Cognition 166 (2017) 358–370

Contents lists available at ScienceDirect

Cognition

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

Original Articles

A universal bias in adult vowel perception – By ear or by eye

Matthew Masapollo [a,c,*], Linda Polka [a,c], Lucie Ménard [b,c]

[a] School of Communication Sciences and Disorders, McGill University, 2001 McGill College, 8th Floor, Montreal, QC H3A 1G1, Canada
[b] Département de Linguistique, Université du Québec à Montréal, Pavillon J.-A. De sève, DS-4425, 320, Sainte-Catherine Est, Montréal, QC H2X 1L7, Canada
[c] Centre for Research on Brain, Language, and Music, McGill University, 3640 de la Montagne, Montreal, Quebec H3G 2A8, Canada

ARTICLE INFO

Article history:
Received 8 December 2016
Revised 3 May 2017
Accepted 1 June 2017
Available online 8 June 2017

Keywords:
Speech perception
Multisensory
Vowels
Universals
Natural referent vowel framework
Prototypes

ABSTRACT

Speech perceivers are universally biased toward "focal" vowels (i.e., vowels whose adjacent formants are close in frequency, which concentrates acoustic energy into a narrower spectral region). This bias is demonstrated in phonetic discrimination tasks as a directional asymmetry: a change from a relatively less to a relatively more focal vowel results in significantly better performance than a change in the reverse direction. We investigated whether the critical information for this directional effect is limited to the auditory modality, or whether visible articulatory information provided by the speaker's face also plays a role. Unimodal auditory and visual as well as bimodal (auditory-visual) vowel stimuli were created from video recordings of a speaker producing variants of /u/, differing in both their degree of focalization and visible lip rounding (i.e., lip compression and protrusion). In Experiment 1, we confirmed that subjects showed an asymmetry while discriminating the auditory vowel stimuli. We then found, in Experiment 2, a similar asymmetry when subjects lip-read those same vowels. In Experiment 3, we found asymmetries, comparable to those found for unimodal vowels, for bimodal vowels when the audio and visual channels were phonetically-congruent. In contrast, when the audio and visual channels were phonetically-incongruent (as in the "McGurk effect"), this asymmetry was disrupted. These findings collectively suggest that the perceptual processes underlying the "focal" vowel bias are sensitive to articulatory information available across sensory modalities, and raise foundational issues concerning the extent to which vowel perception derives from general-auditory or speech-gesture-specific processes.

1. Introduction

In typical conversational interactions, speech perception is inherently multisensory in nature. Indeed, considerable experimental work has shown that the mapping between the acoustic signal and phonetic structure is complex, in that it changes substantially depending on the presence of co-occurring information specified by other sensory modalities (for reviews, see Fowler, 2004; Munhall & Vatikiotis-Bateson, 2004). For instance, acoustic /ba/ syllables presented simultaneously with visemes for /ga/ are more likely to be perceived as /da/ (the "McGurk effect"; McGurk & MacDonald, 1976), and acoustic /ba/ syllables presented in noise simultaneously with cutaneous air puffs are more likely to be perceived as aspirated /pa/ (Gick & Derrick, 2009). Although it is widely agreed that speech perceivers are "informational omnivores" in this regard (Fowler, 2004), exactly how perceivers consume multimodal information to extract stable phonetic percepts has been debated for decades.

Explanations of multisensory influences on speech perception cover a range of theoretical perspectives. According to one theoretical camp, the complexities of multimodal speech perception are substantially reduced by assuming that perceivers track and extract articulatory information (e.g., Fowler, 1996, 2006). In the present article, we explore the viability of this perspective for understanding a highly robust phenomenon in speech perception, namely, directional asymmetries in vowel discrimination (e.g., Polka & Bohn, 2003). We begin by first presenting a brief overview of research on vowel perception asymmetries, which has focused exclusively on auditory vowel perception. This work uncovered a universal perceptual bias in vowel perception, which recently led to the formulation of the Natural Referent Vowel framework (Polka & Bohn, 2011). Within this context, we then motivate several experiments that address the question of how perceivers consume multimodal speech information by exploring vowel perception biases in visual and auditory-visual speech processing.

It has been known for some time that for many vowel contrasts, discrimination is easier when the same pair of vowels is presented

* Corresponding author at: Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, 190 Thayer St., Metcalf Office 151, Providence, RI 02912, United States.
E-mail address: matthew_masapollo@brown.edu (M. Masapollo).

http://dx.doi.org/10.1016/j.cognition.2017.06.001
0010-0277/

in one direction compared to the reverse direction (for a recent review, see Polka & Bohn, 2011). Such directional effects are ubiquitous in phonetic discrimination tasks employing auditory vowel stimuli with both infants and adults. To take an example of the general phenomenon, Bohn and Polka (2001) tested German-learning infants' discrimination of the German /i/-/e/ vowel contrast using the head-turn conditioning procedure (see, e.g., Werker et al., 1998). In this procedure, infants are presented with a repeating background stimulus and are assessed on their ability to discriminate a change from the background to a different stimulus. The presentation of each vowel was counterbalanced, such that half of the infants tested were presented with one direction of change, i.e., half were presented a change from /i/ to /e/, and half were presented a changed from /e/ to /i/. The results revealed that infants performed better at discriminating the change from /e/ to /i/, compared to the reverse.

Comparable asymmetries in infant vowel perception have been reported for a variety of contrasts from across phonetic space such as /I/-/e/ (Bohn & Polka, 2001), /ɛ/-/æ/ (Polka & Bohn, 1996), /ʊ/-/o/ (Bohn & Polka, 2001), and /Y/-/ʊ/ (Polka & Werker, 1994). Moreover, these directional effects have been shown using several different behavioral paradigms (i.e., habituation, operant conditioning; see, e.g., Polka & Bohn, 2011; Polka & Werker, 1994; Pons, Albareda-Castellot, & Sebastián-Gallés, 2012). What is especially interesting is that these effects cannot be explained by simple acoustic dimensions, such as pitch, amplitude or duration, because these variables have been well-controlled in the test stimuli used across studies. In addition, these asymmetries do not appear to be driven solely by language experience because asymmetries have been shown to emerge in infants' discrimination of both native and non-native (foreign language) vowel contrasts. However, Polka and Bohn (2003) pointed out that these asymmetries in general, with few exceptions, can be predicted by considering the relative position of each vowel within phonetic space (defined by $F1$-$F2$). More precisely, a change from a relatively less peripheral vowel (e.g., /e/) to a relatively more peripheral vowel (e.g., /i/) in phonetic space tends to result in significantly better discrimination performance than a change in the reverse direction. This perceptual pattern is illustrated in Fig. 1a, which shows an $F1$-$F2$ plot of many of the vowel contrasts that have been examined in infant vowel discrimination studies with arrows indicating the direction of change that was easier to discriminate. Thus, the evidence indicates that young infants from across cultures come to the task of language acquisition universally biased toward certain vowels which act as perceptual attractors within phonetic space.

At the same time, there is mounting evidence that linguistic experience builds on and fine-tunes this early vowel bias (Dufour, Brunelliere, & Nguyen, 2013; Polka & Bohn, 2011; Tyler, Best, Faber, & Levitt, 2014). This idea is most clearly demonstrated by recent experimental findings reported by Polka and Bohn (2011). In a recent series of studies, these authors compared Danish-learning infants aged 6 and 12 months and Danish-speaking adults on their ability to discriminate a native (i.e., /e/-/ø/) and a non-native (i.e., British English /ɒ/-/ʌ/) vowel contrast, using the conditioned head-turn procedure. Interestingly, they found an asymmetry in Danish infants' discrimination of both a native (i.e., /e/-/ø/) and a non-native (i.e., British English /æ/-/ɯ/) contrast at 6 months of age. By 12 months of age, however, Danish infants showed evidence of an asymmetry only for the non-native contrast. Similarly, Danish adults showed an asymmetry in their discrimination of the non-native contrast, but not the native contrast. A complementary result was also recently reported by Dufour et al. (2013), who observed an asymmetry in Southern French adults' discrimination of a non-native (Standard French /o/-/u/) contrast, but not a native (Standard French /ɔ-u/) contrast. Although the existing data suggests that asymmetries are more limited for native contrasts in adult perceivers, there is some evidence that similar asymmetries do occur for some native contrasts (e.g., /i/-/I/; Cowan & Morse, 1986; Repp & Crowder, 1990; Repp, Healy, & Crowder, 1979).

Polka and Bohn recently offered a model of early phonetic development termed the Natural Referent Vowel (NRV) Framework, which focuses on explicating the nature and development of human vowel perception (Polka & Bohn, 2011). NRV proposes that the speech perception system is initially biased toward vowels with extreme articulatory/acoustic properties and, moreover, that this bias is often experimentally demonstrated as a directional asymmetry. As infants accrue experience listening to the native language, this bias (which we will henceforth refer to as the NRV bias) will then dynamically adjust to increase perceptual sensitivity near the boundaries of native vowel categories. This distortion of perception is predicted to produce a reduction in asymmetries for native vowel contrasts and maintain (or even possibly enhance) asymmetries for non-native contrasts.

While the existence of the NRV bias is by now well-established, the processes that underlie it remain poorly understood. Moreover, the precise nature of the information that such processes operate on has not yet been fully explicated. At this stage in theory development, Polka and Bohn have proposed that asymmetries derive from a universal perceptual attunement to "focal" vowels; that is, vowels whose adjacent formants are close in frequency (Polka & Bohn, 2011; see also, Schwartz, Abry, Boë, Ménard, & Vallée, 2005). Below we unpack this hypothesis, for which there is increasing consensus.

Articulatory and acoustic investigations of vowels have shown that when spectrally adjacent formants move close together in frequency there is a mutual reinforcement of their acoustic energy, such that the amplitude of each formant increases. As a result, when formants converge, acoustic energy becomes concentrated into a narrower spectral region (see Kent & Read, 2002; Stevens, 1989, for a discussion). These convergence points have been referred to as "focal" points in vowel spectra (Schwartz, Boë, Vallée, & Abry, 1997; Schwartz et al., 2005). Importantly, maximal formant convergence, or focalization, is observed for vowels found at the periphery of the vowel space, which also have the most extreme vocal tract postures. For example, F2 and F3 and F4 converge (i.e., are close in frequency) for /i/ (which is the highest front vowel), and F1 and F2 are spectrally close to each other for /a/ (which is the lowest central vowel) as well as /u/ (which is the highest back vowel). Thus, the convergences of formants serve as acoustic information for extreme vocalic articulations.

Indeed, there is experimental evidence that vowels with well-defined spectral prominences caused by formant convergence exhibit increased spectral salience (e.g., Schwartz & Escudier, 1989). In a critical study, Masapollo and colleagues conducted direct tests that examined whether adults from different language backgrounds would show directional asymmetries consistent with NRV predictions while discriminating a series of synthetic /u/ vowels varying systematically in their acoustic peripherality and degree of formant proximity (between F1 and F2, in equal psychophysical steps; Masapollo, Polka, Molnar, & Ménard, 2017). In a preliminary experiment designed to ensure that the stimuli fell within the /u/ category, subjects completed a phonetic identification and goodness-rating task in which they were presented with the vowel stimuli and asked to identify and then rate each stimulus according to how good an exemplar of the /u/ category it was (see, e.g., Iverson & Kuhl, 1995; Miller & Volaitis, 1989). The results were clear in showing that although many of the stimuli were consistently identified as /u/ by both language groups, the best French /u/ exemplar range occupied a more peripheral region of phonetic space compared to the best English /u/ exemplar range. Fig. 1b shows the precise locations of the best exemplar ranges for both