A method for evaluating the relation between sound source segregation and masking

Robert A. Lutfi^{a)} and Ching-Ju Liu

Auditory Behavioral Research Laboratory, University of Wisconsin, Madison, Wisconsin 53706
ralutfi@wisc.edu, chingjuliu@wisc.edu

Abstract: Sound source segregation refers to the ability to hear as separate entities two or more sound sources comprising a mixture. Masking refers to the ability of one sound to make another sound difficult to hear. Often in studies, masking is assumed to result from a failure of segregation, but this assumption may not always be correct. Here a method is offered to identify the relation between masking and sound source segregation in studies and an example is given of its application.

© 2011 Acoustical Society of America

PACS numbers: 43.66.Ba, 43.66.Dc [QJF]

Date Received: September 28, 2010   **Date Accepted:** October 27, 2010

1. Introduction

One of the greatest challenges for contemporary research in hearing is to understand how we are able to identify individual sound sources in a complex mixture of sounds. The sound reaching our ears is the sum of the sound waves emitted by each source, so to identify any one source we must somehow extract from this sum the individual sound elements belonging to that source. One way the perceptual system is thought to achieve this goal is by “segregating” from the mixture those elements that share common acoustic properties with one another, for example, have coincident onsets, are harmonically related, or covary in frequency and amplitude over time. Perceptual segregation based on such common features is recognized as a manifestation of the Gestalt principle of *common fate* (Bregman, 1990, pp. 213–393; Hukin and Darwin, 1995). It is evidenced by a clear subjective impression of the sound as “standing out” from the rest of the mixture. Masking, by comparison, is an objective measure. It relates to the accuracy with which a listener is able to detect or identify one sound (the target) in the presence of another (the noise). Often in the literature an association is assumed between masking and perceptual segregation wherein masking is taken to reflect the listener’s inability to segregate the target from the noise (Kidd *et al.*, 1994; Durlach, 2006). In fact, the inability to segregate target from noise is neither a necessary nor sufficient condition for masking. A noise that fails to be segregated from the target may produce little or no masking if the target-plus-noise percept is easily distinguished from the noise alone. Conversely, a perceptually distinct noise can cause significant masking if it suppresses the target (Sachs and Kiang, 1968; Houtgast, 1972), distracts attention away from the target (Werner and Bargones, 1991; Carlyon and Moore, 1986), or serves as a referent for perception of the target as in a figure–ground relation (Green, 1988, pp. 3–21). One method that has been used to evaluate the relation between segregation and masking requires listeners to report upon their subjective impression of the target and noise as they participate in different conditions of masking (Micheyl and Oxenham, 2010). This approach can provide valuable insights into the causes of masking, but it can also be unreliable as subjective reports are known to be variable and sometimes difficult to interpret. The present paper describes an alternative method that does not require subjective reports or, for that matter, any other form of additional data collection to evaluate the role of segregation in masking

---

^{a)}Author to whom correspondence should be addressed.

experiments. The method is simple to implement and provides a reliable quantitative measure of the degree perceptual segregation that can be related directly to the amount of masking observed in different experimental conditions.

2. Method

The proposed method amounts to a specific application of the conditioned-on-a-single-stimulus (COSS) procedure described by Berg (1990). The COSS procedure and its associated analysis were originally developed by Berg to be used in multi-tone pattern discrimination studies where one wishes to measure the relative decision weight listeners give to different tones in the pattern. The procedure involves introducing small random perturbations in the tones from trial to trial along the dimension to be discriminated. The decision weights are then given by regression coefficients in a general linear model for which the perturbations on the individual tones are predictor variables for the listener's trial-by-trial response. The present application proceeds in much the same way, except that the stimulus is taken to be comprised of only two elements, target and noise, and only one decision weight is computed, the weight on the noise relative to that of the target. The logic of this method is that if masking results from a failure to segregate target from noise, the decision weight on the noise must be positive, like that for the target. A zero or negative decision weight on the noise can only occur if the listener, at least in part, perceives the noise as being separate from the target.

In what follows, the method and analysis are demonstrated by way of example. The listener's task was to identify as target the impact sound produced by the larger of two stretched membranes. The noise on each presentation was the impact sound of a variable-sized plate. In different conditions the noise was gated on simultaneously with the target or after a delay. Performance in every case was poorer when the target and noise were gated on simultaneously consistent with the principle of common fate. However, the decision weights on the noise for some listeners suggested an alternative explanation.

2.1 Stimuli

The stimuli were approximations to the impact sounds of a stretched circular membrane (target) and loosely suspended circular plate (noise). They were synthesized and presented via headphones using first-order analytic equations for the motion of these sources from standard acoustic texts (Morse and Ingard, 1968; pp. 175–221). The resultant impact sounds were a sum of exponentially damped sinusoids representing the individual partials of the sounds. For the membrane the frequencies of the partials were $f_{1M} \times [1.00 \ 1.594 \ 2.136 \ 2.296 \ 2.653 \ 2.918]$; for the plate they were $f_{1P} \times [1.00 \ 2.80 \ 5.15 \ 5.98 \ 9.75 \ 14.09]$. The values of f_{1M} and f_{1P} varied from one presentation to the next, as described in the next section. The power of the first partial was fixed across trials and equated for both sources; the decay modulus of the first partial was fixed at 0.2 s for the membrane and 0.4 s for the plate. Both the power and decay moduli of higher-numbered partials decreased from the first partial proportionally with frequency. So as to keep the trial sequence at a reasonable length, a 5-ms cosine-squared ramp was used to truncate the sounds after 1 s. Sounds were played at a 44 100-Hz sampling rate with 16-bit resolution using a Mark of the Unicorn (MOTU) 896 audio interface. From the interface the sounds were buffered through a Rolls RA62c headphone amplifier and then delivered diotically to listeners over Beyerdynamic DT 990 headphones. A loudness balancing procedure was used (see Lutfi *et al.*, 2008) to calibrate the overall sound power to be approximately 70 dB sound pressure level (SPL) at the eardrum. Listeners were seated individually in a double-walled, Industrial Acoustics Company (IAC) sound-attenuated chamber.

2.2 Procedure

In different conditions the time interval between the onset of the target (membrane) and the noise (plate) was fixed at $\tau = 0$ ms (coincident onsets) or fixed at $\tau = 100$ ms

(target following noise). In each condition the values of f_{1M} and f_{1P} in Hz varied independently and at random on each presentation, as would correspond to changes in the size of the membrane and plate. The specific values were

$$f_{1M} = 500 \cdot \exp(10 \cdot \text{jnd} \cdot z_M), \quad (1a)$$

$$f_{1P} = 250 \cdot \exp(50 \cdot \text{jnd} \cdot z_P), \quad (1b)$$

where $\text{jnd} = \log(1.002)$ is an estimate of the just-noticeable difference in frequency (Wier *et al.*, 1977) and z is a random normal deviate selected independently for the membrane and plate on each presentation. As earlier described, the frequencies of higher-number partials were scaled in each case according to f_{1M} and f_{1P} . In the two-interval forced-choice procedure with feedback, listeners were instructed to select the interval containing the impact sound corresponding to the larger membrane (lower f_{1M}). Note that this procedure differs from standard procedures where the task is to detect a fixed increment in the target. Conditions were run in random order on different days, each day consisting of a single 1-h experimental session of eight blocks of trials, 50 trials per block. Listeners were 1 female and 2 male students of the University of Wisconsin—Madison, aged 25–32 yr. Other listeners had participated in this experiment, but we chose to show the data from these three as they were the most revealing in the application of the method. All listeners had extensive previous experience with the impact sounds in similar two-interval, forced-choice tasks. The results of a standard hearing evaluation showed all to have normal hearing sensitivity of 15 dB hearing level (HL) or better from 250 Hz to 8 kHz (ANSI S3.6-1996).

2.3 Analysis

The general application of our method requires only that the noise be perturbed from one presentation to the next along the same dimension for which the target is judged (object size in the present case). Let X_T and X_N denote the perturbed value along this dimension, respectively, for target and noise. The listeners relative decision weight on the noise is then evaluated using a general linear model in which the probability of a first-interval response, $P(R = 1)$, is given by

$$\log[P(R = 1)] = (1 - |w|)\Delta(X_T + e_T) + w\Delta(X_N + e_N) + e, \quad (2)$$

where Δ is the difference between the first and second intervals (first minus second), the e are various sources of internal noise, and w is the decision weight on the noise relative to the target. Now, if the e are taken to be independent of X_T and X_N , they can be pooled to yield a single error term, ε . Equation (2) then provides a general logistic regression model for estimating w in which ε is the regression error. The values of w are obtained by regressing the listener's trial-by-trial response on the trial-by-trial values of ΔX_T and ΔX_N using the `glmfit` routine of MATLAB v.7.0.1. Resulting values of $-0.5 \leq w \leq 0$ in this case are taken as evidence that the listener is able to segregate target and noise. Resulting values $0 < w \leq 0.5$ are taken as evidence of a failure to segregate, with $w = 0.5$ representing the extreme case wherein the listener effectively makes no distinction between target and noise.

3. Results and conclusions

As in past applications of COSS analysis, the general linear model provided excellent fits to the trial-by-trial data (Deviance < 342 for $df = 398$). We focus, therefore, on the obtained values of w . Figure 1 gives for each listener (symbol shape) the obtained w and corresponding d' for $\tau = 0$ ms (filled symbols) and $\tau = 100$ ms (unfilled symbols). Error bars give the 95% confidence intervals and continuous curves give the upper and lower bounds on the data corresponding to the best performance achievable given a particular value of w and zero internal noise (zero ε). The results show all three listeners

Fig. 1. The relative decision weight on the noise, w , and the corresponding performance level, d' , for $\tau = 0$ ms (filled symbols) and $\tau = 100$ ms (unfilled circles) are plotted for each listener (symbol type). Error bars represent the 95% confidence intervals. Continuous curves give the upper and lower bounds on the data corresponding to zero internal noise in Eq. (2). Data falling on or below the horizontal line provide evidence of perceptual segregation of target and noise.

to obtain a significant release from masking when the onset of the target is delayed relative to onset of the noise. This is the kind of result that in past studies has been interpreted as evidence for source segregation; i.e., a stimulus manipulation imposed that is expected to enhance segregation and the manipulation results in a release of masking. In the present case, however, the decision weights on the noise tell a different story. For only one of the three listeners (circles), the weights are positive when the target is gated on simultaneously with the noise. For the other two listeners the weight is negative or does not significantly deviate from zero. Where the weight does not deviate from zero poor performance appears to result from a tendency of the noise to deflect attention away from the target without itself being given a positive weight. Where the weight is negative poor performance reflects a figure-ground comparison of target and noise for which a large membrane (low f_{1M}) tends to appear smaller in the presence of a larger plate (lower f_{1P}). Clearly, these two listeners are able to use other differences between target and noise to segregate target and noise when their onsets are simultaneous. These are cases where, based on performance accuracy alone, one might have come to a false conclusion regarding the ability to segregate.

Rarely in studies of sound source segregation do target and noise differ along only one dimension. It is often possible, therefore, that listeners may segregate target from noise in ways that are not evident from measures of performance accuracy alone. In such cases, the present method of estimating listener decision weights on the noise provides a means of evaluating the relation between segregation and masking that does not depend on subjective report. The method requires no additional data collection and reliable estimates of the decision weights can be obtained without large numbers of trials (Lutfi, 1995).

Acknowledgment

We would like to thank Dr. Brian C.J. Moore and two anonymous reviewers for helpful comments on a previous version of this manuscript. This research was supported by NIDCD (National Institutes of Deafness and Other Communication Disorders) Grant No. 5R01DC006875-05.

References and links

- ANSI S3.6-1996 (1996). *American National Standard Specification for Audiometers* (American National Standards Institute, New York).
- Berg, B. G. (1990). "Observer efficiency and weights in a multiple observation task," *J. Acoust. Soc. Am.* **88**, 149–158.
- Bregman, A. S. (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA), pp. 213–393.
- Carlyon, R. P., and Moore, B. C. J. (1986). "Continuous versus gated pedestals and the 'severe departure' from Weber's law," *J. Acoust. Soc. Am.* **79**, 453–460.
- Durlach, N. (2006). "Auditory masking: Need for improved conceptual structure," *J. Acoust. Soc. Am.* **120**, 1787.
- Green, D. M. (1988). *Profile Analysis: Auditory Intensity Discrimination* (Oxford University Press, Oxford, England), pp. 3–21.
- Houtgast, T. (1972). "Psychophysical evidence for lateral inhibition in hearing," *J. Acoust. Soc. Am.* **51**, 1885–1894.
- Hukin, R. W., and Darwin, C. J. (1995). "Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification," *Percept. Psychophys.* **57**, 191–196.
- Kidd, G., Jr., Mason, C. R., Deliwala, P. S., Woods, W. S., and Colburn, H. S. (1994). "Reducing informational masking by sound segregation," *J. Acoust. Soc. Am.* **95**, 3475–3480.
- Lutfi, R. A. (1995). "Correlation coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks," *J. Acoust. Soc. Am.* **97**, 1333–1334.
- Lutfi, R. A., Liu, C. J., and Stoelinga, C. N. J. (2008). "Level dominance in sound source identification," *J. Acoust. Soc. Am.* **124**, 3784–3792.
- Micheyl, C., and Oxenham, A. J. (2010). "Objective and subjective psychophysical measures of auditory stream segregation and integration," *J. Assoc. Res. Otolaryngol.* **11**(4), 709–724.
- Morse, P. M., and Ingard, K. U. (1968). *Theoretical Acoustics* (Princeton University Press, Princeton, N.J.), pp. 175–191.
- Sachs, M. B., and Kiang, N. Y. S. (1968). "Two-tone inhibition in auditory-nerve fibers," *J. Acoust. Soc. Am.* **43**, 1120–1128.
- Werner, L. A., and Bargones, J. Y. (1991). "Sources of auditory masking in infants: Distraction effects," *Percept. Psychophys.* **50**, 405–412.
- Wier, C. C., Jesteadt, W., and Green, D. M. (1977). "Frequency discrimination as a function of frequency and sensation level," *J. Acoust. Soc. Am.* **61**, 178–184.