Normalization of the Natural Referent Vowels

D. H. Whalen
City University of New York; Haskins Laboratories; Yale University

Abstract
The Natural Referent Vowel framework makes strong, testable predictions that have already provided fruitful directions for new research. Largely undiscussed, however, is the role that vocal tract normalization plays in the perception of vowels by infants. Two issues arise: First, when presented with a single vowel, how does the infant know whether it is truly a referent vowel or not? Second, if, unlike in all previous studies, vowels attributable to different vocal tracts are perceived, does the infant normalize or not? The first might be answerable with neural imaging. The second can be tested behaviorally, though the design is difficult both mechanically and theoretically.

1. Introduction

The Natural Referent Vowel framework (Polka & Bohn, 2011) treats the articulatorily and acoustically extreme vowels as natural reference points that are especially useful to infants learning language. These NRVs are /i ɑ u/. Being on the edge of the vowel space, they can give an infant anchor points for developing a vowel space of their own. Infants can more easily tell that a vowel has changed when the change is toward the periphery (i.e., toward the NRVs) than in the opposite direction. A variety of experimental results are consistent with NRVs being treated differently from other vowels, as summarized in the 2011 paper.

Anne Mette Nyvad, Michaela Hejná, Anders Højen, Anna Bothe Jespersen & Mette Hjortshøj Sørensen (Eds.), ***A Sound Approach to Language Matters – In Honor of Ocke-Schwen Bohn*** (pp. 81-89). Dept. of English, School of Communication & Culture, Aarhus University.

A previous framework, the Native Language Magnet (NLM) effect (e.g., Kuhl & Iverson, 1995), makes some compatible predictions while leaving open the issue of the universality of the NRVs. NLM predicts that vowels that occur in the ambient language will attract nearby vowel tokens into their perceptual category, while vowel categories from other languages will not. Given that most languages have the NRVs in their inventories, this will lead to compatible predictions between the two accounts in many cases, though Polka and Bohn have results that do not depend on the NRV's existence in the language (e.g., Polka & Bohn, 1996). (It would be interesting to see what happens with languages, such as most of the Algonquian languages, which lack /u/.) There is some evidence that the NRVs have a greater perceptual effect than other native vowels (Polka & Bohn, 2011: 476), with /i/ eliciting more reaction (sucking) than /y/ even for infants in a Swedish environment where both vowels are native. As with most issues concerning acquisition, there is much more work to be done.

Vowel identification is not straightforward for listeners, however. Different vocal tract lengths produce different formant patterns for the same vowel. This is clear both on theoretical grounds (Fant, 1960) and in measurements of men, women and children (Peterson & Barney, 1952). Human listeners compensate for such effects, and they seem to do so both with signal-internal ("intrinsic") and ancillary ("extrinsic") information (Ainsworth, 1975). Intrinsic information is entirely within a single vowel. Extrinsic information relates the token to a speaker's vowel space, or at least the immediate environment. That infants are capable of such normalization is indicated by their success at imitating adult productions with their tiny little vocal tracts, even though their formant values were necessarily different from the adult models (Kuhl & Meltzoff, 1996). Indeed, imitation based on reinterpretation of the input signal, including sensitivity to its visual aspects, into something the infant can produce is a prerequisite for speech acquisition (Studdert-Kennedy, 1986).

Many acoustic normalization procedures have been proposed (for a review, see, e.g., Flynn, 2011). To date, they all perform more poorly than human listeners. Humans, of course, have an advantage in having a couple of million years of evolution helping them out, but the algorithms are also hampered by limitations on the input given to them. Typically, the input includes fundamental frequency (F0) and formant values for the midpoint of a vowel, augmented in some cases by duration information. We have known for decades that this is not the information that human listeners depend most on (e.g., Strange et al., 1976), but the levels of performance

obtained are sufficient that the approach continues to be used. However, our formant measurements are not terribly accurate (Klatt, 1986; Shadle et al., 2016), leading to an initial degradation of performance by the normalization algorithms. For one sizable dataset, a model that included F0 and formant measurements at one time point still performed well below human perceptual levels, while using three time points along with duration led to fairly equivalent performances to those of humans (Hillenbrand et al., 1995). It would seem that infants have their work cut out for them.

The research discussed in the previous paragraph included all the vowels of English, but the NRVs are not always the best identified ones. In the Hillenbrand et al. (1995, p. 3108) study, the vowel /i/ was identified correctly by human listeners the most often (99.6%). The vowel /u/ was also highly identifiable (97.2%), but not as much as /o/ (99.2%). The vowel /ɑ/ was noticeably less accurately identified (92.3%). If we ignore the vowel /ɔ/, which is not distinctive in all American English dialects, the next worst rate of identification was 90.8% for /ʌ/. These are, of course, adult perceptions of an established inventory, and they are identification scores, which may be less revealing than discrimination scores. However, they do not immediately indicate that NRVs have a special status.

This paper will explore two predictions that can be drawn from the NRV position. First, experimental studies that present individual tokens of vowels to infants would seem to require that they recognize each token as being an example of an NRV or not. Is this possible? How can we tell? Second, does the need to normalize for more than one vocal tract reduce the effectiveness of the NRVs? Adult listeners show reduced accuracy with multiple speakers (e.g., Assmann et al., 1982); do infants also have trouble adjusting their categories? Possible ways of addressing those questions will be presented.

2. Normalization of a single vowel

Each time a stimulus is presented to a listener (in the current case, an infant), some phonetic information is extracted. Just what kind of information that is remains underspecified. Is it a true representation of the resonances of the vocal tract that produced it? Is it classified into the NRV category it belongs to? Or is it placed into some vaguely specified acoustic space that is only tangentially related to vowel categories? The ultimate perceptual treatment of ambiguous vowels (e.g., Kuhl et al., 1992) is then somewhat unclear: Are these vowels also normalized on intrinsic grounds, or are they (extrinsically) put into the context of the current speaker?

The nature of the low NRV is itself rather ambiguous. There is a great deal of variation in the low vowel used by any specific language, even though some form of low vowel is, perhaps, universal. Should we expect that the NRV status of [ɑ] (or [ɒ] or [ɐ] or [a]) can be determined on a language-specific basis? Or does that violate the principles of the NRV proposal? Certainly, the lowest vowel that an infant hears from a particular speaker might serve as a reference point, but this implies that extrinsic normalization is at work (which would seem to reduce the "referent" component of NRV) and that the infant can associate utterances consistently with a single speaker. It does seem that infants can recognize new voices, at least those that are speaking the ambient language (Johnson et al., 2011), but such recognition might, of course, depend on source characteristics rather than filter characteristics. Would an infant be startled to hear a newly-familiar voice produce a vowel that seemed outside its range? Or would that signal a new speaker? Our current experimental techniques may be inadequate for answering the question directly, but considerations of how we might approach the question allow further insight into the nature of the NRVs.

The statistical distributions of vowel formants might also influence infants' perception in these experiments, but many of the same issues arise. Are the instances of a formant pattern mapped onto an individual speakers' vowel space? Can the infant keep multiple maps and update them appropriately? Indeed, how do they know to put them into a vowel space at all, rather than just interesting formant patterns? Their babbling (at one year of age) does reflect the ambient language in the trends of formant values (Boysson-Bardies et al., 1989), an effect those authors attribute to the onset of category formation. Statistical explanations were proposed to avoid nativist explanations of acquisition, by allowing the language patterns themselves to provide the information needed for acquisition. Infants have been shown to be sensitive to a number of statistical properties in speech experiments (Kuhl et al., 1992; Maye et al., 2002; Saffran et al., 1996), although it is less clear that short-term sensitivity predicts long-term retention (Gómez, 2017). Just where and how those statistics are stored is also unspecified. Vowel formants would seem to have to be normalized into some universal space or stored along with speaker identity. Further, if there are no categories, what are the statistics applying to? The earliest stages of acquisition would seem to resist an entirely input-based approach. Beyond that, the way in which speaker-specific storage would then affect the infant's own production is not obvious.

The nativist positions that were challenged by statistical approaches were largely reacting to theories that assumed some kind of segment as innate, but non-segmental nativist proposals may also be viable. The Articulatory Organ Hypothesis (the AOH; Studdert-Kennedy & Goldstein, 2003) assumes that certain broad gestures (tongue tip, lips, etc.), "organs," are available to infant perceivers, so that distinctions across organs are more easily perceived than those within. The evidence for the AOH has been primarily examined for consonants, and the results have been largely positive with many problematic cases (Best et al., 2016). For vowels, constrictions of the pharynx, tongue root and tongue tip could provide organs for /ɑ/, /u/ and /i/ respectively. This is, of course, entirely consistent with the NRVs, but seen primarily from the articulatory viewpoint. The essential breadth of the organs in the AOH also allows for the wide variety of low vowels mentioned above to be included in one organ category without removing /ɑ/ (and its neighbors) from primary status. Any success the infant has in recognizing organs in adult speech is as dependent on normalization as in other frameworks, but it may be that this kind of global gesture is more easily computed from the acoustic signal than previously thought, once a fuller (and more realistic) depiction of the acoustic signal is used (Iskarous, 2010).

Whichever framework ultimately provides the best explanation for speech acquisition, it is clear that a great deal of work remains before we will have a satisfactory understanding. If we find a neuroimaging technique that allows us to see a distinct signature for a specific vowel category, perhaps we will be able to see when and where the normalization takes place, how successful it is, and what happens to individual tokens in an experiment. To date, we do not have such signatures, but they may yet be discoverable. If so, exploring the development of categorization will become even more fascinating.

3. Multiple vocal tracts in one experiment

The procedure for testing infant perception of vowels typically involves a single speaker (e.g., Polka & Bohn, 1996) or one speaker per language (e.g., Polka & Werker, 1994). Although this makes the experimental design manageable – infants do not tolerate a huge number of stimuli – it does ensure that the vowels will be maximally informative about a single vocal tract. The imitation studies cited above suggest that infants do, in fact,

perform some kind of normalization, so that they can relate what they hear to what they would produce. How well does this normalization proceed on a token to token basis?

Two possibilities seem likely. The first is that infants normalize primarily on intrinsic grounds, and thus they should be able to identify tokens from different speakers as belonging to the same category. The other, completely opposite possibility is that infants rely greatly on extrinsic grounds (sampling of the total vowel space, for example) and would fail to identify any vowels, including the NRVs, when speakers vary. There are probably other intermediate possibilities, for example, that there is something strongly coherent in the NRV acoustic pattern that is treated as a "magnet" on psychophysical grounds, such as a close proximity of two intense formants (F1 and F2 for [u] and [ɑ], F2 and F3 for [i]). Such an account would be consistent (I think) with the intrinsic normalization variant. In any event, these two starkly contrasting ones suggest a direct test.

We could test these two different predictions by seeing whether having multiple speakers eliminates or reduces the attraction that NRVs have in infant perception. (Multiple talkers were used in Bundgaard-Nielsen et al. (2015), showing that the technique is possible.) In the extreme case, every token presented to an infant could come from a different talker. But even having 10 or 20 talkers would presumably be sufficient to test whether speaker consistency in the input is strictly necessary. In a fantasy world, where infants scheduled themselves in large numbers, we could then imagine doing each of those talkers separately to ensure that the different voices were equally effective. Even in our present world, we could conceivably test 4, or at the very least 2 of the speakers to see if the NRVs had a larger effect with a single speaker than they did with multiple speakers. However, the main issue would not require such an extension: If NRVs are effective with many talkers, it would seem that intrinsic normalization was operative. If they were not effective, then it could be that extrinsic normalization is necessary, but it could also be that infants are not happy with a situation with too many adults and/or a lack of social connection with one or two adults. No doubt there are other possible outcomes and explanations, but in our current state, we don't know how NRVs are normalized.

4. Summary

The NRV proposal is one that makes an admirable number of predictions possible. Because so much of what happens in speech perception, especially at the very beginning of a speaker/listener's life, is currently unknown,

these predictions must be rather broad. Further, the extensive work that is involved in any study of infant perception guarantees that progress will be slow. Ocke Bohn, and his many collaborators, are to be commended for persevering with this and other questions fundamental to our understanding of speech. While he may not reach the ultimate answers before hanging up his headphones, we can hope that Ocke will continue to explore this endlessly fascinating topic.

5. References

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgements. In G. Fant, & M. A. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 103-113). San Francisco: Academic Press.

Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *Journal of the Acoustical Society of America*, *71*, 975-989.

Best, C. T., Goldstein, L. M., Nam, H., & Tyler, M. D. (2016). Articulating what infants attune to in native speech. *Ecological Psychology*, *28*, 216-261.

Boysson-Bardies, B. de, Hallé, P. A., Sagart, L., & Durand, C. (1989). A crosslinguistic investigation of vowel formants in babbling. *Journal of Child Language*, *16*, 1-17.

Bundgaard-Nielsen, R. L., Baker, B. J., Kroos, C. H., Harvey, M., & Best, C. T. (2015). Discrimination of multiple coronal stop contrasts in Wubuy (Australia): A Natural Referent Consonant account. *PLoS ONE*, *10*(12), e0142054.

Fant, G. (1960). Acoustic theory of speech production. The Hague: Mouton.

Flynn, N. (2011). Comparing vowel formant normalisation procedures. *York Papers in Linguistics Series*, *2*, 1-28.

Gómez, R. L. (2017). Do infants retain the statistics of a statistical learning experience? Insights from a developmental cognitive neuroscience perspective. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372.

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

Iskarous, K. (2010). Vowel constrictions are recoverable from formants. *Journal of Phonetics*, *38*, 375-387.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, *14*, 1002-1011.

Klatt, D. H. (1986). Representation of the first formant in speech recognition and in models of the auditory periphery. In P. Mermelstein (Ed.) *Proceedings of the Montreal satellite symposium on speech recognition, 12th international congress on acoustics* (pp. 5-7). Montreal: Canadian Acoustical Society.

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the "perceptual magnet effect". In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 121-154). Timonium, MD: York Press.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America*, *100*, 2425-2438.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. E. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606-608.

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

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*, 175-184.

Polka, L., & Bohn, O.-S. (2011). Natural Referent Vowel (NRV) framework: An emerging view of early phonetic development. *Journal of Phonetics*, *39*, 467-478.

Polka, L., & Bohn, O.-S. (1996). A cross-language comparison of vowel perception in English-learning and German-learning infants. *Journal of the Acoustical Society of America*, *100*, 577-592.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 421-435.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Shadle, C. H., Nam, H., & Whalen, D. H. (2016). Comparing measurement errors for formants in synthetic and natural vowels. *Journal of the Acoustical Society of America*, *139*, 713-727.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, *60*, 213-224.

Studdert-Kennedy, M. (1986). In B. Lindblom, & R. Zetterstrom (Eds.), *Development of the speech perceptuomotor system. Precursors of early speech* (pp. 205-217). New York: M. Stockton Press.

Studdert-Kennedy, M., & Goldstein, L. M. (2003). In M. Christiansen, & S. Kirby (Eds.), *Launching language: The gestural origin of discrete infinity. Language evolution* (pp. 235-254). Oxford: Oxford University Press.