Dimensions of auditory segregation: What do they tell us about levels of auditory processing?

P. L. Divenyi

Veterans Affairs Medical Center, Experimental Audiology Research Laboratory
150 Muir Road, Martinez, California, 94553, United States
pdivenyi@marva4.ebire.org

1. Introduction

Perceptual segregation of multiple, simultaneous auditory streams has been becoming an area of interest to several investigators since the publication of Bregman's (1991) landmark book on the subject. The principal motivation behind these investigations has been to find a closure on the classic, and invariably elusive, problem of the auditory segregation of simultaneous speech signals, i.e., the "cocktail-party effect." Search for psychophysical underpinnings of this unique human ability, however, has been by-and-large limited to research on the segregation of concurrent harmonic complexes, especially vowels (Marin and McAdams 1991; Carlyon 1992; de Cheveigné 1993; Assmann 1996). The present work attempts to approach the problem from a more comprehensive point of view and take a look at how the various auditory dimensions contribute to, and interact in, the segregation of concurrent sounds.

Earlier (Divenyi 1995), we proposed a framework for auditory segregation of sounds concurrently emanating from several sources, that incorporated three main, "cardinal" dimensions: the acoustic characteristics of the signal emanating from each source (including short-term envelope fluctuations down to modulation frequencies of about 20 Hz), the temporal pattern generated by syllabic- and subsyllabic-rate envelope fluctuations within each source, and the spatial location of the sources. For a necessary and sufficient criterion of segregation, we proposed to adopt the dual requirement of (1) having to resolve the difference between the concurrent sources along the diverse dimensions, and (2) correctly associating the particular values perceived along the various dimensions. Thus, if Talker **A** says *X* and Talker **B** simultaneously says *Y*, even if **A**, **B**, *X*, and *Y* are correctly identified, the report that "Talker **A** said *Y* and Talker **B** said *X*" indicates an incorrect association and will not be accepted as signaling segregation. In addition, we postulated that, at some segregation threshold, differences between concurrent signals should trade off against one another.

2. Theory of auditory segregation

For the sake of simplicity, let us consider the case of only two concurrent signals to be segregated. Let us define normalized differences along the three cardinal dimensions to be Δf (for the difference in the combined attribute of spectral profile and pitch), Δt (for the difference in envelope structure), and $\Delta \varphi$ (for the difference in spatial location). According to the tradeoff rule,

$$\Delta f \, \Delta t \, \Delta \varphi = k \qquad (1)$$

but only if the three dimensions are mutually orthogonal, i.e., the process of segregation along the three dimension is independent of one another. In case the dimensions are pairwise correlated, Equation (1) should be replaced by the more general

$$\Delta f\, \Delta t\, \Delta \varphi\ [(1-\rho_{f,t})\,(1-\rho_{f,\varphi})\,(1-\rho_{t,\varphi})]^{-1} = k \qquad (2)$$

where $\rho_{f,t}$, $\rho_{f,\varphi}$, and $\rho_{t,\varphi}$ are the correlations between the respective pairs of dimensions, as they affect segregation. Correlation is to be interpreted as the perceived difference between the two signals along a given dimension being interfered with (helped or impeded) by a perceived difference along another dimension. Along any given dimension, a *resolvable* difference between the two signals is needed, in order for one of the signals not to *mask* the other. Clearly, here we are referring to *informational*, rather than *energetic* masking (Watson 1987). Along a given dimension, and within certain limits, the degree of informational masking is monotonically related to the information-carrying difference along the dimension (Kidd, Mason et al. 1998). This monotonicity was also experimentally demonstrated, as shown in Fig. 1 for the discrimination of pitch difference in three-component harmonic complexes.

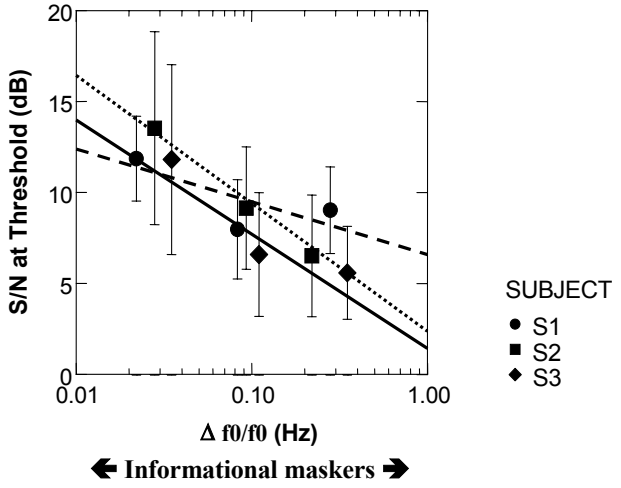

Figure 1. *Informational masking of the discrimination of two sequentially presented three-component harmonic complexes differing in pitch (with f_0 in the 100-Hz region) presented in a masker consisting of multiple series of harmonic components the normalized pitch differences between which covered about four octaves. Data of three experienced listeners; the lines represent linear regression of S/N as a function of the logarithm of $\Delta f_0/f_0$.*

Similar informational signal-to-noise thresholds were obtained for the discrimination of a harmonic complex with components in the 500-1200-Hz range presented successively at different azimuths, mixed with a masker consisting of the same complex emanating from 18 different azimuthal locations (obtained through headphone presentation using the subject's own head-related transfer functions). Informational masking thresholds were also determined for the discrimination of an accelerating and a decelerating pattern of three bursts of a three-component harmonic carrier amplitude-modulated at an average rate of 4.375 Hz and at different modulation indexes, presented in an informational noise generated by four simultaneous versions of the same carrier independently and randomly amplitude-modulated at the same average rate. These two informational masking threshold measurements also indicated that the logarithm of spatial distance and masking, as well as modulation depth (in dB) and masking had a monotonic relationship with a significant linear component. Thus, resolution of differences along any cardinal dimension $\boldsymbol{X}$ can be represented in terms of a simple informational masking psychometric function having the form

$$S/N_X \approx \boldsymbol{a}_X + \boldsymbol{b}_X \log(\Delta x/x) \qquad (3)\ .$$

Since a_X is an additive constant not contributing to the shape of the function, it can be ignored and the normalized difference of x can be expressed as ΔX^{bX}, where ΔX is the linearized S/N_X.

Now, we can use the above relationship and the tradeoff relationship of Eq. (1) to express the predicted tradeoff in the theoretical situation in which two concurrent signals differing along two orthogonal dimensions have to be segregated. The independence of the two dimensions is a fact, since in this situation the two dimensions never physically coexist—their coexistence is merely imagined. Thus, substituting Eq. (3) into Eq. (1) will yield for dimensions X and Y

$$(\Delta x/x)\,(\Delta y/y) \approx \Delta X^{bX}\,\Delta Y^{bY} = k \qquad (4)$$

or

$$b_X \log(\Delta x/x) \approx \log k - b_Y \log(\Delta y/y) \qquad (5)\ .$$

In other words, the theoretical tradeoff function between normalized differences along two orthogonal dimensions is a negative-sloped line in log-log coordinates and can be predicted from the slopes of their respective informational masking psychometric functions. One example of this predictive operation is shown in Figure 2 in which a tradeoff function for segregation based on pitch difference vs. azimuth separation was constructed from informational masking psychometric functions obtained separately for pitch and azimuth.

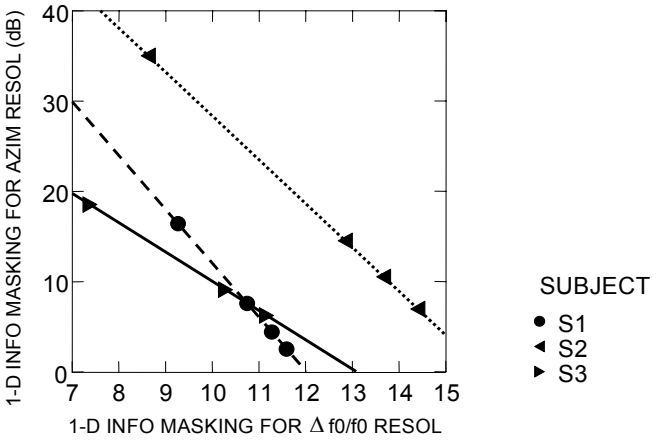

Figure 2: *Two-dimensional tradeoff between azimuth-based and pitch-based segregation, as predicted from informational masking psychometric functions obtained separately for pitch resolution and azimuth resolution. Data for three experienced subjects.*

3. Segregation experiments

We conducted a series of experiments in which experienced young listeners had to segregate two concurrent signals differing, in any condition, along two of our three dimensions.

3.1. Methods

Stimuli were two simultaneous streams of three-component harmonic complexes. The three dimensions were pitch in the 100-Hz f_0 region, temporal structure of low-modulation frequency envelope patterns (accelerating or decelerating), and azimuthal location. The difference between streams along one of the dimensions was fixed within any given condition, whereas the one along the other dimension was adaptively varied, in order to estimate a threshold difference at which segregation was barely possible. The stimuli were presented in a two-alternative forced choice paradigm as follows (see Figure 3): For one of the alternatives, in Interval 1, Stream A contained a signal with value x_1 paired with y_1 on dimensions X and Y, respectively, and in Interval 2, it contained x_1 paired with y_2; simultaneously,

in Interval 1, Stream B contained x_2 paired with y_2 and Interval 2 contained x_2 paired with y_1. For the other alternative, the order of the two intervals was reversed. The subject was instructed to attend to one of the streams, say the one with a higher pitch, and indicate the nature of the change in that stream with respect to the other dimension, say right to left. Thus, segregation (as indicated by correct association of the relative values along each of the two dimensions) was measured with an objective procedure. Feedback was given after every response.

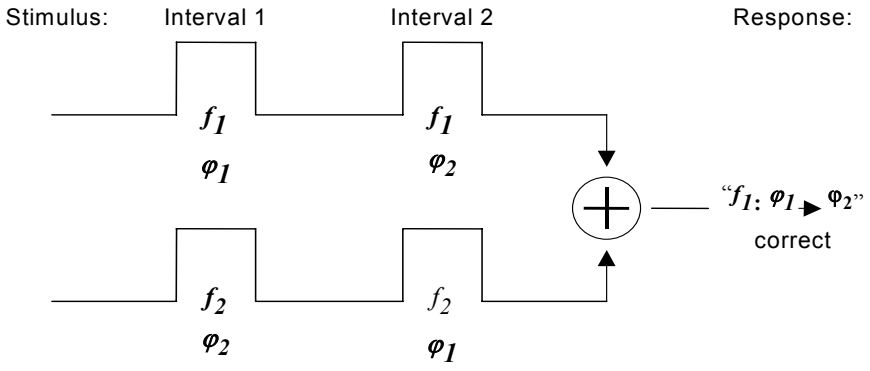

Figure3*: Schematic diagram of a trial. The two dimensions are pitch (f_1, f_2) generated by the same three harmonics of two different fundamental frequencies, and azimuth (φ_1, φ_2) generated by presenting the signals at two different azimuths using the subject's own HRTFs. The instruction to the subject was to listen to the pitch f_1 (usually the higher one) and tell whether the signal having that pitch went from left to right or right to left. The order of the two intervals was random.*

3.2. Results

Thresholds for the segregation of two streams were obtained for the tradeoff of differences along the two dimensions regarding to which the two streams differed from one another: Pitch difference against azimuth difference, pitch difference against temporal structure difference, and azimuth difference against temporal structure difference. For illustrative purposes, data for the tradeoff between pitch difference and azimuthal position difference is displayed in Figure 4a (degree vs. normalized pitch difference) and in Figure 4b (the differences on the abscissa and ordinate of Fig. 4a translated into dB differences using the informational masking conversion from psychometric functions similar to those shown in Fig. 1). The results prompt several observations. First, the three subjects, no matter how experienced, exhibit considerable individual differences that manifest in the divergent slopes. The reason for these differences must lie in that segregation with respect to only two of the many possible dimensions is likely to be accomplished by a given listener on the basis of his/her natural or learned level of capability of grasping and decoding information on a certain dimension. There are subjects who can hear minute pitch differences—those individuals will predominantly base their segregation decisions on pitch. Others can do remarkably well localizing sounds—for those listeners segregation will be predominantly based on azimuthal cues.

The second observation is that, from these data alone, one cannot make any inference as to the correlation, or the lack of it, between the two dimensions present. For this to be possible, we will have to evoke our tradeoff predictions based on estimates of informational masking obtained separately for each dimension. These

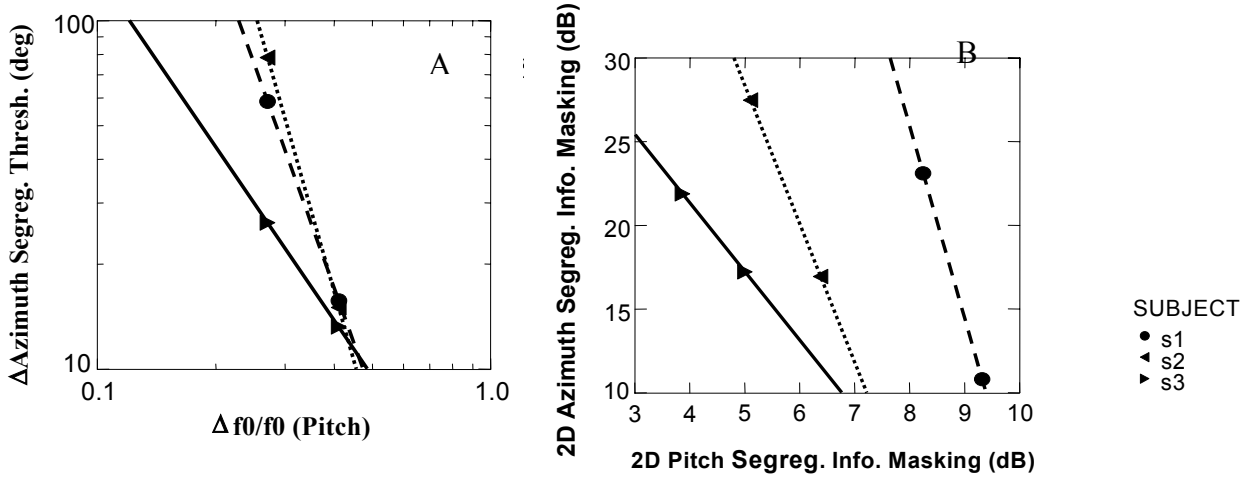

Figure 4: *Tradeoff results of a segregation experiment with the two streams differing along two dimensions, namely, pitch and azimuthal location. Data of three listeners. Panel A: abscissa and ordinate on logarithmic scale; they depict the actual difference in pitch (normalized difference) and azimuth (in degrees). Panel B: the same data represented after dB conversion using the informational masking S/N ratio associated with a given difference in pitch or azimuthal location.*

predictions yield estimates for the tradeoff slopes for each subject and for each condition, which can be compared with those obtained in segregation experiments that involve the same two dimensions. Comparisons of predicted and obtained slopes, again for the pitch-azimuth dimension pair, are illustrated in Figure 5. Clearly, the predicted and obtained slopes are nearly identical only for one subject, whereas for the other two the predicted slopes are considerably shallower than those obtained. With a simple geometric exercise, however, we can make, for each subject, the predicted slope identical to the obtained slope: we can do this by changing the coordinate system from rectangular to polar, i.e., by closing the angle between the abscissa and the ordinate until the two slopes will be the same. At that point, the correlation between the two axes will equal to the cosine of the angle enclosed between them. Thus, this simple operation results in an estimate of the correlation between the two dimensions of the three subjects' results in our two-dimensional segregation experiments.

Table I. Estimated correlations between dimensions in 2D relation

2D Relation	Subject	Correlation ρ
$\Delta\varphi = f(\Delta t)$ (azimuth/temporal structure)	S1	0.217
	S2	0.017
	S3	0.251
$\Delta\varphi = f(\Delta f)$ (azimuth/pitch)	S1	0.003
	S2	0.220
	S3	0.152
$\Delta f = f(\Delta t)$ (pitch/temporal structure)	S1	0.340
	S2	0.307
	S3	0.053

A complete list of the correlations obtained for segregation based on all three dimension pairs is shown in Table I. It appears that the three subjects exhibit rather

marked individual differences that are qualitative as well as quantitative. The qualitative difference between the three listeners dwells in the fact that for each one a different pair of dimensions is nearly orthogonal, i.e., for each subject a different dimension is correlated with the other two, suggesting that that dimension is the one from which the listener gathers most segregation cues. Thus, S1 is a temporal structure specialist, S2 is a pitch listener, and S3 gets most information from location.

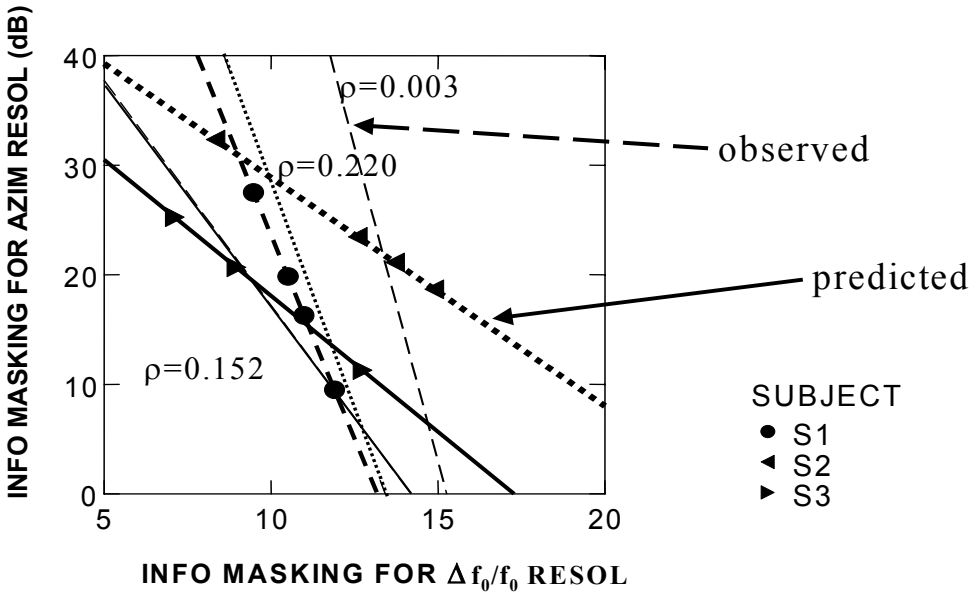

Figure 5: *Comparison of the tradeoff slopes obtained from prediction (heavy lines) as described in Figure 3, and obtained from observed segregation results, as shown in Figure 4. The two dimensions traded are pitch resolution and azimuth resolution. The scales on both axes represent a conversion of a given difference along the pitch or azimuth dimension into a corresponding informational masking estimate. Note that the predicted slopes for Subjects S2 and S3 are shallower than the observed one. To make the predicted slope equivalent to the observed for these two subjects, the rectangular coordinate system must be warped into a polar coordinate system with an angle between the y and the x axes smaller than 90°. The cosine of that angle will equal to the correlation between the dimensions.*

4. Discussion

In the foregoing paragraphs, a theory was presented to propose that auditory segregation of concurrent signals that differ along more than one dimension are the result of segregation processes aimed at each of the dimensions, which may not be independent. It was further proposed that, with minimal assumptions that directly address the relationship between concurrent resolution and informational masking along a given auditory dimension, the correlation between the information gained from cues along the different dimension to achieve segregation may be estimated.

But where do these processes take place? Have we been measuring primitive or schema-driven segregation, as defined by Bregman (1991)? This question, unfortunately, cannot be directly answered from our results. Nevertheless, one can hypothesize without taking any big leap that cues based on pitch or spectrum, as well as those based on location should be readily available to the listener without having to reach for stored information or without heavily taxing his attention. In contrast, one of our dimensions was that of temporal structure consisting of slow, syllabic-rate envelope fluctuations: cues based on that dimension are unlikely to derive from

peripheral processing because the time constants needed to register differences among these structures are within the realm of cortical activity (Schreiner and Urbas 1986). Processing temporal patterns of this order for the segregation of streams differing in pitch or location also means that the pitch and/or location information has to be retained and efficiently retrieved even for the longer duration deciphering the slow pattern-bound cues takes. Hence, a correlation between low-rate temporal structure-based and more peripherally-based segregation should indicate that, at least for our nonlinguistic stimuli and for certain listeners, primitive and schema-driven segregation are quasi inseparable.

As to the individual differences, they are rather the norm than the exception when it comes to experiments requiring complex strategies. What our listeners clearly show is that human beings will always obtain information in a way that maximizes their success. Whenever obtaining such information from cues obtained from one physical dimension is insufficient, they will switch strategy and get the information from another dimension. But the tradeoff relation between the resolution along the different dimension also shows that switching from one dimension to another cannot occur without loss, as stated by information theory (Gábor 1946) and transposed to the present problem by Equations (1) and (2). Thus, ultimately, the laws of physics will impose limitations on the "cocktail-party effect" much as they do on the whole world around us.

5. Acknowledgements

This research has been supported by Grant R01-07998 from the National Institute on Aging and by the Veterans Affairs Medical Research.

6. References

Assmann, P. F. (1996). Modeling the perception of concurrent vowels: Role of formant transitions. J. Acoust. Soc. Am. 100, 1141-1152.

Bregman, A. S. (1991). *Auditory scene analysis*. Cambridge, Mass., Bradford Books (MIT Press).

Carlyon, R. P. (1992). The psychophysics of concurrent sound segregation. *Philos. Trans. .Roy. Soc. .London. B* 336, 347-355.

de Cheveigné, A. (1993). Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. J. Acoust. Soc. Am. 93, 3271-3290.

Divenyi, P. L. (1995). Auditory segregation of concurrent signals: An operational definition. *Proceedings of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics.* New Paltz, New York, IEEE. Section 2.3: 1-4.

Gabor, D. (1946). Theory of communication. J. Inst. Electr. Engin. (London) 93: 429-457.

Kidd, G., Jr., Mason, C. R. and Rothla, T. I. (1998). Identification of brief auditory patterns. J. Acoust. Soc. Am. 103, 3019-3020.

Marin, C. M. H. and McAdams S. (1991). Segregation of concurrent sounds II: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width. J. Acoust. Soc. Am. 89, 341-351.

Schreiner, C. E. and Urbas R. V. (1986). Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). Hear. Res. **21**: 227-241.

Watson, C. S. (1987). Uncertainty, informational masking, and the capacity of immediate auditory memory. *Auditory processing of complex sounds.* W. A. Yost and C. S. Watson (Eds). Hillsdale, New Jersey, L. Erlbaum**,** pp. 267-287.