taken here, is to fit a single function to all the data after adjusting for the difference in the overall difficulty of the two tasks. To this end, a logistic function was first fit separately to the streaming and masking data. The data were then adjusted to equate the intercepts of these fits as a measure of overall difficulty. A logistic function is a natural choice in this case as the confidence ratings are bounded between 0 and 1. The exact form of the function used was

$$P(DKL) = \frac{1-a}{1 + e^{-(\Delta/\sigma - b)/c}} \quad (5)$$

where the free parameters were the slope (c), intercept (b), and upper asymptote ($1-a$) of the function. The upper asymptote was included as a free parameter to reflect our subjects' general unwillingness to report a confidence rating of 100%. Using the nonlinear least squares method and the *fit* function of MATLAB (version r2014b), the fitted functions accounted for 91% and 78% of the total variance, respectively, for the streaming and masking data. For streaming, the estimated values of parameters were $a = 0.083$, $b = 0.143$, and $c = 0.456$; for masking, they were $a = 0.204$, $b = 0.311$, and $c = 0.645$. Figure 3 shows the mean confidence ratings of Figure 2 replotted as a function of Δ/σ for both the streaming (filled symbols) and masking (unfilled symbols) tasks. The masking data have been shifted upward along the y-axis by a constant amount so as to equate the intercepts of the separately fitted functions. The curve drawn through the data represents the function given by Equation 5 fitted to both sets of data simultaneously. The fitted function accounts for 96% of the total variance in the data with parameters $a = 0.082$, $b = 0.128$, and $c = 0.573$. The results tend to support the prediction for these conditions that the confidence ratings for both streaming and masking tasks can be described by a single, monotonically increasing function of Δ/σ.

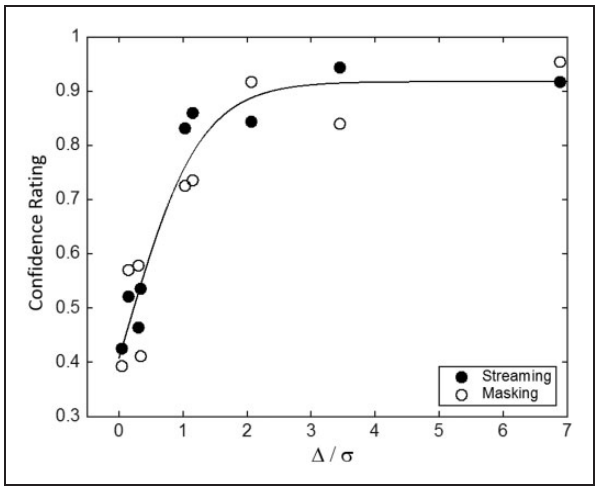

Figure 3. Mean confidence ratings of Figure 2 replotted as a function of Δ/σ for both the streaming (filled symbols) and masking (unfilled symbols) tasks. Confidence ratings for the masking task have been shifted upward by a constant amount to equate overall difficulty for the two tasks (see text for explanation). The curve shown is the nonlinear, least-squares fit of Equation 5 to the data. The fit accounts for 96% of the total variance in the data.

Discussion

Studies investigating the effects of target-masker similarity on masking have led to conjecture of a complementary relation between auditory streaming and masking, streaming being identified with release from masking (Dau, et al., 2009; Durlach et al., 2003; Kidd et al., 1994, 2002, 2003; Micheyl, Kreft, et al., 2013; Micheyl, Shamma, et al., 2013; Oh and Lutfi, 2000). The results of the present study lend support to this conjecture in showing (a) streaming and masking to be a common function of the information divergence (DKL) of tone sequences and (b) similarity and uncertainty effects to be conflated for both. This was the predicted outcome of a detection-theoretic analysis of listener judgments based on the premise that listeners approach the streaming and masking tasks as would an ideal observer. The agreement of the results with these predictions lends support to this basic premise and reinforces the theoretical approach taken here.

The theoretical approach taken here is, to the authors' knowledge, the first to offer an explicit analytic account

of the relation between streaming and masking. It is useful therefore to consider the implications of the analysis for the outcome of past studies that have taken a strictly methodological approach to the problem. We reviewed many of these studies in the introduction and the results generally agree with theory in showing a correspondence of performance in conditions of masking with expectations of perceptual streaming. Still, a simple correspondence provides only weak support. Moreover, only one of these studies complemented the masking results with subjective reports of streaming to confirm the expectations of streaming in the same or similar conditions. The one exception is the study of Micheyl and Oxenham (2010). These authors obtained reports of streaming in a temporal gap discrimination task involving two conditions, one for which perceptual segregation was expected to improve gap discrimination performance, and the other for which it was expected to impair performance. For the data averaged across listeners, the correlation of streaming reports with performance was statistically significant; however, the correlations were quite small. Moreover, for most *individual* listeners the correlations were not statistically significant due both to the small effect size and the large variability in the individual subject reports. The authors conclude that the correlations could not reliably be used to predict individual discrimination thresholds in their experimental conditions.

A different approach to the problem was taken by Lutfi and Liu (2011) and Richards, Carreira, and Shen (2012). These authors attempted to overcome the variability and bias associated with subjective reports by developing an *objective* measure of perceptual segregation, one that could be compared directly to measures of detection and discrimination performance in the same conditions without any additional data collection. Their methods differed in details but were the same in principle. The task in both cases was a masking task. Small perturbations were added to target and masker on each presentation along the dimension of the to-be-detected change in the target (object size for Lutfi and Liu and temporal position in a tone sequence for Richards et al.). Decision weights on the target and masker were then estimated from the correspondence between the perturbations and listener's trial-by-trial response using a regression model (cf. Berg, 1990; Lutfi, 1992). The inference regarding perceptual segregation was based on the decision weight for the masker. If the sign on this value was negative (i.e., different from that for the target), it could only mean that the listener heard the target separately (segregated) from the masker. If it was positive (same as that for target), it could only mean that the listener somehow confused (failed to segregate) target and masker. Using this approach, the authors found, like Micheyl and Oxenham (2010), that not all listeners show the expected correspondence between measured segregation and masking; indeed, some subjects whose masker weight was negative showed as much or even more masking than subjects whose masker weight was positive. Similar results using this method and involving a larger group of subjects have since been reported by Lutfi, Liu, and Stoelinga (2013).

Before concluding, we must offer two caveats regarding the interpretation of the present results. The first is to recognize that the ideal observer analysis offered here does no more than to generate an expectation for how listener judgments may be *related* in streaming and masking tasks. It was not intended, nor should it be taken, to be a specific model of listener performance in these tasks. A model of listener performance would require, at least, an internal noise parameter or some other means of accounting for failures of streaming where tone frequencies are fixed (no variance to bound the value of *KS*). It would require a prediction for how stimulus dimensions are weighted in the two tasks when stimuli differ along two or more dimensions simultaneously (cf. Lutfi, 1995). And, it would require for the masking task a model for how the change in target is to be detected, for which the present analysis says nothing. The analysis is perhaps better considered as a framework for developing a model rather than a specific model in and of itself. The second caveat has to do with experimental design. A major difference between traditional masking and streaming studies not captured by our experiment is the differential effect of response bias in the two tasks. In detection theory, a distinction is made between what the listener hears (sensitivity) and what they say they hear (response bias), the latter being influenced by the particular costs and rewards associated with the listener's response. Streaming studies make no attempt at separating the effects of these two factors on the listener's response but masking studies do. In masking studies, this is achieved by analyzing responses to no-signal trials for which a positive response would be scored as incorrect (see Green & Swets, 1966). In the current experiment, there are no no-signal trials and no incorrect responses. We purposely allowed bias to influence the masking results in the present design so that subjects, stimuli, and procedure would be identical for the both tasks; only the instructions given to listeners differing for the two tasks. We expected that if bias were to have any effect at all in the masking task, it would apply equally to all conditions in that task so that the function relating judgments to *DKL* would be unaffected. That the data are well described by a single function of *DKL* for both masking and streaming appears to support this conjecture.

Lastly, we offer some speculation regarding how the present results might be viewed in relation to past and possibly future work on the cocktail party problem.

We have emphasized the fundamental differences in the two approaches taken to the problem in the past, both in theory and in practice. Notwithstanding, the two approaches can be thought to exist at somewhat different ends of a continuum; streaming entailing the more *predictable*, more *discriminable* properties of individual talkers' speech that allows us to segregate one talker from another, while masking involving the less predictable, less discriminable properties that hinder this ability. The results of the present study reinforce this view by showing both streaming and masking to vary systematically along a single continuum that captures both of these properties in the value of *DKL*. In this way, the results also reinforce the connection between streaming and masking inferred in the literature. But what implications, if any, might the results have for future research? Notably, the past work has been mostly parametric in nature, documenting the degree to which specific *acoustic* differences between signals promote streaming or cause a release from masking. Although clearly much has been learned from these studies, the present results would advocate more strongly for an approach that emphasizes the *statistical* over the acoustical properties of signals, beginning of course with a comparison of the predictions based on *KS* and *DKL*. Particularly relevant to this point is the study by Lutfi, Gilbertson, et al. (2013). These authors describe masking experiments involving multitone pattern discrimination, multitalker word recognition, sound-source identification, and sound localization in which manipulations of masker uncertainty and target-masker similarity had the same effect on performance for the same change in Simpson-Fitter's *da*, an approximation to *DKL* for their conditions. They interpret their results to reflect a general principle of perception that exploits differences in the statistical structure of signals so as to separate figure from ground. The idea is by no means new (cf. Attneave, 1954; Barlow, 1961); but in light of the present results, it might serve as an impetus for research on cocktail party listening that attaches greater importance to the statistical properties of signals.

Acknowledgments

The authors would like to thank Dr. Andrew Oxenham, Dr. Alison Tan, and two anonymous reviewers for helpful comments on an earlier version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by National Institute on Deafness and Other Communication Disorders Grant No. 5R01DC001262–22.

References

ANSI S3.6-2004. (2004). *S3.6-2004, specification for audiometers*. Washington, DC: American National Standards Institute.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*, 183–193.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication*, *Ch. 13*, 217–234.

Bendixen, A., Bohm, T. M., Szalardy, O., Mill, R., Denham, S. L., & Winkler, I. (2013). Different roles of similarity and predictability in auditory stream segregation. *Learning & Perceptions*, *5*(Supplement 2), 37–54.

Bendixen, A., Denham, S. L., Gyimesi, K., & Winkler, I. (2010). Regular patterns stabilize auditory streams. *Journal of the Acoustical Society of America*, *128*(6), 3658–3666.

Berg, B. G. (1990). Observer efficiency and weights in a multiple observation task. *Journal of the Acoustical Society of America*, *88*(1), 149–158.

Bey, C., & McAdams, S. (2003). Postrecognition of interleaved melodies as an indirect measure of auditory stream formation. *Journal of Experimental Psychology*. *Human Perception and Performance*, *29*(2), 267–279.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound* (pp. 213–394). Cambridge, MA: MIT Press.

Chang, A.-C., Lutfi, R. A., & Lee, J. (2015). Auditory streaming of tones of uncertain frequency, level, and duration. *Journal of the Acoustical Society of America*, *138*(6), EL504–EL508.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*(5), 975–979.

Dau, T., Ewert, S., & Oxenham, A. J. (2009). Auditory stream formation affects comodulation masking release retroactively. *Journal of the Acoustical Society of America*, *125*(4), 2182–2188.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., & Kidd, G. Jr. (2003). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *Journal of the Acoustical Society of America*, *114*, 368–379.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Hartman, W. M., & Johnson, D. (1991). Stream segregation and peripheral channeling. *Music Perception*, *9*(2), 155–183.

Kidd, G. Jr., Mason, C. R., & Arbogast, T. L. (2002). Similarity, uncertainty, and masking in the identification of nonspeech auditory patterns. *Journal of the Acoustical Society of America*, *111*(3), 1367–1376.

Kidd, G. Jr., Mason, C. R., Deliwala, P. S., Woods, W. S., & Colburn, H. S. (1994). Reducing informational masking by sound segregation. *Journal of the Acoustical Society of America*, *95*(6), 3475–3480.

Kidd, G. Jr., Mason, C. R., & Richards, V. M. (2003). Multiple bursts, multiple looks, and stream coherence in the release

from informational masking. *Journal of the Acoustical Society of America*, *114*(5), 2835–2845.

Kidd, G. Jr., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). Informational masking. In W. A. Yost, & A. N. Popper (Eds.), *Springer handbook of auditory research: Auditory perception of sound sources* (pp. 143–190). New York, NY: Springer-Verlag.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79–86.

Lutfi, R. A. (1992). Comment on "analysis of weights in multiple observation tasks" [J. Acoust. Soc. Am. 86, 1743–1746 (1989)]. *Journal of the Acoustical Society of America*, *91*, 507–508.

Lutfi, R. A. (1993). A model of auditory pattern analysis based on component-relative-entropy. *Journal of the Acoustical Society of America*, *94*(2), 748–758.

Lutfi, R. A. (1995). Correlation coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks. *Journal of the Acoustical Society of America*, *97*(2), 1333–1334.

Lutfi, R. A., & Doherty, K. A. (1994). Effect of component-relative entropy on the discrimination of simultaneous tone complexes. *Journal of the Acoustical Society of America*, *96*(6), 3443–3450.

Lutfi, R. A., Gilbertson, L., Chang, A.-C., & Stamas, J. (2013). The information divergence hypothesis of informational masking. *Journal of the Acoustical Society of America*, *134*(3), 2160–2170.

Lutfi, R. A., & Liu, C. J. (2011). A method for evaluating the relation between sound source segregation and masking. *Journal of the Acoustical Society of America*, *129*(1), EL34–38.

Lutfi, R. A., Liu, C. J., & Stoelinga, C. N. J. (2008). Level dominance in sound source identification. *Journal of the Acoustical Society of America*, *124*(6), 3784–3792.

Lutfi, R. A., Liu, C. J., & Stoelinga, C. N. J. (2013). A new approach to sound source identification. In B. C. J. Moore, R. D. Patterson, I. M. Carlyon, & H. E. Gockel (Eds.), *Basic aspects of hearing: Physiology and perception* (pp. 203–211). New York, NY: Springer.

Micheyl, C., Kreft, H., Shamma, S., & Oxenham, A. J. (2013). Temporal coherence versus harmonicity in auditory stream formation. *Journal of the Acoustical Society of America*, *33*(3), EL188–E194.

Micheyl, C., & Oxenham, A. J. (2010). Objective and subjective psychophysical measures of auditory stream segregation and integration. *Journal of the Association for Research in Otolaryngology: JARO*, *11*, 709–724.

Micheyl, H., Shamma, S., & Oxenham, A. J. (2013). Hearing out repeating elements in randomly varying multitone sequences: A case of streaming? In B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, & J. Verhey (Eds.), *Hearing—From basic research to applications* (pp 267–274). Berlin, Germany: Springer.

Moore, B. C. J. (2002). Psychoacoustics of normal and impaired hearing. *British Medical Bulletin*, *3*(1), 121–134.

Moore, B. C. J., & Gockel, H. E. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1951), 919–931.

Noether, G. E. (1978). A brief survey of nonparametric statistics. In R. V. Hogg (Ed.), *Studies in statistics* (pp. 39–65). Washington, DC: Mathematical Association of America.

Oh, E., & Lutfi, R. A. (1998). Nonmonotonicity of informational masking. *Journal of the Acoustical Society of America*, *104*(6), 3489–3499.

Oh, E. L., & Lutfi, R. A. (1999). Informational masking by everyday sounds. *Journal of the Acoustical Society of America*, *106*(6), 3521–3528.

Oh, E. L., & Lutfi, R. A. (2000). Effect of harmonicity on informational masking. *Journal of the Acoustical Society of America*, *108*(2), 706–709.

Oxenham, A. J., & Dau, T. (2001). Modulation detection interference: Effects of concurrent and sequential streaming. *Journal of the Acoustical Society of America*, *110*(1), 402–408.

Richards, V. M., Carreira, E. M., & Shen, Y. (2012). Toward an objective measure for a "stream segregation" task. *Journal of the Acoustical Society of America*, *131*(1), EL8–13.

Szalardy, O., Bendixen, A., Bohm, T. M., Davies, L., & Denham, S. L. (2014). The effects of rhythm and melody on auditory stream segregation. *Journal of the Acoustical Society of America*, *135*(3), 1392–1405.

Thompson, R. (1985). A note on restricted maximum likelihood estimation with an alternative outlier model. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, *47*(1), 53–55.

van Norden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences* (Unpublished doctoral dissertation). University of Technology, Eindhoven, The Netherlands.

Vliegen, J., Moore, B. C. J., & Oxenham, A. J. (1999). The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. *Journal of the Acoustical Society of America*, *106*(2), 938–945.

Wang, D., & Brown, G. J. (2006). Fundamentals of computational auditory scene analysis. In D. Wang, & G. J. Brown (Eds.), *Computational auditory scene analysis, principles, algorithms, and applications* (pp. 81–111). Hoboken, NJ: Wiley-IEEE Press.

Wertheimer, M. (1938). Gestalt theory. In W. D. Ellis (Ed.), *A source book of gestalt psychology* (pp. 1–11). London, England: Routledge & Kegan Paul. (Original work published 1924).