Saliency Effects in Distributional Learning

Martijn Goudbeek[1,2] & Daniel Swingley[3]

[1]University of Geneva, Department of Psychology, Geneva, Switzerland;
[2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands;
[3]University of Pennsylvania, Department of Psychology, Philadelphia, USA.
goudbeek@pse.unige.ch, swingley@psych.upenn.edu

Abstract

Acquiring the sounds of a language involves learning to recognize distributional patterns present in the input. We show that among adult learners, this distributional learning of auditory categories (which are conceived of here as probability density functions in a multidimensional space) is constrained by the salience of the dimensions that form the axes of this perceptual space. Only with a particular ratio of variation in the perceptual dimensions was category learning driven by the distributional properties of the input.

1. Introduction

Learning the sounds of a language is a daunting task faced by both infants and the learners of a second language. Distributional learning is a reasonable candidate for a domain-general category discovery mechanism and almost certainly plays a central role in the acquisition of phonetic categories. In the case of the infant, such learning is necessarily unsupervised since infant listening abilities precede their speaking abilities (Juszcyk, 1997; Aslin, Juszcyk, & Pisoni, 1998).

To the extent that the members of two categories occupy distinct regions of perceptual or conceptual space, given sufficient data a distributional learner could, in principle, determine that there are two categories, and discover the important characteristics of those categories. Evidence consistent with distributional learning of visual and auditory categories has been shown even in infants (Younger 1985; Saffran, Aslin, & Newport 1996). Maye, Werker and Gerken (2002) demonstrated such learning in a laboratory setting. They exposed two groups of infants to stimuli on an artificial voice-onset-time (VOT) continuum extending from [da] to unaspirated [ta], a distinction not present in English. One group of infants listened to sounds in which the VOT followed a unimodal distribution and the other group listened to sounds that followed a bimodal distribution. Following this exposure, infants listened to stimuli that were either alternating or non-alternating stimulus sets. Only the infants that were exposed to the bimodal stimuli evidenced a preference for alternating over non-alternating stimuli. Maye and Gerken (2001, 2002) found a similar sensitivity to distributional information for adults with similar stimuli.

Infant learning of phonetic categories takes place without the benefit of feedback. Unsupervised learning has been studied in many visual category learning experiments (Ashby, Queller & Beretty, 1999; Love, 2003) but rarely in the auditory domain. In the present studies, listeners were never given trial-by-trial feedback about their categorizations. There were cues to category membership, however. In all experiments the distributional properties of the stimuli could, in principle, be used to determine the category structure and classify the stimuli accordingly. In addition to distributional information, a condition with a perfectly correlated, but temporally distinct, auditory cue to category membership was also tested.

Speech is a multidimensional signal containing regularities of many different sorts. As a result, any viable account of distributional learning of speech categories must account for the learning of categories that are defined over more than one dimension. To address this, we constructed a two dimensional perceptual space (Shepard, 1957), spanned by the dimensions *formant frequency* and *sweep rate*. Stimulus tokens were inharmonic tone complexes selected from this space of possible sounds. Formant frequency was defined as the frequency of the spectral peak of the signal. Sweep rate was defined in octaves per second as the speed with which the base frequency of the signal (F0) rose with time. The dimensions that constituted the perceptual space were equalized in terms of just noticeable difference (JND), a common procedure in psychophysical experimentation (Thurstone, 1927; Ashby & Perrin, 1988). In experiment 1, the JND for sweep rate equivalent to that found for formant frequency was determined to aid in stimulus scaling for the subsequent categorization experiments.

Formant frequency and sweep rate have been shown to be important in the speech signal. For example, formant frequency plays a crucial role in determining the identity of vowels (Ladefoged & Broadbent, 1957; Hillenbrand, Getty, Clark, & Wheeler, 1995). Fundamental frequency variation is central to prosody and is a particularly important component of child directed speech (e.g., Fernald, 1989).

2. Experiment 1: JNDs

Perceptual spaces are traditionally constructed by equalizing the JNDs of the perceptual dimensions spanning the space

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

(Shepard, 1957). The JND for formant frequency has been determined by Glasberg and Moore (1990) to be 2.98 Hz (or 0.12 ERB, a perceptual counterpart of Hz). In Experiment 1 the JND for sweep rate was determined using a same/different paradigm.

2.1. Method

2.1.1. Subjects

Thirty-seven listeners participated in the pilot experiments. In this and the following experiments, participants were drawn from the subject pool of the Max Planck Institute for Psycholinguistics and received a small payment for their participation. All were students from the University of Nijmegen and reported normal hearing. Listeners were randomly assigned to one of three conditions.

2.1.2. Stimuli

The stimuli were inharmonic tone complexes that varied in formant frequency and sweep rate. Table 1 shows the range of the stimuli tested; the size of the difference in sweep rate between the members of each stimulus pair; the specific sweep-rate differences that were tested in each condition; and the ERB level at which these differences were tested. For example, Condition 3 tested whether subjects could reliably detect differences of 1 and 2 octaves per second at two ERB levels (18.8 and 19.7) with stimuli ranging from 5 octaves per second to 15 octaves per second. A stimulus pair in this condition could thus be (5.0 octaves per second-18.8 ERB versus 6.0 octaves per second – 18.8 ERB) testing the ability to judge a difference of 1 octave per second at 18.8 ERB.

Table 1: Stimulus characteristics per condition and the experimental properties of the 3 conditions.

Cond.	Stimulus Range (oct/s)	Step Size (oct/s)	Differences Tested (oct/s)	No. of pairs	Tested ERB levels
1	2.2-2.8	0.2	0.2/0.4/0.6	192	17.9/20.6
2	5.0-15	0.5	0.5/1.5	400	18.8/19.7
3	5.0-15	1.0	1.0/2.0	400	18.8/19.7

2.1.3. Procedure

All conditions consisted of same/different judgment tasks in which half of the stimulus pairs were *same* trials and half were *different* trials. Listeners were seated comfortably in a sound-attenuated room and listened over Sennheiser headphones (HD 270). If they considered the sounds to be the same, they pressed a button labeled (the Dutch equivalent of) "same". If they considered the sounds to be different, they pressed a button labeled (the Dutch equivalent of) "different".

All conditions took about 30 minutes and participants were offered a break halfway through the experiment. The comparisons were done at two levels of ERB and sweep rate. For example, in Condition 1, the difference between 2.2 octaves per second and 2.4 octaves per second was compared at two ERB levels (18.8 and 19.7). This way, possible interactions between the two dimensions could be investigated. Differences in sweep rate were also compared at different levels to investigate possible differences in JNDs at different levels of sweep rate. For example, in Condition 3 the differences between 5.0 and 6.0 octaves per second and that between 14.0 and 15.0 octaves per second were investigated to assess the discriminability of 1 octave per second.

2.2. Results and discussion

All three conditions yielded hit and false alarm rates that were used to compute the d' values associated with each difference in sweep rate. A d' of about 1 is considered to reflect two just perceptually separable stimuli; our goal was to find a sweep rate with a d' as close to 1 as possible. Table 2 shows the results. ERB of "Low" refers to 17.9 in Condition 1 and 18.8 otherwise; "High" refers to 20.6 in Condition 1 and 19.7 otherwise, as listed in Table 1.

Table 2: Mean d' values for the judged differences in sweep rate at two different ERB levels. d' values of at least 1.0 are given in boldface.

		Condition 1			Condition 2		Condition 3	
		Difference (octave per second)						
	ERB	0.2	0.4	0.6	0.5	1.5	1.0	2.0
d'	Low	0.3	0.9	**1.1**	0.6	**1.3**	**1.9**	**3.1**
	High	0.5	0.7	**1.3**	0.4	**1.4**	**1.5**	**2.9**

A difference in sweep rate between 0.6 octaves per second (Condition 1) and 1.5 octaves per second (Condition 3) has a d' of approximately 1. Condition 1 shows that sweep rate differences below 0.4 octaves per second were difficult to distinguish. Condition 2 and 3 showed that differences in sweep rates greater than 1.5 were very easy to distinguish. Somewhere between 0.6 and 1.5 lies the d' value of 1 sought in this pilot experiment. Because the d's for 0.5 from Condition 2 were considerably lower than 1, we estimated that a sweep rate of about 1.0 octave per second would constitute a JND. This value was then used in the following category learning experiment.

Table 2 also seems to show that this perceptual space is not homogeneous: a high ERB level is associated with higher mean d' values for some sweep rate differences. However, the d's of the different ERB levels do not differ significantly (all $p > 0.18$, $t[max] = 0.95$). Given the absence of a significant difference between the higher and lower ERB rate, we took the simplifying step of using a single JND for sweep rate at all ERB levels.

3. Experiment 2: Category Learning

In Experiment 2, listeners were exposed to members of two categories separable by sweep rate (with category-irrelevant variation in formant-frequency), or separable by formant-frequency (with irrelevant variation in sweep rate). This approach is very similar to paradigms used in visual category learning research. The JND for sweep rate determined in the pilot experiments (1.0 octave per second) and the JND for formant frequency (0.12 ERB) derived from Glasberg and Moore (1990) were used to construct the stimuli.

The experiment consisted of a *learning phase* in which listeners heard sample members of each category, drawn from each category's distribution (left and middle panels of Figure 1) and a *maintenance phase* where the stimuli were positioned in an equidistantly (7 x 7) spaced grid, thus not providing any further distributional information (right panel of Figure 1).

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

Learning of auditory categories was examined under two conditions: 1) exposure to the stimuli with only the distributional properties of the categories as cues to category membership and 2) exposure to the stimuli where each stimulus was followed by an nonlinguistic auditory "label" that served as a perfectly correlated cue to category membership. To equalize both conditions in terms of auditory complexity, each stimulus of condition 1 was followed by an *uninformative* auditory label.

The orientation of the probability density functions of the learning stimuli determined the relevant and irrelevant dimension for the listeners. In the learning phase of Conditions 1 and 2, formant frequency was the relevant dimension (second panel of Figure 1) whereas in the learning phase of in Conditions 3 and 4, sweep rate was the relevant dimension (first panel of Figure 1).

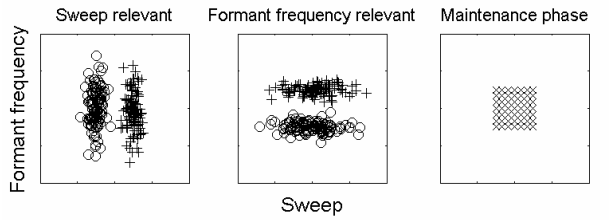

Figure 1: The design of the experiments: learning phases with distributional information and a neutral maintenance phase used to test learning.

Thus, the experiment had a between-subjects 2x2 design: two learning conditions (informative labels versus uninformative labels) by two orientations (sweep relevant versus formant frequency relevant).

3.1. Method

3.1.1. Subjects

Twenty-four participants (six in each condition) were drawn from the MPI subject pool.

3.1.2. Stimuli

The 224 learning stimuli (2 categories x 112 stimuli in each category) as well as the 49 maintenance stimuli were inharmonic sound complexes that differed in both formant frequency and sweep. The labelling sounds were two simple and easily distinguishable sounds (the sound of a book being shut played forward or backward).

3.1.3. Procedure

Listeners were seated in a soundproof booth and listened to the stimuli over Sennheiser headphones (HD 270). In the learning phase, listeners heard the distributionally defined stimuli with informative or uninformative labels, but did not categorize them. Depending on condition, either formant frequency or sweep rate was the relevant dimension of variation in the learning phase.

After the learning phase, listeners' category judgments were tested in a forced-choice category labeling task: in the maintenance phase they had to categorize 196 (49 stimuli x 4 repetitions) maintenance stimuli as they saw fit. This phase was intended to neutrally scan the categorization tendencies of the listeners without providing them with new information about the category distributions.

3.2. Results and discussion

Responses were analyzed using logistic regression (Agresti, 1990). A logistic regression analysis yields a β-weight for each predictor. The dimensions were entered as predictors for the categorization response. Figure 2 shows the mean β-weights of each dimension in each condition. It is clear that sweep rate was the dimension used whether it was the relevant dimension or not. Further, the "labeling sound" was apparently of no assistance in leading listeners to the distributional categories.

An ANOVA with Dimension (relevant versus irrelevant) as within-subjects variable and Orientation (formant frequency relevant versus sweep rate relevant) and Condition (listening versus labeling) as between-subjects variables indicated no significant effects for Dimension ($F [1,27] = 0.004$, n.s.), Orientation ($F [1,27] = 0.46$, n.s.) or Condition ($F [1,27] = 0.47$, n.s.).

Thus, listeners favored the use of rise in fundamental frequency, irrespective of whether this was the relevant dimension, and in spite of the fact that variability in the category structures was approximately equalized in JND units for both dimensions. Apparently, listeners were not sensitive to the different category structures. The interaction between Orientation and Dimension was highly significant ($F [1,27] = 139.71$, $p < 0.000$) indicating the preference for sweep rate, irrespective of whether it was the relevant condition or not.

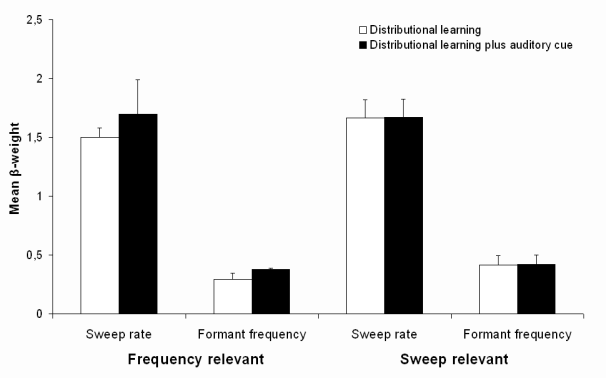

Figure 2: Mean β-weights obtained in the maintenance phase for two category structures (formant frequency. relevant or sweep rate relevant) and under two learning conditions (without label cue, or with it). Error bars represent SE.

This absence of distributional learning warranted reconsideration of the chosen JNDs. Apparently, the JNDs from the same/different paradigm did not transfer to the categorization experiment. In Experiment 3, we systematically manipulated the size of the difference in sweep rate in an attempt to find a "sweet spot" for distributional learning in a perceptual space defined by formant frequency and sweep rate.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

4. Experiment 3: finding the sweet spot

The rationale of Experiment 3 was that by systematically comparing a set of ranges of variation on the sweep dimension, we might determine a "sweet spot" in the ratio of variances in sweep and formant frequency at which distributional information in the exposure stimuli would drive listeners' category learning in the learning phase and subsequent identification decisions in the maintenance phase.

4.1. Method

4.1.1. Subjects

Twenty-four participants (4 in each of the 6 conditions) were drawn from the MPI subject pool.

4.1.2. Stimuli

The stimuli were identical to those used in Experiment 1: inharmonic sounds that differed in formant frequency and sweep rate. Depending on condition, either the variation in formant frequency was relevant for distinguishing the categories and sweep rate was irrelevant or vice versa (see the first and second panel of Figure 1). The conditions also differed in the range of variation on each dimension in units determined by the pilot experiment. With respect to the JND of Experiment 2, where the ratio in variation between the two dimensions was 1:1 (sweep rate:formant frequency), Experiment 3 tested ratios of 0.5:1 ("sweep 2" condition), 0.25:1 ("sweep 4" condition), or 0.125:1 ("sweep 8" condition).

4.1.3. Procedure

The procedure was identical to the informative-labeling conditions of Experiment 1. In the learning phase, listeners heard a stimulus that was immediately followed by an acoustic label that correlated perfectly with category membership. In the maintenance phase, listeners were asked to categorize the stimuli as they saw fit (without the acoustic labels present).

There were six experimental conditions (2 category structures x 3 levels of range of variation in sweep rate) in the experiment. Four listeners participated in each condition.

4.2. Results and discussion

The results from the maintenance phase were again analyzed with a logistic regression analysis yielding a β-weight indicating each subject's use of each dimension. Figure 3 shows the mean β-weights for all six conditions. In the Sweep 2 condition, wherein the range in variation for sweep rate was half as great as it was in Experiment 2, listeners still had a higher β-weight for sweep rate, irrespective of whether it was the relevant dimension or not.

When the range in variation for sweep rate was lowered to 0.125 times that of the original range (i.e., in the Sweep 8 condition), however, the variation in sweep rate was apparently too small and the β-weight for formant frequency was higher than that for sweep rate, independently of its relevance to the category structure.

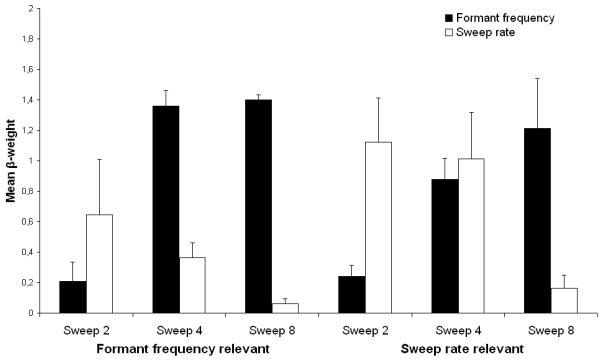

Figure 3: The mean β-weight obtained per orientation (formant frequency or sweep rate relevant) per amount of variation of sweep rate (0.5; 0.25; and 0.125 of the original 1 octave per sec.).

Finally, with a sweep rate of 0.25 octaves per second the relevant dimension was the one that was used most by listeners. In other words, only when the JND for sweep rate was 0.25 of the original, the experimental manipulations were not washed out by the differences in salience of the different dimensions. The effect was still quite small when sweep rate was the relevant dimension, but compared to when formant frequency was relevant, the differences were considerable. An ANOVA with Dimension (relevant versus irrelevant) as within-subjects variable and Orientation (formant frequency relevant versus sweep rate relevant) as between-subjects variable and the β-weights of each dimension a dependent variables indicated a significant main effect of Dimension ($F[1,6] = 5{,}87$, $p < 0.05$). This shows that listeners were able to determine and use the relevant dimension in their categorizations.

In sum, the stimulus set of condition "Sweep 4", constructed by considering 0.25 octaves/sec as comparable to 0.12 ERB, was the only set that yielded significant effects of the category learning phase. Doubling that variation in the sweep-rate dimension (condition Sweep 2) led listeners to attend primarily to sweep rate; halving that variation (Sweep 8) led listeners to attend primarily to formant frequency. The numerical pattern of data suggests that the true sweep-rate match to a JND of 0.12 ERB might be somewhat higher than 0.25 octaves/sec, given that at this rate listeners over-used the frequency dimension when sweep rate was relevant for categorization.

5. General discussion

These experiments used two important dimensions for speech recognition to show that category structures of identical formal separability varied in their learnability according to how variation in the dimensions was scaled. The results also show the validity of the logic of conducting consecutive categorization experiments with differing sweep rates to find the "sweet spot" where the distributional characteristics of the stimuli *do* drive category learning and subsequent identification. There may be such a "sweet spot" for all combinations of perceptual dimensions. However, some dimensions are likely to be more susceptible to learnability differences due to scaling than others.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

Explanations for why some dimensions have a smaller and more unstable sweet spot compared to others may turn on the extent to which a dimension is verbalizable. Ashby, Alfonso-Reese, Turken and Waldron (1998) deployed a similar argument with regard to categorization rules. Rules that are easier to verbalize are used more often and guesed at sooner in unsupervised categorization than rules that are harder to verbalize. Analogously, dimensions that are easier to verbalize could be less vulnerable to relative range of variation effects.

Another line of reasoning is that the difference in dynamic properties is involved in the relative range effect. Dynamic dimensions tend to be salient. Because sweep rate is dynamic and formant frequency is not, the finetuning of these dimensions to one another involves more than just equalizing their JNDs. The JND for sweep rate has to be low enough to compensate for the salience it has compared to formant frequency due to its dynamic properties. JNDs obtained in a discrimination task might not transfer to the JNDs in a categorization task (see e.g. Nosofsky, 1986 and Ashby & Lee, 1991), at least not when distributional learning is at stake.

In sum, while distributional learning may in principle be a universal learning mechanism, it is constrained by saliency biases of the listeners.

6. Acknowledgements

We would like to thank Marloes van der Goot and Maarten Jansonius for their help in recruiting the participants and conducting the experiments. Roel Smits contributed substantially to this work, and we thank Anne Cutler for helpful suggestions on the manuscript.

7. References

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons, Inc.

Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U., & Waldron, E.M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442-481.

Ashby, F. G. & Lee, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General*, *120*, 150-172.

Ashby, F. G., & Perrin, N. A. (1988). Towards a unified theory of similarity and recognition. *Psychological Review*, *95*, 124–130.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*, 1178-1199.

Aslin, R.N., Jusczyk, P.W., & Pisoni, D.B. (1998). Speech and auditory processing during infancy: Constraints on and precursors to language. In D. Kuhn & R. Siegler (Eds.), *Handbook of Child Psychology*, (5th ed., Vol. 2, pp. 147-198). New York: Wiley.

Fernald, A (1989). Intonation and communicative intent in mothers'speech to infants. Is the melody the message? *Child Development*, *60*, 1497-1510.

Glasberg, B.R., & Moore, B.C.J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*, 103-138.

Hillenbrand, J., Getty, L., Clark, M., & Wheeler, K. (1995). Acoustic characteristics of American English vowels, *Journal of the Acoustical Society of America*, *97*, 3099–3111.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Ladefoged, P. & Broadbent, D.E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*, 98-104.

Love, B.C. (2003). The multifaceted nature of unsupervised category learning. *Psychonomic Bulletin & Review*, *10*, 190-197.

Maye, J. & Gerken, L. (2000). Learning phoneme categories without minimal pairs. *Proceedings of the 24th Annual Boston University Conference on Language Development*: 522-533. Sommerville, MA: Cascadilla Press.

Maye, J. & Gerken, L. (2001). Learning phonemes: How far can the input take us? In A. H-J. Do, L. Domínguez, & A. Johansen (Eds.), *Proceedings of the 25th Annual Boston University Conference on Language Development*. *480-490*. Sommerville, MA: Cascadilla Press.

Maye, J., Werker, J., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101-B111.

Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.

Saffran, J., Newport, E., & Aslin, R. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606--621.

Shepard, R.N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325-45.

Thurstone, L.L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 278-286.

Younger, B. (1985). The segregation of items into categories by ten-month-old infants. *Child Development*, *56*, 1574-1583.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review