Temporal predictability as a grouping cue in the perception of auditory streams

Vani G. Rajendran, Nicol S. Harper, and Benjamin D. Willmore
Department of Physiology, Anatomy and Genetics, University of Oxford, Sherrington Building, Parks Road, Oxford OX1 3PT, United Kingdom
vani.rajendran@univ.ox.ac.uk, nicol.harper@dpag.ox.ac.uk, benjamin.willmore@dpag.ox.ac.uk

William M. Hartmann
Department of Physics and Astronomy, Michigan State University, 567 Wilson Road, East Lansing, Michigan 48824
hartman2@msu.edu

Jan W. H. Schnupp
Department of Physiology, Anatomy and Genetics, University of Oxford, Sherrington Building, Parks Road, Oxford OX1 3PT, United Kingdom
jan.schnupp@dpag.ox.ac.uk

Abstract: This study reports a role of temporal regularity on the perception of auditory streams. Listeners were presented with two-tone sequences in an A-B-A-B rhythm that was either regular or had a controlled amount of temporal jitter added independently to each of the B tones. Subjects were asked to report whether they perceived one or two streams. The percentage of trials in which two streams were reported substantially and significantly increased with increasing amounts of temporal jitter. This suggests that temporal predictability may serve as a binding cue during auditory scene analysis.

PACS numbers: 43.66.Mk, 43.66.Ba, 43.66.Lj [QJF]
Date Received: March 29, 2013 **Date Accepted:** May 28, 2013

1. Introduction

The auditory system, like other sensory systems, receives a flood of often noisy stimuli from which it must parse and extract relevant information. Part of this process involves perceptually grouping sequential sounds that are likely to come from a single source, a process known as auditory streaming. A classic experimental paradigm used to investigate streaming uses sequences of two tones (A and B) that alternate in a simple repeating A-B-A-B-$\cdots$ pattern. When the frequency separation (ΔF) between the A and B tones is large enough, human listeners switch from perceiving a single, perceptually grouped A-B-A-B stream to perceiving two separate, perceptually segregated A----A$\cdots$ and --B----B$\cdots$ streams (van Noorden, 1975).

A number of sensory cues have been identified that the auditory system uses to either bind or segregate elements of acoustic input (Moore and Gockel, 2012; Hartmann and Johnson, 1991). Nevertheless, the principles and mechanisms governing stream segregation are not yet fully understood, and a number of competing hypotheses have been put forward. As sounds are more likely to separate into two streams when more widely separated in frequency, and given that the neural representation of frequency is tonotopically arranged, one hypothesis is that separation of neural populations along a neurotopic axis facilitates the formation of separate streams (Pressnitzer *et al.*, 2008). However, others (Sheft, 2007; Elhilali *et al.*, 2009) point out that tones that are well-separated in frequency are perceived as a single stream of complex tones if they are synchronous in time. On this basis, they propose "temporal coherence" (or

"average coincidence," i.e., the synchronous temporal modulation of neural activity across a feature space such as sound frequency) as a model for stream segregation (Shamma *et al.*, 2011). Finally, a third hypothesis, the predictive hypothesis, proposes that perceptual streams are formed or maintained by grouping together those aspects of a sound sequence that form predictable patterns (Bendixen *et al.*, 2010).

Theories on how the auditory system may extract regularities to form predictive representations of sensory input have been discussed in a number of reviews (Winkler *et al.*, 2009, 2012). These accounts emphasize the importance of time in auditory processing, especially the rate of alternation of tones. However, surprisingly little has been done in understanding the role of temporal regularity or predictability of the tones themselves. Andreou and colleagues (2011) reported that detecting a pattern in a temporally irregular tone sequence in the presence of a distracting tone sequence was easier if the distracting sequence was regular. They hypothesized that this was the case because it is easier to perceptually segregate the temporally regular distractor from the irregular target (Andreou *et al.*, 2011). In the experiments described here, we used a more direct approach—we measured the frequency of reporting a two-stream percept as a function of both frequency separation and temporal irregularity. We found that a two-factor logistic regression with both these parameters better explains the data than one that relies on frequency separation alone. Thus, we show that the likelihood that an A-B-A-B-$\cdots$ tone sequence will be perceived as two separate streams, rather than just a single stream, increases systematically with increasing temporal irregularity in the sequence of B tones.

2. Methods

The experimental methodology was approved by the local Ethical Review Committee of the Experimental Psychology Department of the University of Oxford, and conforms to the ethical standards laid down in the 1964 Convention of Helsinki.

Stimulus sequences consisted of one low-frequency tone and one high-frequency tone, referred to below as A and B, respectively, that alternated in an A-B-A-B-$\cdots$ pattern. All tones were 50 ms in duration, including 10-ms raised-cosine onsets and offsets. The frequencies of tones A and B were chosen to fall within the middle of the range of normal human hearing. In any one A-B-A-B-$\cdots$ sequence, the frequency of A was fixed at one of five values chosen uniformly from {891, 944, 1000, 1059, 1122} Hz. The frequency of the B tones was fixed at 1, 4, or 10 semitones above the frequency of A, yielding three different ΔF conditions.

The temporal pattern of the tone sequences is illustrated in Fig. 1. A and B tones could either alternate regularly (separated by 50-ms silent intervals), or have the same overall pattern and rate, but with a degree of temporal jitter, τ, added to the onset time of each B tone. The temporal jitter, τ, was chosen independently for each B tone from a uniform distribution over an interval $(-T,T)$, where T is the maximal jitter. Five different temporal jitter conditions were tested, with maximum jitter values T of {0, 4, 8, 16, or 32} ms. The temporal jitter, τ, thus had mean 0 and standard deviation (root-mean-square value) of $J = T/\sqrt{3}$. Throughout, we will use the standard deviation, J, as the quantitative measure of jitter. Since T never exceeded the 50-ms silence between neighboring tones in the regular pattern, the A and B tones never overlapped (see Fig. 1), thereby avoiding the confound of complex tones.

In our experiments, we tested responses to each permutation of the 3 frequency separations (ΔF) and 5 jitter conditions (J), resulting in 15 ΔF by J combinations. For each ΔF by J combination and each of the five possible A-tone frequencies, we created a 3-s long tone sequence. Each experimental subject was tested with a stimulus set comprising eight presentations of each of these sequences, presented in randomized order. Thus, each subject performed 40 trials for each ΔF by J combination. Tone sequences were generated and presented, and responses collected, using custom written MATLAB code. Sequences were presented diotically, at 60 dB sound pressure level, over Sennheiser (Wedemark, Germany) HD 650 headphones, which have a flat frequency response over

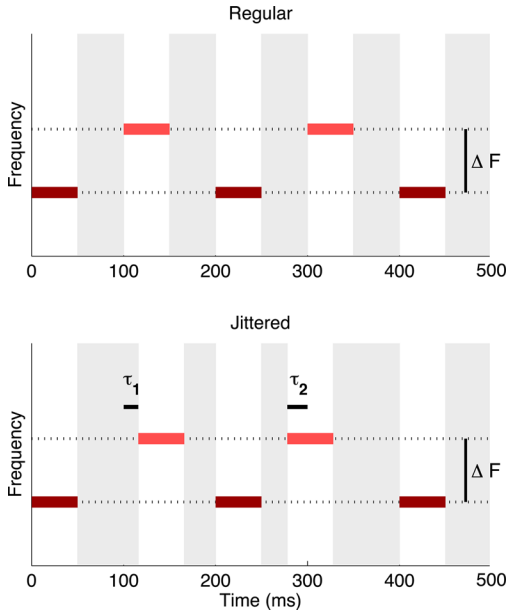

Fig. 1. (Color online) Schematic diagram of experimental paradigm. Stimulus sequences consisted of alternating tones at frequency A and B, separated by ΔF. B tones were flanked by 50-ms silent gaps (left). Temporal jitter was added to the B stream by shifting each B tone forward or backward in time by τ. Jitter τ was drawn from a uniform distribution bounded by maximum jitter parameter T for each B tone in a sequence.

the relevant frequency range. The sound sequences are available at http://www.auditoryneuroscience.com/streamingandjitter (last viewed 10 June 2013).

Experiments were conducted in a double-walled soundproof room. The eight participants tested comprised four naive subjects and four subjects involved with this study. Naive subjects first underwent an instructional period during which auditory stream segregation was explained and examples were played. One well-documented phenomenon with alternating tone sequences is the "build-up" effect of streaming, or the strong tendency for a tone sequence to be heard initially as a single stream, but then to break up into two separate streams, rapidly if the frequency separation is large, and more slowly or not at all if it is small (Anstis and Saida, 1985). Because of the build-up phenomenon and potential differences in task interpretation, subjects were consistently instructed, after each 3-s A-B-A-B-$\cdots$ sequence, to type the numeral "1" on a computer keyboard if they were able to perceive the tone sequence as a single stream throughout, or alternatively to type "2" if the percept had broken into two streams by the end of the sequence.

3. Results

Figure 2 illustrates, for each of the eight subjects, the percentage of reports of a two-stream percept for each stimulus condition. The different symbols (square, circle, triangle) represent data obtained at increasing ΔF. The amount of temporal jitter, quantified as the standard deviation of the distribution from which individual random time shifts, τ, were drawn, is plotted on the abscissa. Figure 2 demonstrates that, while individual subjects differed in how often they reported each stimulus condition as one or two streams, all eight subjects reported two streams on a larger fraction of trials in the conditions with greater amounts of temporal irregularity in the B tone sequence. Stream segregation also increased substantially with increasing ΔF, as was expected given that such dependence on ΔF has been reported in many previous studies (e.g., van Noorden, 1975; Pressnitzer *et al.*, 2008).

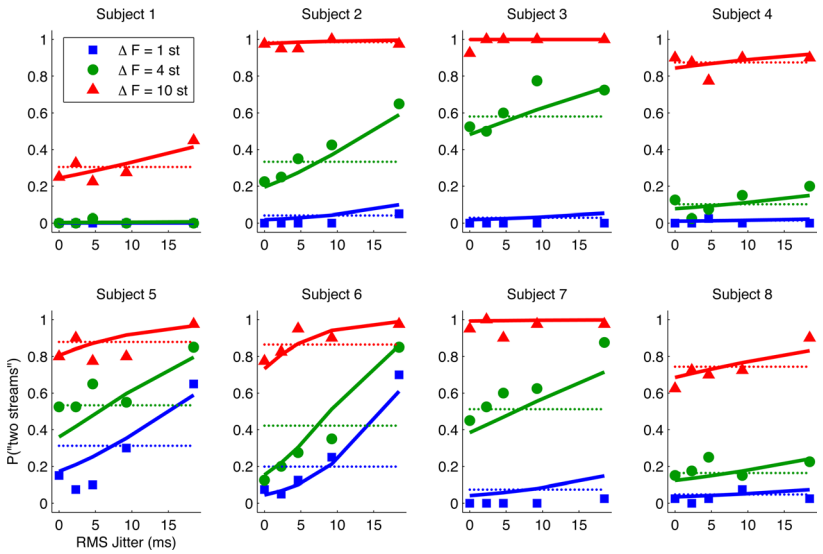

Fig. 2. (Color online) Symbols: raw data of observed fraction of trials (out of $n = 40$), where two streams were reported as a function of standard deviation of jitter J. Solid lines: fits of the simple logistic regression model of Eq. (1). Dotted lines: fits of the null model given by Eq. (2). Squares, circles, and triangles show data for $\Delta F = 1$, 4, and 10 semitones, respectively.

To quantify the influence of frequency separation and temporal jitter on perceived stream segregation, and to assess its statistical significance, we fitted a simple logistic regression model to the data. The form of the model is given by

$$\log\left(\frac{P}{1-P}\right) = b_{\text{offset}} + b_{\text{jitter}}J + b_{\Delta F}\Delta F. \quad (1)$$

Here, P is the fraction of trials in which two streams were reported. Hence, the fraction of one-stream reports must equal $1 - P$, and therefore the "odds ratio" of reporting two streams is $P/(1 - P)$. The logistic regression model assumes that the order of magnitude of this odds ratio [the "log-odds" $\log(P/(1 - P))$] grows proportionally with increasing jitter or frequency separation, and the effects of J and ΔF are assumed to be additive. The model parameter, b_{offset}, captures a subject's individual baseline propensity to hear two streams rather than one, and b_{jitter} and $b_{\Delta F}$ are the coefficients that capture the subject's sensitivity to jitter, J, and frequency separation, ΔF, respectively. They are estimates of the amount by which the log odds of perceiving two streams increases for each millisecond increase in temporal jitter or each semitone increase in the frequency separation, respectively.

The continuous lines in Fig. 2 show the fits of this logistic regression model to the observed data for each subject. It is clear that the fits are not perfect; for example, the model systematically overestimates the probability with which Subject 7 should perceive tone sequences separated by only one semitone as two streams. However, considering the great simplicity of the model (only strictly linear and additive effects, no interaction terms, all ΔF and J combinations described by one set of parameters for each subject), the fits do well in capturing the trends in the data.

Table 1 summarizes the coefficients and the standard errors of the best-fit coefficient estimates. It shows that both J and ΔF coefficients were positive for all subjects, i.e., all subjects were more likely to perceive two streams if either frequency separation or temporal irregularity of the B tone stream increased. We then sought to confirm whether the observed effects of increasing jitter (positive values for b_{jitter}) were statistically significant, both for each individual subject, and across the entire subject

Table 1. Best-fit (maximum likelihood) estimates of the coefficients (+/− standard error) of the logistic regression model in Eq. (1) for each subject.

Subject	b_{offset}	b_{jitter} (ms^{-1})	$b_{\Delta F}$ (semitone^{-1})
1	−8.86 (+/−1.64)	0.0424 (+/−0.0227)	0.773 (+/−0.164)
2	−4.86 (+/−0.406)	0.0966 (+/−0.021)	0.86 (+/−0.0788)
3	−5.29 (+/−0.595)	0.0597 (+/−0.0218)	1.31 (+/−0.146)
4	−5.27 (+/−0.426)	0.0407 (+/−0.0237)	0.696 (+/−0.0478)
5	−1.91 (+/−0.207)	0.105 (+/−0.0162)	0.333 (+/−0.0312)
6	−3.52 (+/−0.286)	0.19 (+/−0.0198)	0.453 (+/−0.0365)
7	−4.04 (+/−0.378)	0.0754 (+/−0.0197)	0.891 (+/−0.0868)
8	−3.8 (+/−0.303)	0.0444 (+/−0.0179)	0.458 (+/−0.0339)

population. One principled method for doing this is to perform a deviance test (Berry *et al.*, 2001) comparing the residual deviance of the best fit of the model given in Eq. (1) against that of a simpler model, which in this case represented the null hypothesis that jitter has no influence on streaming judgments. The simpler model posits that subjects' perceptions can be explained entirely by ΔF and random noise alone, as summarized by the simplified logistic regression model given by Eq. (2),

$$\log\left(\frac{P}{1-P}\right) = b_{\text{offset}} + b_{\Delta F}\Delta F. \tag{2}$$

The dashed lines in Fig. 2 show the fits of Eq. (2) to the data. For generalized linear models such as those given by Eqs. (1) and (2), the difference in residual deviance is chi-square distributed, with the number of degrees of freedom equal to the difference in the number of parameters of the two models to be compared (here, one). Thus, the statistical significance level of parameter b_{jitter} in Eq. (1) can be calculated as $1-\chi^2(D)$, where χ^2 is the cumulative chi-square distribution with one degree of freedom, and D is the difference between the residual deviances of the models given by Eqs. (1) and (2). Table 2 summarizes the residual deviances and the resulting p-value for the statistical significance of temporal jitter on the stream perception for each subject. In six of the eight subjects, the effect of jitter is highly statistically significant, and in the remaining two, the effect only narrowly misses statistical significance.

To test the significance of the influence of temporal jitter on the population as a whole, we set up further logistic regression models which pooled the responses from all subjects and treated subject identification as a factor variable. These models incorporated just one $b_{\Delta F}$ parameter, and one or zero b_{jitter} parameters, across all subjects.

Table 2. Deviance of logistic regression model with and without jitter from the raw data for each individual. The deviance when jitter was included was not significantly different from the deviance when jitter was excluded for two subjects. p-values were computed from the chi-square distribution. * indicates statistical significance at $p<0.05$; ** at $p<0.01$.

Subject	Deviance with jitter	Deviance without jitter	p-value	sig.
1	5.80	9.26	0.06	ns
2	19.12	41.28	$<10^{-5}$	**
3	42.63	50.49	0.005	**
4	14.10	17.09	0.084	ns
5	34.19	81.30	$<10^{-11}$	**
6	12.67	135.58	$<10^{-14}$	**
7	63.77	79.08	$<10^{-4}$	**
8	12.53	18.76	0.013	*

Comparing the model that included b_{jitter} against that without using the deviance test showed that, across the entire subject group, the influence of temporal jitter on stream perception was highly statistically significant with a $p < 10^{-16}$. The model fit to the whole population dataset yields best-fit parameter estimates which indicate that, on average and over the range of stimulus conditions tested, the log-odds of a subject hearing two streams grow with increasing ΔF at a rate of 0.5736 (+/−0.0165) per semitone, and with increasing temporal jitter at a rate of 0.0490 (+/−0.0038) per ms.

4. Discussion

We report the novel finding that increasing temporal irregularity between one set of tones in an A-B-A-B paradigm increases the likelihood that the sequence will segregate into two streams. Although temporal predictability has been suggested as relevant to streaming (Okada and Kashino, 2008; Winkler *et al.*, 2012), it has not yet been thoroughly explored. There has also been some study of the effect of jittering both sets of tones, where little effect was found (French-St. George and Bregman, 1989; Micheyl and Oxenham, 2010). Andreou *et al.* (2011) used a proxy measure of probability of detecting two streams—whether subjects succeeded in detecting an amplitude modulated pattern embedded in a target sequence of temporally irregular tones. They observed improved performance only when the temporally irregular target sequence of tones was presented together with a temporally regular distractor sequence. Another study similarly demonstrated that jittered familiar melodies are more easily recognized when a distractor tone sequence is isochronous rather than jittered (Devergie *et al.*, 2010). Our findings extend upon the above two results, but by taking the more direct approach of simply asking subjects to report whether they perceived one stream or two, as is commonly done in streaming psychoacoustics studies (Denham *et al.*, 2010; Pressnitzer *et al.*, 2008).

Such a direct approach might be considered less objective—a subject's responses will depend not only on the stimuli, but also on their interpretation of the instructions and internal criteria—but we designed our study to compensate for this potential confound. Trials with varying ΔF and jitter values were randomly interleaved and presented in rapid succession such that subjectivity of judgment criteria could be assumed to be either constant or, in the worst case, sources of random noise, which can be accommodated easily in the analytic framework of logistic regression. Thus, we were able to conclude that listeners indeed perceived two streams more frequently as the amount of temporal jitter increased, because a model that treats both jitter and ΔF as factors contributing to stream segregation accounts for the data significantly better than a model based on ΔF alone. This result raises the question of whether the "rhythmic separation" between the A and B tone sequences due to temporal jitter is more or less effective than frequency separation as a stream segregation cue. Our data do not speak to this because frequency separation in semitones is not directly comparable to rhythmic separation parameterized as temporal jitter in ms, but it may be possible to investigate this in the future if both frequency and rhythmic differences are measured in comparable units of psychoacoustic discriminability.

A further follow-up to this study would be to look at the effect of adding temporal jitter to both the A and B sequences. As mentioned above, some studies suggest that jittering both sequences produces little effect (French-St. George and Bregman, 1989; Micheyl and Oxenham, 2010). In the general streaming framework set forth by Moore and Gockel (2012), differences between sounds tend to induce stream segregation, and if A and B tone sequences are both jittered by the same amount, there would be no difference in regularity between them. However, the effect of the increased overall temporal unpredictability in this framework is unclear. It would be of interest to resolve this open question through a carefully designed study that examines the appropriate parameter range to determine the perceptual effect of jittering both tone sequences with varying amounts of jitter.

One explanation may be that temporal regularity itself is an overlooked but inherently salient feature to the auditory system that is quickly identified, resulting in the immediate grouping of the regular tones. Based on our findings, we propose that

temporal regularity is a form of predictability that tends to bind tone sequences into one stream, and adding temporal jitter violates to an increasing degree the expected pattern of regularity that would otherwise serve as a binding cue. The mechanisms behind this effect remain unclear, and it will be important to evaluate the performance of existing models of auditory streaming in light of our findings.

5. Conclusions

The results in this letter demonstrate that temporal regularity is a binding cue between tone sequences, and that this breaks down to an increasing degree as one sequence becomes more irregular in time. The auditory system constantly copes with temporal irregularities in natural auditory scenes. It remains to be seen whether this influence of temporal irregularity on streaming can be captured by any of the existing computational models. However, it is clear that temporal predictability plays an important role in the perception of auditory streams, and needs to be considered for any complete understanding of auditory scene analysis.

Acknowledgments

This work was supported by the Wellcome Trust and the Air Force Office of Scientific Research.

References and links

Andreou, L.-V., Kashino, M., and Chait, M. (**2011**). "The role of temporal regularity in auditory segregation," Hear. Res. **280**, 228–235.
Anstis, S., and Saida, S. (**1985**). "Adaptation to auditory streaming of frequency-modulated tones," J. Exp. Psychol. Hum. Percept. Perform. **11**, 257–271.
Bendixen, A., Denham, S. L., Gyimesi, K., and Winkler, I. (**2010**). "Regular patterns stabilize auditory streams," J. Acoust. Soc. Am. **128**, 3658–3666.
Berry, G., Matthews, J. N. S., and Armitage, P. (**2001**). *Statistical Methods in Medical Research* (Blackwell Science Inc., Oxford), pp. 488–495.
Denham, S. L., Gyimesi, K., Stefanics, G., and Winkler, I. (**2010**). "Stability of perceptual organisation in auditory streaming," in *The Neurophysiological Bases of Auditory Perception*, edited by E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis (Springer, New York), pp. 477–488.
Devergie, A., Grimault, N., Tillmann, B., and Berthommier, F. (**2010**). "Effect of rhythmic attention on the segregation of interleaved melodies," J. Acoust. Soc. Am **128**, EL1–EL7.
Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., and Shamma, S. A. (**2009**). "Temporal coherence in the perceptual organization and cortical representation of auditory scenes," Neuron **61**, 317–329.
French-St. George, M., and Bregman, A. S. (**1989**). "Role of predictability of sequence in auditory stream segregation," Percept. Psychophys. **46**, 384–386.
Hartmann, W. M., and Johnson, D. (**1991**). "Stream segregation and peripheral channeling," Music Percept. **9**, 155–183.
Micheyl, C., and Oxenham, A. J. (**2010**). "Objective and subjective psychophysical measures of auditory stream integration and segregation," J. Assoc. Res. Otolaryngol. **11**, 709–724.
Moore, B. C. J., and Gockel, H. (**2012**). "Properties of auditory stream formation," Philos. Trans. R. Soc. London, Ser. B **367**, 919–931.
Okada, M., and Kashino, M. (**2008**). "Temporal jitter affects perceptual transitions in auditory streaming," in *Annual Meeting of the Association for Research in Otolaryngology*, Phoenix, Arizona.
Pressnitzer, D., Sayles, M., Micheyl, C., and Winter, I. M. (**2008**). "Perceptual organization of sound begins in the auditory periphery," Curr. Biol. **18**, 1124–1128.
Shamma, S. A., Elhilali, M., and Micheyl, C. (**2011**). "Temporal coherence and attention in auditory scene analysis," Trends Neurosci. **34**, 114–123.
Sheft, S. (**2007**). *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer, New York).
van Noorden, L. (**1975**). "Temporal coherence in the perception of tone sequences," Doctoral dissertation, Eindhoven University of Technology, Leiden (The Netherlands).
Winkler, I., Denham, S., Mill, R., Bohm, T. M., and Bendixen, A. (**2012**). "Multistability in auditory stream segregation: A predictive coding view," Philos. Trans. R. Soc. London, Ser. B **367**, 1001–1012.
Winkler, I. N., Denham, S. L., and Nelken, I. (**2009**). "Modeling the auditory scene: Predictive regularity representations and perceptual objects," Trends Cogn. Sci. **13**, 532–540.