

Categorical perception of speech without stimulus repetition

Jack C. Rogers, Matthew H. Davis

MRC Cognition and Brain Sciences Unit, Cambridge, UK

jack.rogers@mrc-cbu.cam.ac.uk, matt.davis@mrc-cbu.cam.ac.uk

Abstract

We explored the perception of phonetic continua generated with an automated auditory morphing technique in three perceptual experiments. The use of large sets of stimuli allowed an assessment of the impact of single vs. paired presentation without the massed stimulus repetition typical of categorical perception experiments. A third experiment shows that such massed repetition alters the degree of categorical and sub-categorical discrimination possible in speech perception. Implications for accounts of speech perception are discussed.

Index Terms: speech perception, perceptual learning.

1. Introduction

Categorical perception traditionally implies that two stimuli on opposite sides of the category boundary are heard as different whilst two stimuli with an equivalent physical difference, but on the same side of a category boundary, will be heard as the same [1]. Although better discrimination of between- than within-category differences is a classic and widely-replicated finding in studies of speech perception, the question of whether discrimination is achieved by using a categorical representation of the speech sound remains controversial [2]. If this is the case then this would suggest a very different mode of perception than is classically observed for pitch and/or loudness judgments of stimuli for which discrimination performance typically far exceeds categorisation [3].

One reason for supposing that a different mode of perception might apply to speech stimuli is that listeners must maintain a large number of stable perceptual categories (e.g. phonemes) in the face of an often highly variable acoustic input. It has been shown, for instance, that many different acoustic cues are sufficient to convey a specific phonemic contrast [4], with no single cue being necessary for correct perception. Such findings suggest that categorical perception effects arise from a flexible and dynamically-changing perceptual system rather than from bottom-up stimulus sensitivity. Consistent with this, a number of contextual factors have been shown to influence phonemic category boundaries. For instance, changes to the range of speech stimuli presented in a specific experiment can alter the observed category boundary [5]. Furthermore, the degree of tuning or sharpness of the perceptual categorisation function depends on the distribution of speech tokens presented to participants [6]. Under conditions of uncertainty, listeners will make optimal perceptual decisions based on Bayes rule (cf. [7]).

Lexical information provides one source of information that can potentially guide perceptual tuning. This is consistent with changes to phoneme category boundaries observed in the “Ganong effect” [8]. Here the effect of lexical context results in ambiguous phonemes being perceived differently if one interpretation would make a familiar word. For example, an ambiguous /g/-/k/ segment would be perceived as /g/ in the

context of the word *gift* and /k/ in the context of the word *kiss* [8]. Presentation of as few as twenty Ganong-like stimuli can produce long-term changes to phoneme category boundaries when measured outside of a lexical context [9]. A controversial interpretation of these findings is that top-down processes are responsible for generating and maintaining categorical perception in the face of highly variable speech input [10].

Given the rapid operation of these perceptual learning processes, the traditional method of assessing categorical perception in which a limited set of stimuli are repeated multiple times might overestimate categorical and non-categorical influences on speech perception. Here we use an automated audio-morphing technique to generate a large number of high-quality phonetic continua derived from naturally-recorded speech tokens. These morphed stimuli can be characterised on a linear acoustic scale where a 0% stimulus is identical to speech token 1, 100% is identical to token 2, and a 50% stimulus is equally similar to the two source speech tokens. This percentage scale allows responses to be averaged over multiple different stimuli, including different phonetic changes in different lexical contexts. In this way we can explore perceptual discrimination and categorisation of spoken syllables in the absence of massed repetition of any specific stimulus or phonetic continua.

1.1. Automatic morphing of speech materials

1.1.1. Morphing method

“Straight” is a channel vocoder [11] that provides high quality speech analysis and resynthesis by decomposing speech into three representations: (1) a periodic glottal source, (2) aperiodic excitation, and (3) a dynamic spectral filter. By averaging these three representations for two speech tokens in different proportions we can create natural sounding intermediate tokens (blending the spectral filter of the source tokens in different proportions). For this procedure to work, however, corresponding time-points in both tokens must be averaged. This requires time-consuming and error-prone hand-marking of aligned positions in the two speech tokens. For monosyllables differing in a single phonetic feature, good alignment can be achieved automatically using dynamic time warping. This involves computing a time-slice by time-slice similarity matrix (cosine distance) between the spectral profiles of the two source tokens, then computing the maximum similarity path through this matrix using dynamic programming. Evenly spaced positions in token 1 (at 50ms intervals) can then be mapped onto a maximally-similar corresponding position in token 2. In combination, this provides an automated method of creating phonetic continua and allows us to use the percentage blend between two tokens as a dependent measure in combining responses to different continua.

2. Method

2.1. Materials

The experimental materials were generated by morphing a set of 320 monosyllabic single-feature minimal pairs from 3 lexical conditions (80 word pairs, e.g. *blade* - *glade*, 160 word-nonword pairs, e.g. *blouse* - *glouse*, *bown* - *gown*, and 80 non-word pairs e.g. *blem* - *glem*). Phonetic changes were made at syllable onset (*blade* - *glade*), offset (*tub* - *tug*), or medially (*flute* - *fruit*) and changes were made to voicing (/b/-/p/, /s/-/z/), manner (/b/-/w/) or place (/b/-/d/-/g/) of articulation. The original syllables were recorded by a male native English speaker, “Straight” morphing used to generate 10 intermediate tokens at 10% steps from 5% (similar to word 1 in the pair) to 95% (similar to word 2). Based on categorisation performance (Experiment 1), a subset of 96 pairs were chosen for a discrimination experiment (Experiment 2). This subset consisted of 24 word pairs, 48 word-non-word pairs (counterbalancing for lexical and phonetic status) and 24 non-word pairs, all with category boundaries between 35% and 65% and equal numbers of onset and offset changes. In Experiment 3, this set of stimuli was further reduced to two word pairs (*cone* - *tone* and *oat* - *oak*).

2.2. Experiment 1: Categorisation

2.2.1. Procedure

Twenty native British-English speakers each listened to 3200 morphed syllables (10 tokens for all 320 minimal pairs) with the different phonetic continua intermixed and presented in a pseudo-random order (approximately 100 minutes of testing with regular breaks). Two response alternatives were presented on screen 500ms after the offset of each speech token, with the critical phoneme underlined (e.g. blade - glade). Participants also indicated if they heard something other than the two response alternatives.

2.2.2. Results

Individual responses faster than 300ms or slower than 5000ms (0.5% of the data) and trials in which neither of the two possible alternatives was heard (0.7% of trials) were excluded. The proportion of responses matching either of the end points of the continuum was averaged over both continua and participants (shown in Figure 1). Results show the expected sigmoidal profile, with the majority of responses matching the most similar word option and a relatively sharp transition between the two response alternatives at the intermediate steps in the continuum. Differences between three lexical conditions are suggested with an increased number of ‘real’ word responses for the word–non-word pairs (Figure 1). This shift is analogous to the classic “Ganong effect” ([8]), but is observed here without stimulus repetition that could induce long-term shifts in the phonetic category boundary.

To quantify this lexical shift we used logistic regression to compute the position of the category boundary for each phonetic continua (averaging over responses from all participants), and for each participant (averaging over responses to phonetic continua within each lexical condition). These regressions allow us to use the point in the morphed continua at which 50% responses were generated as the dependent measure (see inset graph Figure 1), computing ANOVAs by participants (*F1*) and items (*F2*). Analysis revealed a significant main effect of lexical condition (*F1* (2,

38) = 25.27, $p < 0.001$; *F2* (2,317) = 8.36, $p < 0.001$) with a highly-significant lexically-driven shift in the category boundary for word-nonword pairs. Within the sets of pairs we can distinguish those in which phonetic changes occurred at the onset or offset of the syllable (excluding pairs with medial changes). Analysis revealed a significant main effect of onset versus offset change for subjects (*F1* (1, 19) = 18.05, $p < 0.01$) but not items (*F2* (1, 298) = 2.49, $p = 0.116$), with no significant interaction between lexical condition and position of phonetic change (*F1* and *F2* < 1). The lack of any significant interaction demonstrates that lexical effects on phoneme perception are equivalent when lexical information precedes or follows the ambiguous phoneme. Hence, this finding suggests that phoneme identification is in this case delayed until after words are recognised.

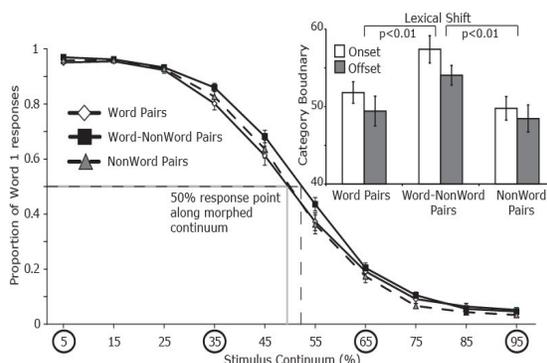


Figure 1. Proportions of responses matching word 1 along the phonetic continuum for all 3 lexical conditions averaged over onset and offset changes. Inset bar-graph shows estimated position of the category boundary as a function of lexical status for phonetic changes at syllable onset or offset (error bars show the standard error of the mean over items).

2.3. Experiment 2: Discrimination

2.3.1. Procedure

Twenty new participants made a speeded ‘same/different’ judgement on two syllables from 96 of the minimal pairs used in Experiment 1 (see Materials). Each syllable was presented in one of 4 morph proportions (5, 35, 65 and 95, circled in Figure 1). Hence, participants heard either two identical syllables successively or two acoustically-different syllables with a 30%, 60% or 90% acoustic difference. Categorical perception would lead to differential responses for 30% acoustic differences that do or don’t cross phoneme category boundaries (i.e. 35-65 trials vs. 5-35 and 65-95 trials). There were 16 trials for each of the 96 phonetic continua (1536 trials in total). Participants were instructed to indicate whether they heard two identical syllables (same), or not (different) on each trial. The inter-stimulus interval on specific trials was either 200ms or 800ms, and trials were counterbalanced across the two ISIs in 2 experimental versions. This ISI manipulation allows us to assess the time-course of within- and between-category discrimination. Previous studies have shown that acoustically-detailed syllable representations are only transiently held in echoic memory since within-category influences are weaker at longer ISIs (cf. [12]).

2.3.2. Results

Here we focus on responses to pairs of stimuli that contain a 30% acoustic difference but are heard within a single phonetic category (5/35, or 65/95 pairs) as well as pairs that contain an equivalent acoustic difference but that cross the category boundary (35/65 pairs). As shown in Figure 2, participant's responses show the classical categorical perception effect, with an increase in discrimination sensitivity (d') for stimulus pairs that straddle the category boundary in each of the lexical conditions and for both ISIs (200/800ms).

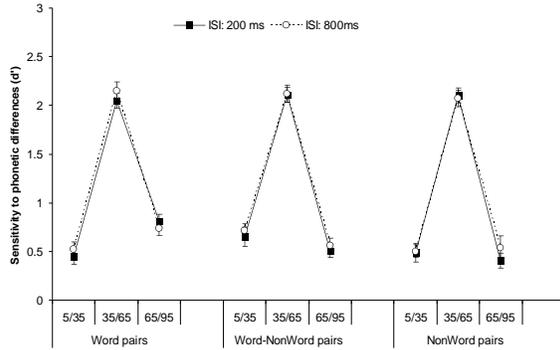


Figure 2. Discrimination sensitivity for 30% acoustic changes within (5/35, 65/95), or between phonetic categories (error bars show the standard error of the mean over subjects ($n=20$)).

A 2 (ISI: 200/800ms) \times 3 (lexical conditions) \times 3 (trial-type: 5/35, 35/65, 65/95) repeated measures ANOVA revealed a significant main effect of within- versus between-category discrimination ($F_1(2, 38) = 514.43, p < 0.001$; $F_2(2, 276) = 327.67, p < 0.001$). Those pairs that cross the category boundary are more reliably distinguished. Analysis revealed no main effect of lexical condition (F_1 and $F_2 < 1$) or ISI (F_1 and $F_2 < 1$) implying that discrimination was equivalent overall in all three lexical conditions. Interestingly, there was a significant interaction between lexical condition and trial-type for subjects ($F_1(4, 76) = 4.86, p < 0.01$) but not items ($F_2 < 1$), post-hoc comparisons suggesting better within-category discrimination for word stimuli than non-word stimuli. In contrast to previous data [12], there was no interaction between ISI and trial-type (F_1 and $F_2 < 1$), implying that participants were equally insensitive to within-category variation at both short (200ms) and long (800ms) ISIs.

2.4. Experiment 3: Effects of stimulus repetition

2.4.1. Procedure

To assess whether the insignificant main effect of ISI on within-category sensitivity was due to limited stimulus repetition in Experiment 2, a third experiment was run using just two word pairs. The same procedure was used as in Experiment 2 but there were 416 presentations of each of two word pairs, rather than 16 presentations of 96 pairs in Experiment 2. A new set of 20 participants were tested.

2.4.2. Results

Within and between-category discrimination of pairs, with a 30% acoustic change, is shown in Figure 3 (right-most graph) along with the results from Experiment 2, and predicted discrimination performance from Experiment 1. For Experiment 3, analysis by items was not possible but a 2 (ISI: 200/800ms) \times

3 (trial type: 5/35, 35/65, 65/95) repeated measures ANOVA on subjects responses revealed the expected improvement in between-category discrimination ($F(2, 38) = 300.44, p < 0.001$). There was also a significant main effect of ISI ($F(1, 19) = 14.44, p < 0.01$), but no interaction between trial-type and ISI ($F < 1$). Thus it appears that participants showed improved discrimination of both within- and between-category differences at a short ISI. These results imply that initial acoustic comparisons preceding phonetic perception can contribute to between as well as within-category discrimination, consistent with certain previous findings [12], [13].

2.5. Stimulus and task effects on speech perception

All three experiments measured the perception of categorical and sub-categorical acoustic-phonetic changes to speech stimuli. By combining the data from the three experiments we can assess perceptual discrimination under different task conditions (categorisation vs. discrimination, i.e. Expt 1 vs. Expt 2), and different stimulus presentation conditions (with/without massed stimulus repetition, i.e. Expt 2 vs. Expt 3). Figure 3 presents data from all three experiments in the form of predicted or observed discrimination performance on word pairs with a 30% acoustic change (within- and between-category). In the case of Experiment 1, discrimination performance was predicted from categorisation responses for the 24 word pairs used in Experiment 2.

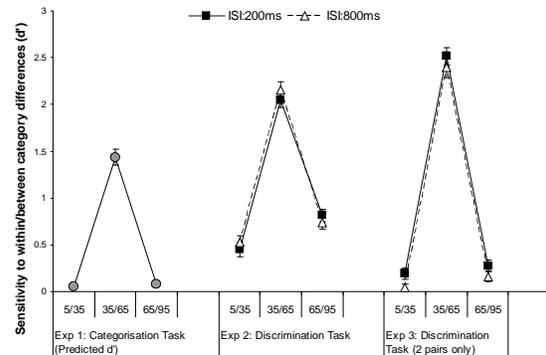


Figure 3. Discrimination sensitivity (d') across all 3 experiments. Experiment 1 predicted discrimination (d') values derived from categorisation performance for word pairs. Experiment 2 observed discrimination (d') values for 24 word pairs presented with minimal stimulus repetition. Experiment 3: discrimination (d') values for 2 word pairs with multiple repetitions. Results from short and long ISIs presented separately for Expt 2 & 3. (Error bars show the standard error of the mean across subjects).

By comparing the results from Experiment 1 and 2, it appears that discrimination of both within- and between-category differences shows the same profile, but was more accurate than predicted by categorisation responses. This effect can be quantified statistically by comparing observed discrimination performance for participants in Experiment 2 with the predicted performance from Experiment 1 for all 96 minimal pairs. Results revealed the expected effect of trial type (5/35, 35/65, 65/95) reflecting significantly enhanced between-category discrimination ($F_1(2, 76) = 929.26, p < 0.001$; $F_2(2, 372) = 454.46, p < 0.001$). However, this was modulated by both a main effect of Experiment ($F_1(1, 38) = 215.34, p < 0.001$; $F_2(1, 186) = 54.21, p < 0.001$) and a trial-type by experiment interaction ($F_1(2, 76) = 3.83, p < 0.05$; $F_2(2, 372) = 3.14, p < 0.05$). This reflects the fact that

discrimination more generally, and particularly within-category discrimination, is enhanced by the presentation of stimulus pairs in succession. This may be explained as resulting from the opportunity for direct acoustic comparisons between stimulus pairs in Experiment 2.

In addition to these short-term effects of stimulus repetition that enhance discrimination, comparisons of Experiments 2 and 3 also revealed significant effects of massed repetition of specific speech tokens. Stimulus repetition produced two opposite effects on discrimination – an enhancement of between-category discrimination and a reduction of within-category discrimination (see right two panels of Figure 3). This was confirmed with a between-experiment ANOVA revealing a main effect of trial-type ($F(2, 76) = 485.15$, $p < 0.001$), and a significant trial type by experiment interaction ($F(2, 76) = 23.69$, $p < 0.001$) reflecting these two opposite effects of stimulus repetition on discrimination in Experiment 3. The results of Experiment 3 suggest that these effects are modulated by ISI with reduced sensitivity to acoustic changes at longer ISIs. The effect of ISI on these acoustic influences appears to arise as a consequence of massed stimulus repetition reflected in a significant interaction between ISI and experiment ($F(1, 38) = 7.74$, $p < 0.01$). Thus, the apparent reduction of acoustic/echoic influence on discrimination may not arise through a process of passive acoustic decay, but rather through a form of top-down reinterpretation or perceptual ‘cleaning-up’, enhanced by stimulus repetition.

3. Discussion

Results show that many classic categorical perception findings can be replicated in the absence of massed stimulus repetition. Experiment 1 shows a strong effect of lexical status on phonetic categorisation, with a bias towards real words for acoustically ambiguous word-non-word tokens close to the category boundary. The lack of any change in the size of this Ganong effect [8], as a function of whether phonetic changes occurred at syllable onset or offset, suggests that these lexical influences on categorical perception are not produced on-line. If this were the case, we would predict larger effects at syllable offset when lexical information is already available. Instead, the significant Ganong effect at syllable onset is consistent with a post-perceptual locus of this lexical influence, perhaps due to top-down processes.

Discrimination responses (d') to successive syllables presented in Experiment 2 show both better discrimination of between-category changes and some residual discrimination of within-category changes. Comparing predicted discrimination values from categorisation responses in Experiment 1 with discrimination responses in Experiment 2 revealed a significantly improved discrimination in Experiment 2. This finding suggests that deciding whether two acoustically different stimulus pairs are the same or different is enhanced by an acoustic comparison process only available when two speech tokens are presented in quick succession. Interestingly, this effect was not modulated by ISI in Experiment 2 suggesting that this acoustic influence on discrimination does not decay over the time intervals tested here. Ongoing work is exploring whether presentation of distractor stimuli (vowel sounds) between pairs of syllables serves to reduce this acoustic influence on discrimination of paired syllables.

The results from Experiment 3 confirm that many of the effects observed in the existing literature on categorical perception appear to depend on stimulus repetition. For instance, stimulus repetition serves to reduce discrimination of within-category differences, and to enhance between-category

discrimination. Furthermore, in this experiment discrimination was modulated by the length of delay (ISI) between the first and second syllables. Both within- and between-category discrimination was enhanced at short ISIs. This finding contrasts with the null effect of ISI in Experiment 2 and suggests that certain effects in the behavioural literature concerning the time-course of categorical perception ([12], [13]) might result from top-down processes that are enhanced by stimulus repetition.

These differences between results obtained with and without stimulus repetition clearly illustrates the importance of studying speech perception under naturalistic listening conditions that do not involve massed stimulus repetition. The automated audio morphing procedure employed here opens up new research opportunities in two ways. Firstly, we can produce large numbers of natural sounding speech tokens in an automated fashion, such that speech perception experiments can be conducted without repeating stimuli. Secondly, by generating phonetic continua along a linear acoustic scale between two naturally-recorded tokens we have a natural means by which to combine and compare perception of stimuli in different lexical or phonetic contexts.

4. Acknowledgements

We would like to thank Hideki Kawahara for providing the Straight vocoder that enabled these experiments. We also acknowledge funding from the UK Medical Research Council, U.1055.04.013.

5. References

- [1] Liberman, A. M., Harris, K.S., Hoffman, H.S., and Griffith, B.C. “The discrimination of speech sounds within and across phoneme boundaries”. *J. Exp. Psychol.* 54, 358-368, 1957.
- [2] Harnad, S., “Categorical Perception: The Groundwork of Cognition”. Cambridge University Press, Cambridge, UK, 1986.
- [3] MacMillan, N.A. “Beyond the categorical/continuous distinction: a psychophysical approach to processing modes”. In: Harnad, S. (Ed.), *Categorical Perception*. Cambridge University Press, Cambridge, UK, 1986.
- [4] Bailey, P.J., and Summerfield, Q. “Information in speech: observations on the perception of [s]-stop clusters”. *J. Exp. Psychol. Hum. Percept. Perform.* 6, 536-563, 1980.
- [5] Brady, S.A., and Darwin, C.J. “Range effect in the perception of voicing”. *J. Acoust. Soc. Am.* 63, 1556-1558, 1978.
- [6] Clayards, M., Tanenhaus, M.K., Aslin, R.N., and Jacobs, R.A. “Perception of speech reflects optimal use of probabilistic speech cues”. *Cognition*, 108, 804-809, 2008.
- [7] Norris, D. and McQueen, J.M. “Shortlist B: A Bayesian model of continuous speech recognition”. *Psychological Review*, 115(2), 357-395, 2008.
- [8] Ganong 3rd, W.F. “Phonetic categorization in auditory word perception”. *J. Exp. Psychol. Hum. Percept. Perform.* 6, 110-125, 1980.
- [9] Norris, D., McQueen, J.M., and Cutler, A. “Perceptual learning in speech”. *Cogn. Psychol.* 47, 204-238, 2003.
- [10] Davis, M.H., and Johnsrude, I.S. “Hearing speech sounds: Top-down influences on the interface between audition and speech perception”. *Hearing Research* 229, 132-147, 2007.
- [11] Kawahara, H., Masuda-Katsuse, I., and de Cheveigne, A. “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction”. *Speech Commun.*, 27, 187-207, 1999.
- [12] Howell, P. “Syllabic and phonemic representations for short-term memory of speech stimuli”. *Percept. Psychophys.* 24, 496-500, 1978.
- [13] Pisoni, D.B., and Tash, J. “Reaction times to comparisons within and across phonetic categories”. *Percept. Psychophys.* 15, 285-290, 1974.