ISCA Archive
http://www.isca-speech.org/archive

Auditory-Visual Speech Processing
(AVSP'98)
Terrigal - Sydney, Australia
December 4-6, 1998

INTEGRATION OF AUDIOVISUALLY COMPATIBLE AND INCOMPATIBLE CONSONANTS IN IDENTIFICATION EXPERIMENTS.

Louis D. Braida, Kaoru Sekiyama, and Ann K. Dix

Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139

1. INTRODUCTION

Although studies of speech perception often focus on the interpretation of the acoustic speech waveform, in many situations the face of the talker can be seen by the listener and the perception of facial actions can greatly influence the interpretation of this signal. When the acoustic signal is unclear, observed facial actions can improve intelligibility [14]. On the other hand, even when the acoustic signal is highly intelligible, observing mismatched facial actions can reduce the effectiveness of acoustic speech cues. In the "McGurk Effect" [8, 7], a classic example of the latter influence, an artificial stimulus – constructed from the acoustic component of one consonant and the facial actions of another consonant that differs in place of articulation – may be perceived as a third consonant or as a consonant cluster.

The purpose of this paper is to demonstrate that these two influences reflect common aspects of a single audiovisual integration process that can be described quantitatively. To simplify the demonstration we focus on the case of stimuli consisting of consonant segments presented in identification experiments in which subjects are required to identify stimuli from among a known closed set of items.

2. INTEGRATION MODEL

Consonants are assumed to be identified on the basis of the sample value of a vector of cues $\vec{X} = \langle a_1, a_2, \ldots, a_M, v_1, v_2, \ldots, v_N \rangle$ that has both auditory $(a_1, a_2, \ldots, a_M)$ and visual $(v_1, v_2, \ldots, v_N)$ components. When consonant C_j is presented, the components of $\vec{X}$ are independent identically distributed Gaussian random variables with means $(A_{j1}, A_{j2}, \ldots, A_{jM}, V_{j1}, V_{j2}, \ldots, V_{jN})$ and a common standard deviation $\sigma = 1.0$. Each consonant is thus associated with a *stimulus center* specified by the mean value of the cue vector for that consonant. The subject is assumed to assign a response by determining which *response center* or *prototype* $\vec{R}_i = \left\langle r_{i1}, r_{i2}, \ldots, r_{i(M+N)} \right\rangle$ is closest to the cue vector on a given stimulus presentation.

In previous work [2] we considered the problem of predicting how well compatible audiovisual stimuli can be identified based only on knowledge of the confusion matrices obtained in separate audio and video identification experiments. We assumed that the subject bases decisions only on the auditory (or visual) components of the cue vector in unimodal experiments. Stimulus centers in these experiments can be estimated from the observed confusion matrices using a type of metric multidimensional scaling. Identification scores in audiovisual experiments can be predicted reasonably well by assuming that subjects use a response center for each consonant that coincides with the stimulus center for that consonant, i.e. $\vec{R}_i = \vec{S}_i$. When the stimuli are presented with equal frequency, the coincident center case leads to the highest possible identification score for a given set of stimulus centers. Small deviations of the response centers from these optimal locations generally have only minor effects on the identification score.

3. EXPERIMENTS

The identification experiments were part of a larger study concerned with variations in audiovisual integration seen across individuals with different national characteristics. Listeners were asked to specify the stimulus from the set /b, d, g, p, t, k, n, m, n/. The stimuli were derived from C-/α/ syllables spoken in isolation by two native speakers of Japanese (1m and 1f) and two native speakers of American English (1m and 1f). Stimuli were constructed from two utterances of each consonant produced by each speaker. In addition to auditory ($\mathcal{A}$), visual ($\mathcal{V}$) and compatible audiovisual ($\mathcal{AV}$) stimuli, incompatible audiovisual ($\mathcal{AV}^*$) stimuli were constructed by dubbing acoustic waveforms onto video recordings of other consonants produced by the same speaker. The resulting $\mathcal{AV}^*$ discrepant pairings (/bg,db,gb,pk,tp,kp,mn,nm/) were incompatible only with respect to place of articulation. In all cases, the acoustic speech waveform was degraded by lowpass filtering and additive noise.

Twenty eight young Japanese adults were tested.

Stim.	Response		
AV	b	d	g
b–	56.3	28.3	15.4
d–	30.0	45.7	24.3
g–	25.4	31.1	43.4
–b	97.8	1.7	0.4
–d	1.0	85.4	12.9
–g	1.1	42.5	56.4
bb	99.3	0.4	0.2
dd	0.3	87.4	12.4
gg	0.9	26.9	72.2
bg	1.5	59.8	38.7
db	96.3	3.1	0.7
gb	96.7	1.1	2.2

Table 1. Reduced confusion matrices for the hypothetical /b,d,g/ stimuli. Response proportions are reported as percentages.

The $\mathcal{A}$ and $\mathcal{V}$ presentation modes were tested separately, but $\mathcal{AV}$ and $\mathcal{AV}^*$ stimuli were mixed in the audiovisual condition. Each subject was presented each stimulus 16 times (four for each speaker) in each mode. In the analysis that follows, results from the different speakers and subjects were combined to form four confusion matrices, one for each of the $\mathcal{A}$, $\mathcal{V}$, $\mathcal{AV}$, and $\mathcal{AV}^*$ conditions, shown in Tables 2–5 below.

4. ANALYSIS

To illustrate the application of the model, the four confusion matrices were simplified by combining stimuli and responses across voicing and manner categories. This produced matrices similar to those that might have been obtained in hypothetical identification experiments in which the stimuli and responses were restricted to /b/, /d/, and /g/. The resulting simplified matrices are shown in Table 1.

This simplification allows the unimodal matrices to be fit by unidimensional (scalar) versions of the model (i.e., $M = N = 1$) and audiovisual matrices to be predicted by two-dimensional versions ($D = M + N = 2$) of the model. The cue spaces for the unimodal conditions are orthogonal straight lines (Fig. 1). The cue spaces for the audiovisual conditions are planes that contain these lines.

The stimulus centers for the unimodal consonants are marked by diamonds, squares, and circles, with filled symbols on the horizontal axis marking the centers for the $\mathcal{A}$ stimuli and open symbols on the vertical axis marking the centers for the $\mathcal{V}$ stimuli. According to the model described above, the stimulus centers for the compatible audiovisual stimuli are predicted to be B, D, and G. Thus, for example, the stimulus center (marked by B) for the AV compatible /bb/ has a horizontal coordinate that is the same as that for the auditory /b/ and a vertical coordinate that is the same as that for the visual /b/. The stimulus centers for the compatible $\mathcal{AV}$ stimuli are generally close to the corresponding response centers [2], which, for clarity, are not shown in Fig. 1.

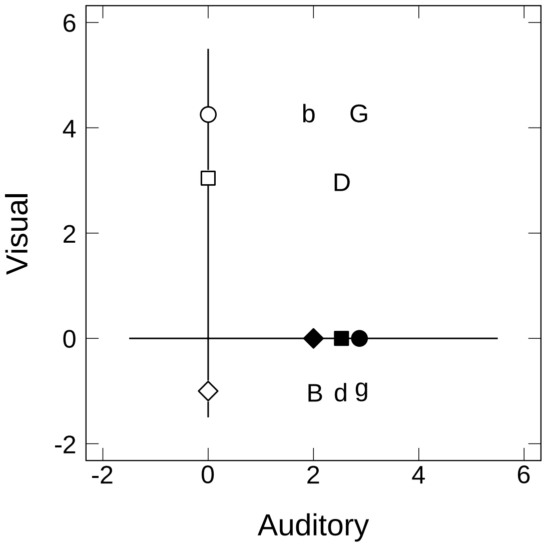

Figure 1. Stimulus centers corresponding to the reduced confusion matrices for the hypothetical three-stimulus identification experiments. In the unimodal experiments, filled diamonds mark the stimulus centers for auditory /b/, filled squares for auditory /d/, and filled circles for auditory /g/. Open diamonds mark the stimulus centers for visual /b/, open squares for visual /d/, and open circles for visual /g/. Upper case letters mark the stimulus centers for the compatible audiovisual stimuli and lower case letters mark those of the incompatible audiovisual stimuli.

Note that the separations between the stimulus centers B, D, and G are greater than those between the corresponding stimulus centers on either the horizontal or vertical axes. This increase in separation corresponds to an improvement in correct identification scores from 49% ($\mathcal{A}$) and 83% ($\mathcal{V}$) to 88% ($\mathcal{AV}$).

The positions of the centers for the incompatible $\mathcal{AV}^*$ stimuli are marked by b for /bg/, d for /db/, and g for /gb/. Thus, for example b, the stimulus center for the incompatible bimodal stimulus composed of an auditory /b/ and a visual /g/, has a horizontal coordinate that is the same as the auditory /b/ and a vertical coordinate that is the same as that for the visual /g/.

Assuming that the audiovisual response centers are located near the stimulus centers for the compatible

$\mathcal{AV}$ stimuli, then the reponses elicited by the incompatible /bg/ stimulus are expected to be predominantly “d”s and “g”s because b is closer to D and G than to B. Similarly, the incompatible /db/ and /gb/ stimuli are expected to elicit primarily “b” responses. These expectations are borne out in the confusion matrices for the incompatible bimodal stimuli shown in Table 1. Whereas 51% of the responses to the $\mathcal{A}$ stimuli reflect errors in reception of place of articulation, nearly 98% of the responses to the $\mathcal{AV}^*$ stimuli are place errors.

5. PREDICTIONS

Formal predictions for the 8-stimulus audiovisual confusion matrices were derived from the unimodal $\mathcal{A}$ (Table 2) and $\mathcal{V}$ (Table 3) confusion matrices. Three dimensional configurations ($N = M = 3$) were derived independently to account for the patterns of errors seen in these matrices.

Stim.	Response							
AV	b	d	g	k	m	n	p	t
b–	56 56	29 29	11 11				2 2	1 1
d–	14 13	42 42	44 42					
g–	21 21	30 30	47 48			2 1		
k–				40 40			29 29	30 30
m–					71 72	26 25		
n–			1 1		50 50	48 48		
p–				34 33			37 38	28 29
t–				27 28			25 25	47 46

Table 2. Confusion matrix for the $\mathcal{A}$ stimuli. Response proportions are reported as percentages, with proportions less than 1% suppressed for clarity. For each stimulus, observed proportions are reported in the first row, percentages corresponding to the fit of the three-dimensional model to the matrix are reported in in the second.

The configuration of stimulus centers for the $\mathcal{A}$ stimuli in Fig. 2 shows a tight clustering of stimulus centers for consonants with different places of articulation but the same voicing/manner of articulation. For the $\mathcal{V}$ stimuli, the centers are also clustered, but the clusters contain centers for consonants with different voicing/manner of articulation. Consonants with different places of articulation have centers in different clusters. For both unimodal conditions, derived identification accuracy is relatively close to that observed: 48.6 vs 48.4% for the $\mathcal{A}$ stimuli and 32.8 vs 32.4% for the $\mathcal{V}$ stimuli. As can be seen in Tables 2 and 3, however, the fit to the confusion patterns evident in the $\mathcal{A}$ matrix (errors on the order of one percentage point or less) is generally better than the fit to the $\mathcal{V}$ matrix (some errors of five and six points).

Stim.	Response							
AV	b	d	g	k	m	n	p	t
–b	30 30	0 1			26 26	1 0	41 41	
–d	1 1	30 35	9 7	6 7	1 0	14 16		38 33
–g	0 1	13 10	24 25	28 28		12 11		21 25
–k		10 8	30 28	31 32		11 10		18 20
–m	25 25				29 30		45 44	
–n			7 8	4 5		28 26		34 34
–p	30 27				20 23		49 48	
–t	2 1	35 36	4 7	7 7		12 16		39 33

Table 3. Confusion matrix for the $\mathcal{V}$ stimuli.

The stimulus centers used to fit the $\mathcal{A}$ and $\mathcal{V}$ matrices were then used to construct the stimulus centers for the $\mathcal{AV}$ and $\mathcal{AV}^*$ stimuli. One set of eight response centers was assumed to apply to both the compatible and incompatible audiovisual stimuli. These centers had the same auditory coordinates as the response centers for the $\mathcal{A}$ condition and the same visual coordinates as the response centers for the the $\mathcal{V}$ condition. Thus no additional parameters were used to make predictions for the audiovisual confusion matrices.

Monte Carlo techniques were used to derive predictions for the patterns of responses to the $\mathcal{AV}$ stimuli (Table 4) and for the responses to the $\mathcal{AV}^*$ stimuli (Table 5). For the compatible stimuli, the integration model predicts the improved identification accuracy fairly well (83.3% vs 87.1%). Larger improvements would have been predicted if the response centers had been closer to the stimulus centers. For the incompatible stimuli, the model predicts strong tendencies for the acoustic consonants to be misidentified. The few sizable discrepancies in the distribution of errors (e.g., the responses to the /bg/ stimulus) can be reduced by small shifts in the locations of the response centers.

6. DISCUSSION

The approach to studying audiovisual integration described in this paper has certain limitations that need

Stim.	Response							
AV	b	d	g	k	m	n	p	t
bb	98 93	0 2					2 4	
dd	0 0	75 74	24 24					
gg	1 1	24 31	73 66			1 1		
kk			1 0	70 64				28 34
mm					99 96		0 2	
nn		1 1			0 3	98 93		2 2
pp	1 1						98 97	0 1
tt				13 14			0 1	86 83

Table 4. Confusion matrix for the $\mathcal{AV}$ stimuli. For each stimulus, observed proportions are reported in the first row, predicted percentages are reported in the second.

Stim.	Response							
AV	b	d	g	k	m	n	p	t
bg	1 3	52 38	45 55	0 1				1 3
db	93 88	4 7	2 3				1 1	
gb	92 89	2 5	4 2		1 2		0 1	
kp							98 96	0 2
mn				1 1	2 4	94 89		2 5
nm	1 1				93 98	5 1	1 0	
pk				68 63				31 35
tp					0 1		99 95	1 1

Table 5. Confusion matrix for the $\mathcal{AV}^*$ stimuli.

to be addressed in future research.

First, the modelling approach is directed at predicting subject behavior rather than at describing the mechanisms responsible for that behavior. The finding that compatible audiovisual stimuli are identified roughly as well as would be expected on the basis of optimal processing of the available audio and visual cues neither specifies how those cues are derived from the stimuli nor how the integration is accomplished. It does, on the other hand, place constraints on the integration process. For example, the identification of audiovisual stimuli might make use of intermodal cues, such as time intervals between acoustic and visual events, which would not be available in the unimodal conditions. However such cues can contribute to identification accuracy only to only a minor degree.

Second, the modelling approach requires that the stimuli used in unimodal experiments be confused with one another to a measurable degree in order to determine the structure of the unimodal cue spaces. As a result, it was necessary to study the McGurk effect using degraded acoustic stimuli, although the effect has been observed even when no degradations are applied to the acoustic speech waveforms. In future work it may be possible to overcome this limitation by including stimuli that have been subject to a range of levels of degradation in the identification experiments.

Third, the use of identification experiments, in which the the response set is restricted by the experimenter, to study integration processes means that it is not possible to predict the occurrence of unanticipated responses. Although such responses are extremely rare when unimodal or compatible bimodal stimuli are presented, they can constitute a substantial fraction of the responses when labial visual consonants are paired with nonlabial auditory consonants. Although the occurrence of such responses can be accomodated within the framework presented in Sec. 2 by introducing additional response centers in the multimodal space, this modification would have no predictive value. Clearly the conditions that give rise to such responses deserve further study.

Fourth, the technology for estimating model parameters is still fairly primitive. In particular, maximum likelihood estimation techniques are not currently available for deriving the parameters. As a result, more responses must be obtained in each experimental condition to attain a specified level of accuracy. Simulation studies have shown that each stimulus must be presented roughly 40 times just to reduce the bias in parameter estimates to levels comparable to the variability of those estimates associated with the underlying Bernoulli statistics, so that an even larger number of presentations would be desirable. It is unclear whether subjects maintain stable response patterns in the extended experiments required to measure the necessary confusion patterns. The results reported in this paper were derived by combining responses from 28 subjects. The resulting data sets were adequate to estimate model parameters but are not necessarily applicable to any individual subject. Moreover, even if all subjects based responses on the same stimulus centers, model parameters would underestimate their ability to distinguish between stimuli if different subjects employ different response centers.

7. CONCLUSIONS

The ability of the integration model to predict the pattern of responses to both the compatible and incompatible audiovisual stimulus pairings suggests that acoustic and visual cues for consonants are combined according to very general principles that are not specific to the perception of speech. The key aspects of the model's ability to make these predictions are the statistical independence of the auditory and visual components of the perceptual cues and the use of (near) optimal response locations.

The importance of basing decisions on statistically independent aspects of perceptual cues has long been recognized in studies of sensory communication. In studies of the ability to identify components of complex displays, it has been shown that more elements can be identified accurately if a large number of stimulus dimensions, e.g., intensity, frequency, duration are used to encode the stimuli [10]. Analysis [1] of more recent identification experiments using tactile stimuli [12] has demonstrated that this improvement in accuracy results from an increase in the dimensionality of the perceptual space, as opposed to, for example, an increase in separations in a space of fixed dimensions. Similar conclusions were reached by Lockhead [6] who analyzed the problem of how to best use a second perceptual dimension to encode a fixed number of stimuli. The statistical independence of the auditory and visual components of the perceptual cues used to identify audiovisual stimuli is consistent with the lack of perceptual interference (e.g., cross-modal masking) between these two modalities. Note that the occurrence of the McGurk effect, which demonstrates that the presentation of visual stimuli can influence the way subjects respond to auditory stimuli, does not call the assumption of statistical independence of *cue components* into question. This influence has been shown to result instead from the mean properties of the cues.

The assumption that the response centers that subjects use to identify audiovisual stimuli are located at or near optimal locations, i.e., close to the multimodal centers for the compatible stimuli, is consistent with the extensive real-life experience that subjects have with compatible audiovisual stimuli. When subjects attempt to identify unidimensional stimuli with which they are much less familiar, they can shift the locations of response centers under the control of experimental variables such as payoffs [5] and presentation probabilities [4]. The extent to which such mechanisms can induce shifts in response center when complex, highly learned, audiovisual stimuli are identified is unknown, but merits investigation.

The ability of the integration model to predict performance in audiovisual identification experiments underscores the importance of being able to predict the structure of the unimodal cue spaces, in particular, to specify the locations of the stimulus centers when the stimuli are described only in physical terms. This problem has received considerable attention, particularly in the auditory case [11]. Some progress has also been made in characterizing the effects of applying degradations to the acoustic stimuli. For example, analysis of the perceptual confusions between consonants measured by Miller and Nicely [9] suggests that changes in the signal to noise ratio have little effect on the relative spacing between stimulus centers [13]. Increasing the S/N by 6 dB roughly doubles the distances between all centers [3]. Unfortunately similar characterizations of the properties of the visual stimulus centers are comparatively sparse.

Inspection of Fig. 1 suggests that the magnitude of the McGurk effect is likely to be determined by the relative distinctiveness of the auditory and visual stimuli. Consider how much larger is the separation between B and b, which measures the distinctiveness of the $\mathcal{V}$ stimuli /b/ and /g/, than that between b and G, which measures the distinctiveness of the $\mathcal{A}$ stimuli /b/ and /g/. As noted in Sec. 4, proximity of b to G rather than to B is responsible for the failure to identify the incompatible /bg/ stimulus as "b". As seen in Fig. 2, labial stimuli are more distinct from the non-labial stimuli in the $\mathcal{V}$ set than in the $\mathcal{A}$ set. When response centers are located near the compatible stimulus centers, stimuli constructed by pairing labial visual consonants with non-labial auditory consonants, are more likely to elicit responses appropriate for labial stimuli than for non-labial stimuli.

According to this analysis, it should be possible to predict the magnitude of the McGurk effect from measurements made under unimodal presentation conditions alone. In particular, the magnitude of the McGurk effect should be reduced by reducing the ability to distinguish between labial and non-labial consonants in the $\mathcal{V}$ condition. We have attempted to do so using combinations of uniform spatial filtering, glare, and reduced brightness, but have met with only modest success. In future work we plan to consider other visual degradations, such as inhomogeneous visual noise, to achieve this end.

8. ACKNOWLEDGEMENTS

This work was supported in part by grants from the National Institutes of Health, the Japanese Ministry of Education, Science and Culture, the Nissan Science Foundation, and the Japan Science Society.

9. REFERENCES

[1] L. D. Braida. Development of a model for multidimensional identification experiments. *J. Acoust. Soc. Am.*, 84:S142, 1988.

[2] L. D. Braida. Crossmodal integration in the identification of consonant segments. *Q. J. Exp. Psych.*, 43(a)3:647–677, 1991.

[3] L. D. Braida. Integration models of intelligibility. In C. W. Nixon, editor, *Speech Communication Metrics and Human Performance*, pages 129–144. USAF Armstrong Laboratory, 1996.

[4] S. Chase, P. Bugnacki, L. D. Braida S. Chase, P. Bugnacki, L. D. Braida, and N. I. Durlach. Intensity perception. xii. effect of presentation probability on absolute identification. *J. Acoust. Soc. Am.*, 73:279–284, 1982.

[5] R. P. Lippmann, L. D. Braida, and N. I. Durlach. Intensity perception. v. effect of payoff matrix on absolute identification. *J. Acoust. Soc. Am.*, 59:129–134, 1976.

[6] G. R. Lockhead. Identification and the form of multidimensional discrimination space. *J. Exp. Psych.*, 85(1):1–10, July 1970.

[7] J. MacDonald and H. McGurk. Visual influences on speech perception processes. *Percep. Psychophys.*, 24(3):253–257, 1978.

[8] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, December 1976.

[9] G. A. Miller and P. Nicely. An analysis of perceptual confusions among some english consonants. *J. Acoust. Soc. Am.*, 27(328-352), 1955.

[10] I. Pollack and L. Ficks. Information of elementary multidimensional auditory displays. *J. Acoust. Soc. Am.*, 26:155–158, 1954.

[11] L. C. W. Pols, L. J. Th. van der Kamp, and R. Plomp. Perceptual and physical space of vowel sounds. *J. Acoust. Soc. Am.*, 46(2(2)):458–467, 1969.

[12] W. M. Rabinowitz, A. J. M. Houtsma, N. I. Durlach, and L. A. Delhorne. Multidimensional tactile displays: Identification of intensity, frequency, and contactor area. *J. Acoust. Soc. Am.*, 82(4):1243–1252, October 1987.

[13] R. N. Shepherd. Psychological representation of speech sounds. In E. E. David and P. B. Denes, editors, *Human communication: A unified view*. McGraw Hill, New York, 1972.

[14] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.*, 26:212–215, 1954.

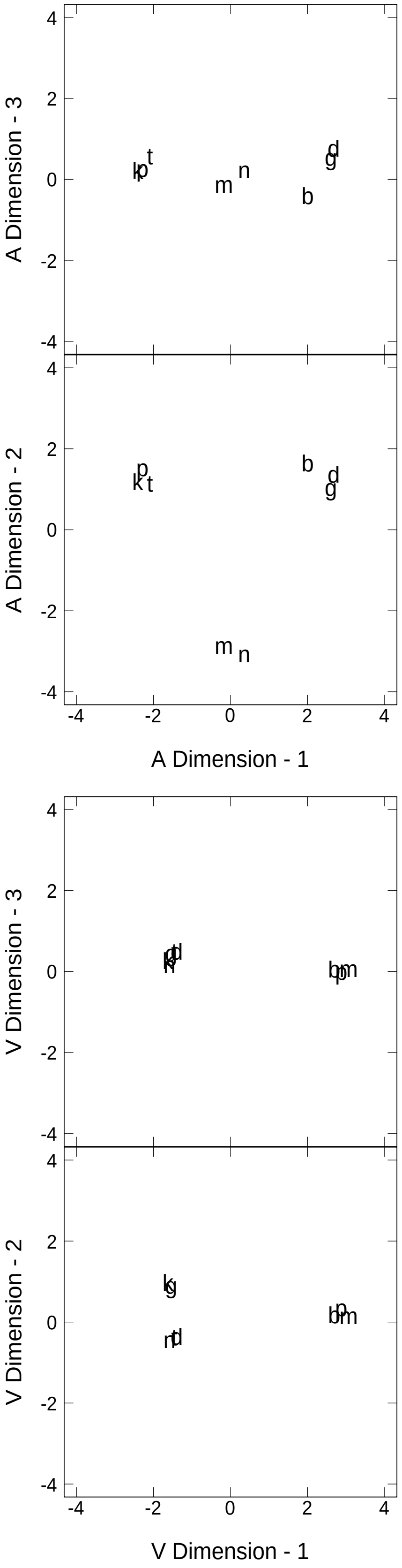

Figure 2. Three-dimensional cue spaces for the stimuli used in the $\mathcal{A}$ (upper panels) and $\mathcal{V}$ (lower panels) experiments.