Journal of Phonetics 61 (2017) 1–12

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Research Article

An evidence accumulation model of acoustic cue weighting in vowel perception

Gabriel Tillman [a,*], Titia Benders [b,c], Scott D. Brown [a], Don van Ravenzwaaij [a,d]

[a] School of Psychology, University of Newcastle, Callaghan, NSW 2308, Australia
[b] ARC Centre or Excellence in Cognition and its Disorders, Macquarie University, Australia
[c] Department of Linguistics, Macquarie University, Australia
[d] Faculty of Behavioural and Social Sciences, University of Groningen, Netherlands

ARTICLE INFO

Article history:
Received 3 June 2016
Received in revised form
26 October 2016
Accepted 1 December 2016
Available online 23 December 2016

Keywords:
Phoneme categorization
Linear ballistic accumulator
Response time

ABSTRACT

Listeners rely on multiple acoustic cues to recognize any phoneme. The relative contribution of these cues to listeners' perception is typically inferred from listeners' categorization of sounds in a two-alternative forced-choice task. Here we advocate the use of an evidence accumulation model to analyze categorization as well as response time data from such cue weighting paradigms in terms of the processes that underlie the listeners' categorization. We tested 30 Dutch listeners on their categorization of speech sounds that varied between typical /ɑ/ and /aː/ in vowel quality (F1 and F2) and duration. Using the linear ballistic accumulator model, we found that the changes in spectral quality and duration lead to changes in the speed of information processing, and the effects were larger for spectral quality. In addition, for stimuli with atypical spectral information, listeners accumulate evidence faster for /ɑ/ compared to /aː/. Finally, longer durations of sounds did not produce longer estimates of perceptual encoding time. Our results demonstrate the utility of evidence accumulation models for learning about the latent processes that underlie phoneme categorization. The implications for current theory in speech perception as well as future directions for evidence accumulation models are discussed.

1. Introduction

Phonemes are linguistic representations with an acoustic counterpart that can be characterized in a multidimensional acoustic space. Values along each acoustic dimension can serve as cues for listeners to recognize a speech sound as a particular phoneme. The cues, such as first (F1) and second (F2) formant frequency, duration, and fundamental frequency, are acoustic and continuous. Yet, these cues map onto phonological representations that may not be continuous, i.e., the phonemes. Phonemes can be viewed as clusters of exemplars in a multidimensional phonetic space (Pierrehumbert, 2001), or as abstract representations that are connected to a range of values along multiple phonetic dimensions (Boersma, 2007). Speech perception is the process of mapping the continuous acoustic information onto the phonological categories (Holt & Lotto, 2010).

Each phoneme correlates with multiple acoustic dimensions (Lisker, 1986) and multiple acoustic cues influence each phoneme categorization (Holt & Lotto, 2006). Some cues contribute strongly to a listener's decision and some cues contribute weakly to the decision – a phenomenon called *cue weighting*. Cue weighting in speech perception often reflects the reliability of the cues for the recognition of phonological categories in the ambient language (Holt & Lotto, 2010).

Researchers investigate cue weighting using a range of methods: computational statistical modeling (McMurray, Aslin, & Toscano, 2009; Toscano & McMurray, 2010), eye-tracking (Reinisch & Sjerps, 2013), neuro-physiological measurements (Lipski, Escudero, & Benders, 2012), in normal-hearing and hearing-impaired populations (Winn, Chatterjee, & Idsardi, 2012; Winn, Rhone, Chatterjee, & Idsardi, 2013), and most commonly, with behavioral data from phoneme categorization tasks (Repp, 1982). In the latter, researchers systematically vary the acoustic cue values of sounds that are played to participants and observe the effects on phoneme categorization. Cue weighting is measured by how much each cue contributes to the categorization response and is therefore based on a measure at the end of processing and decision-making. To use categorization data to learn how acoustic cues are connected with phonological categories, we have to make the assumption that categorization data directly reflects the mapping of the experimentally manipulated cues onto the phonological categories. However, there are two fundamental issues with this assumption.

* Corresponding author.
E-mail address: gabriel.tillman@newcastle.edu.au (G. Tillman).

0095-4470/$ - see front matter © 2016 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2016.12.001

The first problem is that cue weighting is measured for a phoneme contrast and does not give us the association between cues and each category separately (i.e., the cue-to-*one*-phoneme mapping). For example, a cue that is strongly associated with one phoneme in the contrast and only loosely associated with the other phoneme can appear to be indiscriminately 'heavily weighted', because the cue contributes relatively strongly to the decision between these two phonemes. Given this confound, it is difficult to infer how much each acoustic cue contributes to each individual phoneme in the contrast.[1]

The second problem is that researchers only observe the association between experimentally manipulated cues and overt behavioral responses (i.e., the cue-to-response association), which means they need to assume that this association directly reflects the cue-to-phoneme mapping. Yet, a strong cue-to-phoneme mapping may not manifest as a strong cue-to-response association. One reason for a weak association between cues and responses despite a strong mapping could be that listeners do not have good access to the cue. Perhaps the cue is not always loud enough to be perceived or perhaps the cue appears late in the speech signal. Cues that appear later in the signal might be strongly associated with a phoneme, but may not appear as such in a categorization task because earlier appearing cues have already been processed and potentially determined the response (cf. McMurray, Clayards, Tanenhaus, & Aslin, 2008; Reinisch & Sjerps, 2013). In order to address this issue, it is necessary to learn more about how listeners process the acoustic cues. For instance, is cue weighting as inferred from categorization data driven by differences in when cues are available in time, or by listeners processing one acoustic cue faster than another? In any case, researchers need a way to investigate such latent processes in order to derive more accurate conclusions about acoustic cue weighting in terms of cue-to-phoneme mapping.

Both problems limit our ability to use categorization data to learn about how listeners map acoustic information onto phonological categories. Therefore, we need a method to account for how acoustic cues are cognitively processed for each phoneme in the contrast. Below we discuss response times (RT) and eye-tracking, which are alternative measures to categorization data that give insight into the processing of acoustic information, but neither of these measures address both issues.

First, researchers can use the RT associated with phonological decisions to investigate phoneme perception. For example, researchers have investigated processing differences between non-identical and identical phonemes (Pisoni & Tash, 1974) and have determined that phoneme categorization decisions depend more on a phoneme's position in acoustic space than their perceived category goodness (Miller, 2001).

However, there are difficulties with analyzing either choice data or RT in isolation. We know that the accuracy of a decision depends on how fast the decision is made – in other words, a participant's speed-accuracy trade-off setting (e.g., Heitz, 2014; Luce, 1986; Wickelgren, 1977). Without any insight into the trade-off settings used by participants, researchers may draw incorrect conclusions from choice or RT data alone. Furthermore, to analyze RT researchers typically average over all observations for each participant in order to subject the means to a statistical test, such as ANOVA. Analyzing the RT in this manner can lead to researchers drawing incorrect conclusions (e.g., Ashby, Maddox, & Lee, 1994; Curran & Hintzman, 1995; Heathcote, Brown, & Mewhort, 2000) and does not allow researchers to learn about the latent cognitive processes involved in speech perception. For example, an RT of 700 ms on a given trial suggests that 700 ms was needed to perceptually encode the sound, decide what phoneme was heard, and execute a motor response. But, we cannot know how long each of these processes takes from analyzing mean RT with linear models. Given that RT is a measure at the end of processing, analyzing RTs alone only inform researchers about the cue-to-response association but not the cue-to-phoneme mapping.

Eye tracking is an another useful measure that is frequently used to observe how listeners process experimentally manipulated cues online (e.g., Allopenna, Magnuson, & Tanenhaus, 1998). For example, eye-tracking can be used to infer whether the order in which acoustic cues become available to listeners affects listeners' interpretation of the speech signal (McMurray et al., 2008; Reinisch & Sjerps, 2013). In fact, McMurray et al. (2008) showed that listeners do not wait for cues that are available later in a speech signal (e.g., vowel duration) to begin using earlier available cues (e.g., voice onset time). Moreover, Reinisch and Sjerps (2013) showed that listeners use vowel spectral cues before vowel duration cues, because listeners need to wait for the vowel offset before they have full information about the duration.

Eye-tracking data, like RTs, are typically averaged over all observations for each participant, meaning that the aforementioned objections against inferences from averaged data hold for eye-tracking data as well. Furthermore, eye-tracking data are subject to the first confound of categorization data discussed in detail above. That is, they can give insight into cue-weighting, but do not give the cue-to-one-phoneme mapping for phoneme contrasts.

Categorization, RT, and eye-tracking are all useful methods in speech perception research, but none of them address both the cue-to-one-phoneme mapping and the cue-to-phoneme mapping issues discussed above. In this paper, we advocate the simultaneous analysis of phoneme categorization data with their associated RTs using an evidence accumulation model (e.g., Brown & Heathcote, 2008; Ratcliff & McKoon, 2008; Usher & McClelland, 2001). The following section describes what evidence accumulation models are and what they can add to the current speech perception literature.

2. Evidence accumulation models

Since their advent (e.g., Stone, 1960), evidence accumulation models have been applied to many different fields (see Donkin & Brown, 2016, for a review) – including recognition memory, brightness discrimination, lexical decision, consumer choice, workload capacity, optimal decision-making, implicit association, the effects of alcohol on decision-making, and the neural mechanisms of decision-making (Eidels, Donkin, Brown, & Heathcote, 2010; Evans & Brown, 2016; Forstmann et al., 2008; Hawkins et al., 2013; Ratcliff, 1978; Ratcliff & Rouder, 1998; van Ravenzwaaij, van der Maas, & Wagenmakers, 2011;

[1] We are interested in how much each cue contributes to each phoneme in the contrast, which is not the same thing as investigating how much an acoustic cue contributes to a particular phoneme outside the context of the contrast.