Published in final edited form as:

Adv Exp Med Biol. 2013 ; 787: 203–211. doi:10.1007/978-1-4614-1590-9_23.

A New Approach to Sound Source Segregation

Robert A. Lutfi,

Auditory Behavioral Research Laboratory, Department of Communicative Disorders, University of Wisconsin, Madison, WI 53706, USA

Ching-Ju Liu, and

Auditory Behavioral Research Laboratory, Department of Communicative Disorders, University of Wisconsin, Madison, WI 53706, USA

Christophe N.J. Stoelinga

Auditory Behavioral Research Laboratory, Department of Communicative Disorders, University of Wisconsin, Madison, WI 53706, USA

Abstract

We rely critically on our ability to ‘hear out’ (*segregate*) individual sound sources in a mixture. Yet, despite its importance, little is known regarding this ability. Perturbation analysis is a psychophysical method that has been successfully applied to related problems in vision (Murray RF. *J Vision* 11:1–25, 2011). Here the approach is adapted to audition. The application proceeds in three stages: First, simple speech and environmental sounds are synthesized according to a generative model of the sound-producing source. Second, listener decision strategy in segregating target from nontarget (noise) sources is determined from decision weights (regression coefficients) relating listener judgments regarding the target to lawful perturbations in acoustic parameters, as dictated by the generative model. Third, factors limiting segregation are identified by comparing the obtained weights and residuals to those of a maximum-likelihood (ML) observer that optimizes segregation based on the equations of motion of the generating source. Here, the approach is applied to test between the two major models of sound source segregation: target enhancement versus noise cancellation. The results indicate a tendency of noise segregation to preempt target enhancement when the noise source is unchanging. However, the results also show individual differences in segregation strategy that are not evident in the measures of performance accuracy alone.

1 Introduction

We possess a remarkable ability to ‘hear out’ and identify individual sound sources in everyday, multisource acoustic environments. Understanding this ability has been a preoccupation of psychoacoustic research, but it is a tremendous challenge (cf. Bregman 1990). The pressure waveform arriving at the ear is a confluence of information from different sound sources, so determining precisely how a listener is able to identify each source in the mixture requires sophisticated methods. The approach to date has largely been

based on a traditional *performance model* – this is the view that changes in performance accuracy or threshold in different conditions can be used to infer the listener's decision processes in those conditions. The approach is widely practiced, but it can easily lead to false conclusions regarding the data. Recent studies involving the identification of simple sound sources in quiet have shown that listeners can and regularly do adopt different listening strategies that yield the same performance accuracy and, conversely, that the same listening strategy often produces very different levels of performance (Lutfi and Liu 2007, 2011a, b). These studies call into question the conclusions that can be drawn regarding listening strategy from performance measures alone, and they underscore the challenge of understanding listening in multisource acoustic environments. Here we undertake an alternative approach that overcomes some of the limitations of the performance model.

2 Perturbation Analysis

The approach amounts to a variation in methods that have been used to determine listener decision strategy in multitone pattern discrimination studies (Berg 1989). The procedure, referred to as *perturbation analysis*, involves introducing small random perturbations in the tones from trial to trial along the dimension to be discriminated. Listener decision strategy is then assessed from estimates of decision weights on the tones given by regression coefficients in a general linear model for which the perturbations are predictor variables for the listener's trial-by-trial response. The present application proceeds in much the same way, but with two important differences. First, the stimulus is taken to be comprised of only two elements, target and noise. The listener's decision strategy for segregating target from noise is determined from the relative decision weights given to the different material, geometric and/or driven properties that serve to distinguish target from noise. Second, the target and noise are either speech sounds or the impact sounds of simple resonant sources. The perturbations in acoustic parameters are applied indirectly by perturbing the physical parameters of a mechanical model of these sources used to synthesize the sounds (Klatt 1980; Morse and Ingard 1968; pp. 175–221). This modification provides the variation in acoustic parameters necessary to derive listener decision weights without violating the lawful relations that exist among these parameters. It also prevents the listener from performing better than chance by simply discriminating a change in a fixed acoustic waveform, what would otherwise be described as a simple discrimination or detection task.

The goal in each case is to determine precisely *how* the listener makes use of the information in sound to segregate target from noise. This is achieved by formulating a general decision model of listener's trial-by-trial judgments regarding some physical property (or properties) of the target and then estimating the parameters of this model by regression. Here we demonstrate using a simple example from a previously published study (Lutfi and Liu 2011b). The listener's task was to identify as target the impact sound produced by the larger of two circular stretched membranes (the mechano-acoustic analogy to speech would be judging the gender of a speaker by vocal tract length; cf. Patterson et al. 2008). The noise was the impact sound of a variable-size, circular plate presented simultaneously with the target. Let p_T and p_N denote the perturbation in size on each trial, respectively, for target and noise (normally distributed with zero mean). A general decision rule for this task is then given by respond 'larger target' if and only if

$$w_T(\mu_T + p_T + e_T) + w_N(\mu_N + p_N + e_N) > \beta + e_\beta \quad (23.1)$$

where w_T and w_N are, respectively, the decision weights on the target and the noise; μ_T and μ_N denote the nominal sizes of target and noise (before perturbation); β is a response criterion; and e_T , e_N , and e_β are zero-mean normal deviates representing various imperfections in the decision process (internal noise). The internal noise terms are taken to be independent of p_T and p_N and are so pooled to yield a single error term, e (a step justified by the fit of the model). The parameters of the decision rule are then estimated by regressing the listener's trial-by-trial response on the values of p_T and p_N , where the influence of internal noise is estimated from the regression error, σe . In practice, the regression is performed using the `glmfit` routine of MATLAB v.7.0.1 with logistic link function,

$$\text{logit}[P(\text{Larger})] = c_0 + c_1 p_T + c_2 p_N + e, \quad (23.2)$$

where $P(\text{Larger})$ is the probability that the target presented is identified as the larger of the two and c_0 is an estimate of β . (Note that the regression is performed on the type of response, not on whether the response is correct or incorrect.) The regression coefficients c_1 and c_2 are finally used to estimate the decision weight on the noise *relative* to that of the target,

$$\hat{w}_N = \frac{c_2}{c_1 + |c_2|} \approx \frac{w_N}{w_T + |w_N|}, \quad (23.3)$$

where $c_1 > 0$. The approach allows for separate evaluation of the effects of the decision weights and internal noise on performance and so provides a stronger test of listener decision strategy than an overall performance metric alone can provide. It also allows the decision weights to be evaluated in different conditions relative to that of a maximum-likelihood (ML) observer that maximizes performance. For the example described here, the decision weight on the noise for the ML observer is $w_2 = 0$.

3 Application: Target Enhancement Versus Noise Cancellation

Here we consider a specific application of the approach designed to test different models of sound source segregation in noise, noise in this case referring to other sound sources. These models generally fall into two classes: those based on enhancement of the target and those based on cancellation of the noise (Bregman 1990; de Cheveigné et al. 1995; Durlach 1963; Meddis and Hewitt 1992; Piechowiak et al. 2007). Much of the work aimed at testing these models has been done on the identification of concurrent pairs of harmonic and inharmonic vowels (see de Cheveigné et al. (1995) for a review). This work would seem to support a predominant role of noise cancellation; however, the tests are based on measures of performance accuracy which, as noted, may not best reflect the processes underlying segregation. In the present study, the relative contribution of target enhancement and noise cancellation is determined by the sign and magnitude of listener weights on the noise.

The listener's task was size discrimination as in the example described above. A two-interval forced-choice procedure with feedback was used. The impact sounds of the membrane (target) and plate (noise) were synthesized using first-order analytic equations for the motion of these sources (Morse and Ingard 1968). The result in each case was a sum of exponentially damped sinusoids. For the membrane the frequencies were 500, 797, 1,068, 1,148, 1,327, and 1,459 Hz; for the plate they were 250, 700, 1,287, 1,495, 2,437, and 3,522 Hz. The amplitudes and decay moduli decreased proportionally with frequency beginning with a decay modulus of 0.2 s for the membrane and 0.4 s for the plate. The frequencies were perturbed from one presentation to the next as would correspond to changes in the size. The perturbations were normally distributed in just-noticeable (jnd) units, with one jnd being equal to $\log(1.002)$ (Wier et al. 1977). The standard deviation of perturbations for the plate was fixed at 10 jnds; for the membrane it took on values $\sigma = 10\text{--}80$ jnds, somewhat different for each listener. Sounds were played at a 44,100-Hz sampling rate with 16-bit resolution and were calibrated to be approximately 70 dB SPL at the eardrum (see Lutfi et al. 2008). They were delivered diotically over beyerdynamic DT 990 headphones to listeners seated individually in a double-walled, IAC sound-attenuated chamber. Listeners were five students of the University of Wisconsin-Madison, aged 24–36 year. All had normal hearing sensitivity (ANSI S3.6-1996) and extensive previous experience with the task.

Now consider the predictions of the two classes of models for the noise weights in Eq. (23.1). Fundamentally, all noise cancellation models detect or recognize the target as a change from a baseline established by the noise. Probably, the simplest example is the old-versus-new (figure-ground) heuristic of Bregman (1990). Models that implement this process computationally involve some form of noise equalization followed by subtraction (e.g., Durlach 1963). For the present application, this can be shown algebraically to be equivalent to giving a negative weight to the noise in Eq. 23.1. We expect then that $-1 < \hat{w} < 0$. Target enhancement models, by comparison, attribute noise interference to imperfections in the enhancement process that result in partial enhancement of the noise. A common example is the auditory filter model for tone-in-noise detection (Patterson and Nimmo-Smith 1980). These models predict the weight on the noise to be positive, so that we expect $0 < \hat{w} < 1$. (The analytic proof for both predictions is given fully in the recent publication of Lutfi and Liu (2011b).) Note that the decision model given by Eq. (23.1) also allows for the evaluation of a third alternative in which the noise simply serves to distract attention away from the target, without itself being given any weight (cf. Lutfi and Wightman 1996; Carlyon and Moore 1986; Werner and Bargones 1991). In this case $\hat{w}_N$ is expected to be near the optimal value of zero with less than optimal performance dictated by an increase in internal noise e resulting from the distraction.

Figure 23.1 gives the estimated noise decision weight, $\hat{w}_N$, for each listener (symbol type) as a function of performance level in d' , each point representing 400 trials. Repeated symbols are for different values of σ_T . The horizontal and dashed lines are the predictions of models as labeled. For four of five listeners, the weights are positive, indicating target enhancement; however, for the remaining listener, they are negative, indicating noise cancellation. In a second condition, involving a different group of listeners, the task was to choose the membrane hit with harder force; the noise was a plate hit with variable force. The noise weights for all listeners in this condition indicated noise cancellation (Fig. 23.2). With the

exception of the one listener in Fig. 23.1, the results from Figs. 23.1 and 23.2 suggest that noise cancellation occurs whenever the noise source is unchanging (force of impact condition). This would make sense since for a constant noise source, only the amplitudes of modes change; the frequencies of the modes are fixed. Such conditions would be expected to allow for the most effective noise equalization and cancellation, perhaps through adaptation or inattention. We will refer to this as the *fixed-noise hypothesis*; it is fundamentally a version of the old-versus-new heuristic described by Bregman (1990).

A subsequent study was undertaken to further test this hypothesis. The impact sounds were replaced with synthesized vowel sounds (Klatt 1980). A new group of listeners was asked to judge the size of a speaker (length of the vocal tract) uttering the vowel/ae/ in the presence of a variable-size speaker uttering the vowel/a/. The frequencies of the spectral prominences of the vowels in this case varied in much the same way as they did for the variable-size membrane and plate of the first study (cf. Patterson et al. 2008). The results shown in Fig. 23.3 supported target enhancement, as expected. In a second task listeners were asked to judge the force on the vocal folds for/ae/ ('loudness') in the presence of/a/spoken with variable force. The levels of the spectral prominences in this case varied in much the same way as they did for the membrane and plate struck with variable force. The results, given in Fig. 23.4, show listeners in this case to be split between target enhancement and noise cancellation.

The results fail to provide consistent support for the fixed-noise hypothesis, but in so doing, they underscore the difficulty in inferring listener decision strategy from performance accuracy alone. Note, in particular, that in Fig. 23.4, where the performance of two listeners is nearly identical (cyan and red squares), the decision weights reflect clearly different listening strategies ($\hat{w}_N$ for the two listeners being different in sign). Conversely, where the performance of two other listeners is quite different (yellow and blue diamonds), the decision weights indicate similar decision strategies ($\hat{w}_N$ for the two listeners being both positive in sign).

4 Summary Comments

We conclude by mentioning two other unexpected outcomes in application of this approach involving the identification of single sound sources in isolation. The first is *level dominance*. This is the finding that the level of a partial, even in the absence of any significant masking, is often more predictive of the listener's decision weight on that partial than the amount of information it conveys for identification (Lutfi et al. 2008). The result is surprising because it is exactly opposite to that expected from a maximum-likelihood observer. Level dominance has not, to the authors' knowledge, been studied in the hard of hearing, but it might be expected to have an influence on the benefit of amplification provided by hearing aids. The second unexpected finding is that the acoustic information for basic source attributes, such as material and size, is highly vulnerable to information loss at the cochlea (Lutfi 2001; Lutfi and Stoelinga 2010). This too would seem to have implications for understanding the impact of hearing loss on everyday listening when hearing sensitivity is reduced. Taken together, it is hoped that the results of these and future studies applying this approach can be used to advance development of a computational model of normal listening

in multisource acoustic environments, one that can ultimately be used to generate new and testable hypotheses regarding the problem of listening in such environments for the hard of hearing.

Acknowledgments

Figure 23.1 was reproduced from Lutfi and Liu (2011b) with permission from the Journal of the Acoustical Society of America. This research was supported by NIDCD grants R01DC006875 and R01DC01262.

References

- ANSI. ANSI S3.6-2004. Specification for audiometers. American National Standards Institute; New York: 2004.
- Berg BG. Analysis of weights in multiple observation tasks. *J Acoust Soc Am.* 1989; 86:1743–1746.
- Bregman, AS. Auditory scene analysis: the perceptual organization of sound. Bradford Books/MIT Press; Cambridge, MA: 1990.
- Carlyon RP, Moore BCJ. Continuous versus gated pedestals and the severe departure from Weber's law. *J Acoust Soc Am.* 1986; 79:453–460. [PubMed: 3950198]
- de Cheveigné A, McAdams S, Laroche J, Rosenberg M. Identification of concurrent harmonic and inharmonic vowels: a test of the theory of harmonic cancellation and enhancement. *J Acoust Soc Am.* 1995; 97:3736–3748. [PubMed: 7790652]
- Durlach N. Equalization and cancellation theory of binaural masking-level differences. *J Acoust Soc Am.* 1963; 35:1206–1218.
- Klatt DH. Software for a cascade/parallel formant synthesizer. *J Acoust Soc Am.* 1980; 67:971–995.
- Lutfi RA. Auditory detection of hollowness. *J Acoust Soc Am.* 2001; 110:1010–1019. [PubMed: 11519569]
- Lutfi RA, Liu CJ. Individual differences in source identification from synthesized impact sounds. *J Acoust Soc Am.* 2007; 122:1017–1028. [PubMed: 17672650]
- Lutfi RA, Liu CJ. A method for evaluating the relation between sound source segregation and masking. *J Acoust Soc Am.* 2011a; 129:EL34–38. [PubMed: 21302979]
- Lutfi RA, Liu CJ. Target enhancement and noise cancellation in the identification of a rudimentary sound source in noise. *J Acoust Soc Am.* 2011b; 129:EL52–56. [PubMed: 21361412]
- Lutfi RA, Stoelinga CNJ. Sensory constraints on the identification of the geometric and material properties of struck bars. *J Acoust Soc Am.* 2010; 127:350–360. [PubMed: 20058982]
- Lutfi RA, Liu CJ, Stoelinga CNJ. Level dominance in sound source identification. *J Acoust Soc Am.* 2008; 124:3784–3792. [PubMed: 19206804]
- Meddis R, Hewitt MJ. Modeling the identification of concurrent vowels with different fundamental frequencies. *J Acoust Soc Am.* 1992; 91:233–245. [PubMed: 1737874]
- Morse, PM.; Ingard, KU. Theoretical acoustics. Princeton University Press; Princeton: 1968. p. 175-121.
- Patterson RD. Auditory filter shapes derived with noise stimuli. *J Acoust Soc Am.* 1976; 59:640–654. [PubMed: 1254791]
- Patterson, RD.; Smith, DRR.; van Dinter, R.; Walters, TC. Size information in the production and perception of communication sounds. In: Yost, WA.; Popper, AN., editors. Springer handbook of auditory research: auditory perception of sound sources. Springer; New York: 2008. p. 43-76.
- Piechowiak T, Ewert SD, Dau T. Modeling comodulation masking release using an equalization-cancellation mechanism. *J Acoust Soc Am.* 2007; 121:2111–2126. [PubMed: 17471726]
- Werner LA, Bargones JY. Sources of auditory masking in infants: distraction effects. *Percept Psychophys.* 1991; 50:405–412. [PubMed: 1788029]
- Wier CC, Jesteadt W, Green DM. Frequency discrimination as a function of frequency and sensation level. *J Acoust Soc Am.* 1977; 61:178–184. [PubMed: 833369]

Fig. 23.1. Noise decision weights $\hat{w}_N$ versus d' for five listeners (symbols with 95 % confidence intervals) judging membrane size in the presence of a variable-size plate. Curves give predictions for three models of interference. Repeated symbols represent different values of σ_T

Fig. 23.2.

Same as Fig. 23.1 except task is to judge force of membrane impact in the presence of a plate hit with variable force

Fig. 23.3.

Same as Fig. 23.1 except task is to judge size of speaker uttering the vowel/ae/ in the presence of a variable-size speaker uttering the vowel/a/

Fig. 23.4. Same as Fig. 23.3 except task is to judge spoken force ('loudness') of vowel/ae/in the presence of a vowel/a/spoken with variable force