Phonological abstraction without phonemes in speech perception

Holger Mitterer[1], Odette Scharenborg[1,2] & James M. McQueen[1,3]

[1]Max Planck Institute for Psycholinguistics, Nijmegen
[2]Radboud University Nijmegen, Donders Institute for Brain, Cognition, and Behaviour, Nijmegen
[3]Radboud University Nijmegen, Behavioural Science Institute and Donders Institute for Brain, Cognition, and Behaviour, Centre for Cognition, Nijmegen

Recent evidence shows that listeners use abstract prelexical units in speech perception. Using the phenomenon of lexical retuning in speech processing, we ask whether those units are necessarily phonemic. Dutch listeners were exposed to a Dutch speaker producing ambiguous phones between the Dutch syllable-final allophones approximant [r] and dark [l]. These ambiguous phones replaced either final /r/ or final /l/ in words in a lexical-decision task. This differential exposure affected perception of ambiguous stimuli on the same allophone continuum in a subsequent phonetic-categorization test: Listeners exposed to ambiguous phones in /r/-final words were more likely to perceive test stimuli as /r/ than listeners with exposure in /l/-final words. This effect was not found for test stimuli on continua using other allophones of /r/ and /l/. These results confirm that listeners use phonological abstraction in speech perception. They also show that context-sensitive allophones can play a role in this process, and hence that context-insensitive phonemes are not necessary. We suggest there may be no one unit of perception.

How do listeners bridge the divide between acoustic input and lexical meaning? One view is that there are intermediate “units of perception”—most commonly phonemes—that link acoustics to a phonologically abstract lexicon; another view is that the signal is mapped directly onto a lexicon comprising acoustically-detailed representations. Evidence from a perceptual-learning paradigm supports the former view (McQueen, Cutler, & Norris, 2006; Mitterer, Chen, & Zhou, 2011; Sjerps & McQueen, 2010) but more than 40 years of research has failed to reveal what the units of perception are. We argue here that the combination of a new stimulus-construction method with this perceptual-learning paradigm provides a new way to approach this issue. In a first application of this research strategy, we show that prelexical phonological abstraction does not require phonemes and suggest that there may be no universal, context-insensitive unit of speech perception.

During the early years of research on speech perception, it was generally agreed that there was some form of intermediate unit. Research from the seventies to the early nineties of the last century saw efforts trying to delineate the grain size of this basic unit (reviewed in Goldinger & Azuma, 2003). Paradigms used included monitoring tasks (Savin & Bever, 1970) and selective adaptation (reviewed by Remez, 1987). In each case, it turned out that the results were not decisive in arguing for or against any particular basic unit, usually because other assumptions in the chain of inference were disputed

Correspondence concerning this article should be sent to: Holger Mitterer, University of Malta, Department of Cognitive Science, Msida, MSD 2080, Malta or by eMail to holger.mitterer@um.edu.mt. The research of Odette Scharenborg is partly sponsored by the Max Planck International Research Network on Aging. Odette Scharenborg is now at the Centre for Language Studies, Radboud University Nijmegen, The Netherlands.

(see, e.g., Marslen-Wilson & Warren, 1994; McQueen, Norris, & Cutler, 1999, with respect to subcategorical mismatches).

Episodic models of speech perception (Goldinger, 1998; Hawkins, 2003; Port, 2010) questioned the existence of prelexical units altogether. However, findings that listeners are able to generalize perceptual learning about an unusual pronunciation of a speech sound to other lexical items containing that sound (McQueen et al., 2006; Mitterer et al., 2011; Sjerps & McQueen, 2010) have shown that at least some form of prelexical unit is required. In these experiments, participants heard an ambiguous sound (e.g., between /f/ and /s/, [$^{s}/_{f}$]) in /s/-final words (e.g., "platypu[$^{s}/_{f}$]", where *platypus* is an English word and *platypuf* is not). Simulations with a strictly episodic model predict that this experience should have little effect on the perception of other words (Cutler, Eisner, McQueen, & Norris, 2010). But findings show that listeners exposed to "platypu[$^{s}/_{f}$]" retune their perception of /s/ in other words (e.g., they interpret "nai[$^{s}/_{f}$]" as "nice" rather than "knife"). The learning must be prelexical and entail phonological abstraction (i.e., that there are units representing the critical sound) for it to be applied to new words.

The perceptual-learning paradigm has thus brought us back to the old question about the nature of these abstract prelexical representations. As we argue here, this paradigm also offers a new way to approach this old question because it can reveal which units play a functional role in speech perception. Under the assumption that the units which mediate learning are the same as those which mediate speech perception, we can use this paradigm to ask whether these units are necessarily context-insensitive phonemes or may sometimes be context-sensitive allophones, that is, whether the units can vary.

A popular working assumption, both in models of human and automatic speech recognition (Scharenborg, Norris, ten Bosch, & McQueen, 2005), is that prelexical units are similar to the linguistic concept of the phoneme, that is, they are representations of individual sounds which are independent of context and position. Three studies have directly evaluated whether learning about speech segments generalizes over context and/or positions, as the phonemic hypothesis predicts. First, Jesse and McQueen (2011) showed that learning about a syllable-final fricative generalizes to the perception of syllable-initial fricatives. While this may speak for context- and position-independence, Jesse and McQueen pointed out that their data are compatible with allophonic units: the fricative contrast that was tested (/s/-/f/) is relatively context-insensitive, and the same physical fricatives were indeed used across positions.

Kraljic and Samuel (2006) found generalization of learning about voicing in stops (distinguishing [b,d,g] from [p,t,k]) from one contrast (e.g., /b/-/p/) to another (e.g., /d/-/t/). This finding would be in line with the assumption that prelexical units are context-invariant features (Lahiri & Reetz, 2002). But because the acoustic cue to voicing is very similar across places of articulation, this finding is compatible with any account in which perceptual learning is tightly bound to the acoustic patterns in the input, and thus with a variety of representational options.

Third, Dahan and Mead (2010) found context- and position-specific effects in adaptation to noise-vocoded speech. This could be interpreted as evidence that prelexical units are allophonic. But Dahan and Mead admit that the effects could have been driven in large part by a bias to report words at test that resembled the structures heard during training, since the participants struggled to understand the noise-vocoded speech (only about half of the words were identified correctly). Moreover, their stimulus set was a mixed bag, including segments that change drastically between onset and offset position (such as stops in American English) and segments that change hardly at all (e.g., nasals or voiceless fricatives).

These three studies indicate that what is required to distinguish between different formats of prelexical representation is a contrast with considerable and clear-cut allophonic variability. The /r/-/l/ contrast in Dutch is such a contrast. For /l/, allophony is purely positional. In onset position, an alveolar lateral approximant [l] is used; the velarized counterpart [ɫ] ("dark l") is used in offset position. For /r/, the realization depends on its position and the speaker. In onset position, some speakers prefer the alveolar trill [r] and some the uvular trill [ʀ] (Van Bezooijen, 2005). In offset position only, there is the additional option of using the alveolar approximant [ɹ].[1]

We therefore examined lexically-guided retuning of the Dutch /r/-/l/ contrast in syllable-final position and tested generalization over different allophonic implementations and different positions. To deal with the fact that /l/ and /r/ are generally strongly coarticulated with the surrounding vowels (Ladefoged & Maddieson, 1996), we used syllable-based audio morphing to generate ambiguous syllables (Kawahara, Masuda-Katsuse, & de Cheveigné, 1999). This innovation allowed us to expose listeners to ambiguous examples of /r/- and /l/-final words. Pretests established which steps of the continuum were ambiguous (the morphing method does not necessarily create the most ambiguous stimuli in the middle of the continuum). In the main experiment, one group of listeners heard good examples of /l/-final words (e.g., [ɑksɛptabəɫ] *acceptabel*, "acceptable") and /r/-final words with an ambiguous last syllable (e.g., [wɪntə^{ɹ}/$_{ɫ}$], *winter*, "winter"). Another group got the opposite exposure, with good examples of /r/-final words ([wɪntəɹ]) and ambiguous examples of /l/-final words ([ɑksɛptabə^{ɹ}/$_{ɫ}$]). Both groups were then tested on their perception of ambiguous sounds on three types of nonword continua. One continuum used the same allophones as were used in exposure (approximant [ɹ] and dark [ɫ]; full-match continuum), one continuum used the same allophone for the /l/-endpoint but a trill [r] for the /r/-endpoint (partial-match continuum), and the third used two allophones (light [l] and trill [r]; no-match continuum) that were different from those used in exposure.

[1] Allophony here is language specific, the UPSID database (Maddieson, 1984) lists six languages using [ɹ] and [r] as separate phonemes.

If prelexical processing necessarily abstracts to phonemic units, exposure should affect the perception of all continua, with more /l/ responses after exposure to ambiguous segments in /l/ positions and more /r/ responses after exposure to ambiguous segments in /r/ positions. If a partial acoustic match is sufficient to trigger application of learning, some learning should be observed when one allophone is shared between exposure and test. If, however, prelexical processing can abstract to allophonic units, exposure should affect only the perception of the continuum consisting of exposure allophones.

Method

Participants

86 native speakers of Dutch from the participant pool of the Max Planck Institute for Psycholinguistics were paid to take part. There were 16 in the exposure-material pretest, 35 in the test-material pretest and 35 in the main experiment.

Materials and Procedure

Exposure Phase. We selected 200 Dutch words and created 200 nonwords. Forty words ended in /l/ and 40 ended in /r/; there were no /l/'s or /r/'s elsewhere in these words or in the 120 filler words and 200 filler nonwords. Since the sounds [ɫ] and [ɹ] color the pronunciation of the preceding vowel, all critical words ended in /əl/ or /ər/. Word frequency and number of syllables were matched for the /l/- and /r/-final words. There were 11 different pairs of final syllables ending with [Cəɹ/ɫ], where 'C' is one of 11 syllable-initial consonants, including /t/ (the consonant used in the test continuum, see below). All 400 stimuli were recorded by a female native speaker of Dutch. Tokens of the critical syllables were excised and segmented manually into onset, nucleus, and coda portions. This segmentation was used in a time-aligned version of the morphing algorithm in STRAIGHT (Kawahara et al., 1999). Using this algorithm, 11 versions of these syllables were generated with mixture levels ranging from 0% [Cəɹ] and 100% [Cəɫ] to 100% [Cəɹ] and 0% [Cəɫ] in steps of 10%. These 11 sets of syllables were each combined with a pair of word onsets, making word-nonword and nonword-word continua. In an exposure-material pretest, 16 participants labeled the last segment of these stimuli as /l/ or /r/. Based on the results, an ambiguous step of each syllable was selected and used to make the ambiguous exposure items for the main experiment. Two selection criteria were used: overall ambiguous categorization (about 50% /l/- and /r/-responses overall) and a strong lexical bias within each pair of continua.

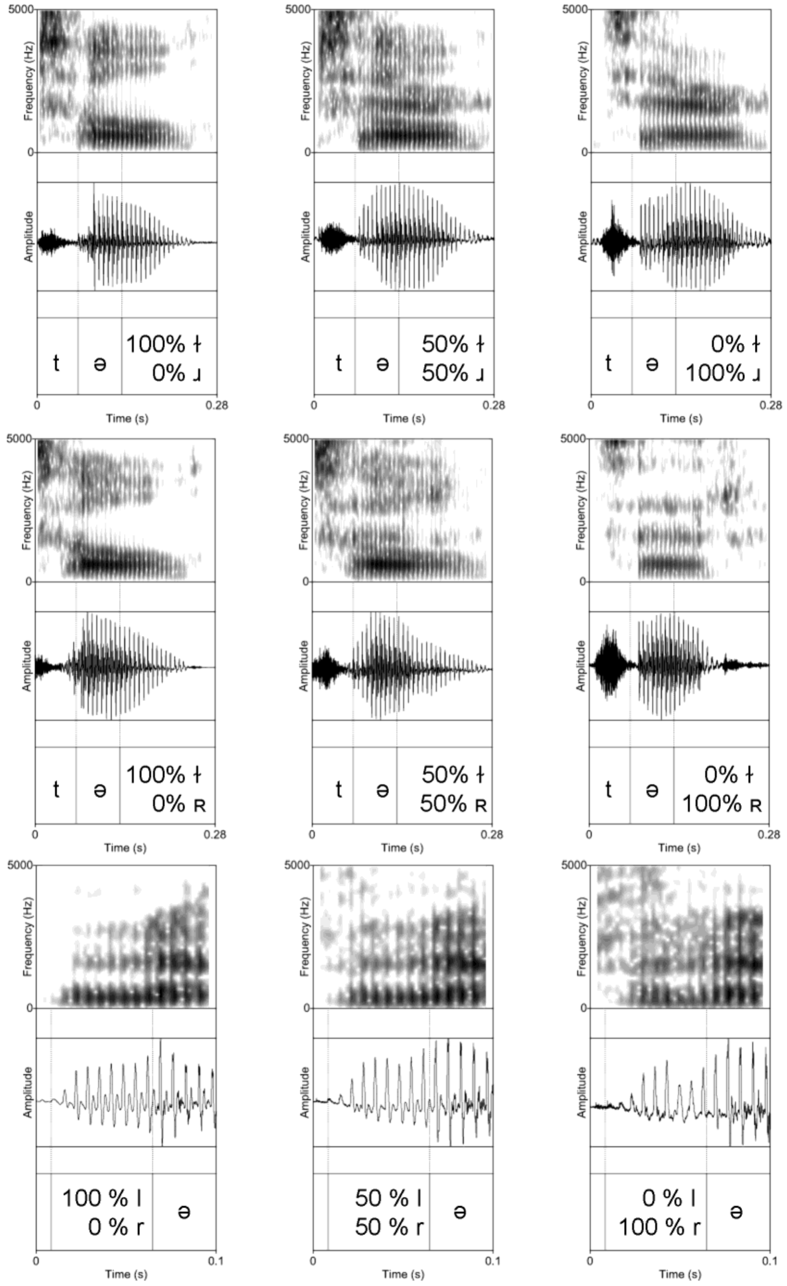

Figure 1. Example spectrograms and oscillograms from the three test continua. The three panels in the top row show three steps from the full-match continuum, the middle row shows the partial-match continuum, and the bottom row shows the no-match continuum.

Test Phase. The Dutch nonwords *kwipter*, *kwiptel*, *repaas*, and *lepaas* were recorded by the same speaker. The /r/-final nonword was recorded in two versions: with an alveolar trill [r] and with an approximant [ɹ]. The /r/-initial nonword was recorded only with an alveolar trill. Three morphed continua were created (see examples in Figure 1): 1. dark /l/ to approximant /r/, [kwɪptəɫ] - [kwɪptəɹ] (same allophones as in exposure; the full-match continuum), 2. dark /l/ to trill /r/, [kwɪptəɫ] - [kwɪptər] (only [ɫ] allophone same as in exposure; the partial-match continuum), and 3. light /l/ to trill

/r/ in onset position, [ləpas]- [rəpas] (different allophones from exposure; the no-match continuum). For the full-match continuum, the exposure-material pretest indicated that step 5 (a 50%-50% mixture of the /r/ and /l/ sounds) was the most ambiguous. The test-material pretest asked whether this was also the case for the other two continua. Thirty-five participants therefore categorized steps 2, 4, 5, 6, and 8 of all three continua. Figure 2 shows the results. The resulting phoneme identification function on the dark-/l/-to-approximant-/r/, full-match continuum was reasonably steep and centered around 50%, as was to be expected on the basis of the exposure-material pretest. For the dark-/l/-to-trill-/r/, partial-match continuum, there were very few /l/ responses overall. To counteract this bias, the more /l/-like steps 5, 6, 7, 8, and 10 were selected for the main experiment. For the light-/l/-to-trill-/r/, no-match continuum, the resulting identification function was shallow around 50%. An extended version of the continuum (steps 0, 2, 5, 8, and 10) was therefore selected.

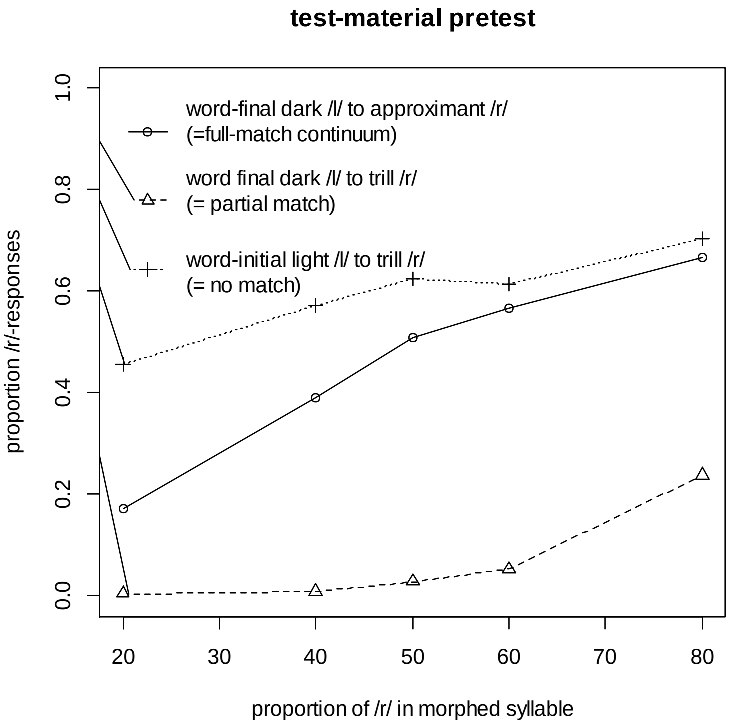

Figure 2. Results of the test-material pretest using stimuli with the mixing proportions from 80% /l/ + 20% /r/ to 20% /l/ + 80% /r/. Based on the results, the same steps of the exposure-material pretest stimuli were used in the test phase, more /r/-like stimuli were used for the partial-match continuum and a more extended continuum was used for the no-match continuum.

In the main experiment, there was first a lexical-decision task with a between-subject manipulation. Half of the participants heard the 40 /l/-final words with unambiguous /l/ and the 40 /r/-words with the ambiguous syllables selected in the exposure-material pretest. The other participants heard the same ambiguous syllables in the /l/-final words and the /r/-final words with unambiguous /r/. Both groups heard all 320 fillers. Experimental and filler stimuli were presented in random order. During the subsequent test phase, all participants heard three blocks containing each of the 15 test stimuli (three continua with five levels) twice, presented in a newly-randomized order in each block. They categorized them as ending (or starting, depending on the continuum) with /l/ or /r/.

Results

Due to a technical error, the lexical-decision data of one participant were lost. Five participants accepted less than 50% of the ambiguous items as words (filler accuracy: 97.5%) and were excluded from further analysis (cf. Norris, McQueen, & Cutler, 2003). The remaining participants accepted 88% and 95% of the ambiguous /r/- and /l/-final items as words, respectively. As Figure 3 shows, at test, the exposure condition influenced perception of the dark-/l/-to-approximant-/r/, full-match continuum (i.e., the trained allophones, see Panel A) but neither of the other continua (Panels B and C). Listeners who were exposed to the ambiguous sound in the /r/-final words labeled the full-match continuum more often as /r/ than listeners who were exposed to the same ambiguous sounds in /l/-final words, but no exposure effects were observed on the other continua.

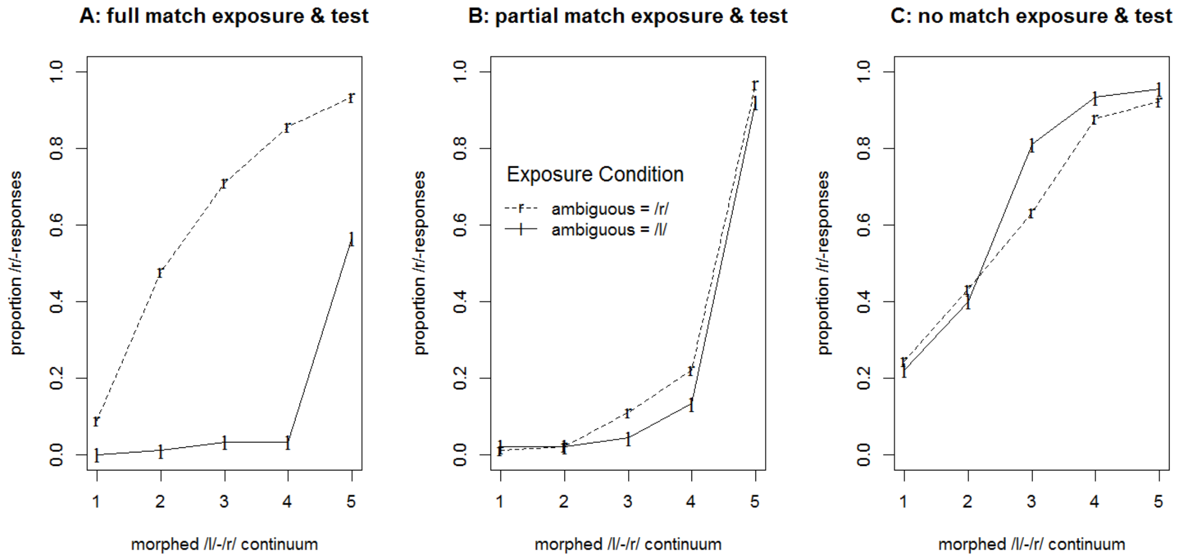

Figure 3. Labeling functions of the two exposure groups for the three test continua. A: data for the continuum of allophones matching those heard during exposure (full-match continuum). B: data for the continuum with one matching and one mismatching allophone (partial-match continuum). C: data for the continuum with allophones fully mismatching those heard during exposure (no-match continuum). An exposure effect (the group exposed to ambiguous sounds in /r/-final words giving more /r/-responses) is shown only in Panel A.

An ANOVA on the logOdds of the /l/-responses with the predictors Exposure (between participant), and Continuum and Step (within participant) revealed a three-way interaction (F(8, 224) = 7.19, $p < 0.001$). Follow-up analyses for each continuum revealed a significant effect of Step only (partial-match: F(4, 112) = 233.21, $p < 0.001$, no-match: F(4, 112) = 65.70, $p < 0.001$) for the two generalization continua; neither the main effect of Exposure nor its interaction with Step was significant (partial-match: F(1, 28) = 1.01 and F(4, 112) = 0.70; no-match: F(1, 28) = 0.40 and F(4, 112) = 1.33, all *p*s > 0.2). But there was a main effect of Exposure and an interaction with Step for the continuum using the exposure allophones (full-match; F(1, 28) = 95.63 and F(4, 112) = 14.91, *p*s < 0.001). There were significant effects of Exposure in separate analyses for each Step of this continuum (F_{min}(1, 28) = 5.09, p < 0.05 at Step 1). The interaction was hence caused by different effect sizes over the continuum, with, as often observed, contextual influences largest in the middle. Lexically-guided retuning thus only occurred in the full-match condition.

Discussion

The lexically-guided retuning paradigm can delineate which units listeners use in prelexical processing because it reveals what kind of abstract phonological representations play a functional role in mapping the acoustic input onto the mental lexicon. We tested whether perceptual retuning for one allophonic implementation of the /r/-/l/ phonemic contrast in Dutch has repercussions for other allophonic implementations of the same contrast. This was not the case. Note that this failure to generalize to other allophones cannot be due to a perceived speaker switch depending on the allophone of /r/ (Eisner & McQueen, 2005): A speaker that uses the approximant in offset position does not necessarily use this variant consistently in this position (see Figure 1 in Van Bezooijen, 2005). Moreover, the fact that the test stimuli sampled from different parts of the continua across conditions (resulting in the different categorization function shapes in Figure 3) is not a plausible reason for the failure of generalization, as the effect of exposure was significant for all steps of the exposure-allophone, full-match continuum and thus not limited to the most ambiguous steps. We hence conclude that context-insensitive phonemes cannot be the only units listeners use.

The current data extend the range of contrasts for which lexically-guided retuning has been shown to an approximant/lateral contrast with distributed perceptual cues. They confirm, for this new type of contrast, that listeners' generalizations of perceptual learning are tightly bound to the acoustic patterns they experience (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2006). But this study goes beyond earlier demonstrations by specifying the nature of the prelexical units involved. We present evidence that context-specific allophones are units used in prelexical processing, and thus that listeners can perceive spoken words without using phonemic representations.

Existing theories of spoken-word recognition need to be modified to account for these findings. In TRACE (McClelland & Elman, 1986) and Shortlist (Norris & McQueen, 2008) the assumption is that the units of speech perception are context-independent phonemes. In FUL (Lahiri & Reetz, 2002), the units are assumed to be context-independent features. Contrary to both these suggestions, the current data suggest that the units need not be context-independent; they can be context-dependent allophones.

The present data do not show, however, that listeners use *only* allophones at the prelexical level. They suggest, more generally, that listeners use units which allow them to make functional generalizations to overcome the invariance problem in speech perception; prelexical units may thus not be restricted to one particular type of prelexical unit (as, e.g., proposed by Wickelgren, 1969). A segment like /f/, which is quite stable over context, may afford more context-independent coding than the context-dependent /r/ and /l/, and segments such as /s/, which is heavily influenced by the presence of rounded vowels, may fall in between. Whether a unit is used in prelexical processing may thus be related to how consistently that unit is produced in different contexts. This proposal motivates a new research program investigating the stability of units in production and their use in perception. The present research shows that allophones are used in prelexical phonological abstraction and hence that phonemes are not *the* unit of perception. That is, there may indeed be no universal, context-insensitive unit.

References

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory phonology 10* (pp. 91–111). Berlin: de Gruyter.

Dahan, D., & Mead, R. L. (2010). Context-conditioned generalization in adaptation to distorted speech. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 704–728. doi:10.1037/a0017449

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238. doi:10.3758/BF03206487

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279. doi:10.1037//0033-295X.105.2.251

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of Phonetics*, *31*, 305–320. doi:10.1016/S0095-4470(03)00030-5

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*, 373–405. doi:10.1016/j.wocn.2003.09.006

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review*, *18*, 943–950. doi:10.3758/s13423-011-0129-2

Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction. *Speech Communication*, *27*, 187–207. doi:10.1016/S0167-6393(98)00085-5

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*, 141–178. doi:10.1016/j.cogpsych.2005.05.001

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review*, *13*, 262–268. doi:10.3758/BF03193841

Ladefoged, P., & Maddieson, I. (1996). *Sounds of the world's languages*. Oxford: Blackwell Publishers.

Lahiri, A., & Reetz, H. (2002). Underspecified recognition. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 637–676). Berlin: Mouton de Gruyter.

Maddieson. (1984). The design of the UCLA Phonological Segment Inventory Database (UPSID). In *Patterns of sounds*. Cambridge University Press. Retrieved from http://dx.doi.org/10.1017/CBO9780511753459.012

Marslen-Wilson, W. D., & Warren, P. (1994). Levels of perceptual representations and processes in lexical access: Words, phonemes, and features. *Psychological Review*, *101*, 653–675.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*, 1113–1126. doi:10.1207/s15516709cog0000_79

McQueen, J. M., Norris, D., & Cutler, A. (1999). Lexical influence in phonetic decision-making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1363–1389. doi:10.1037/0096-1523.25.5.1363

Mitterer, H., Chen, Y., & Zhou, X. (2011). Phonological abstraction in processing lexical-tone variation: Evidence from a learning paradigm. *Cognitive Science*, *35*, 184–197. doi:10.1111/j.1551-6709.2010.01140.x

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238. doi:10.1016/S0010-0285(03)00006-9

Port, R. (2010). Rich memory and distributed phonology. *Language Sciences*, *32*, 43–55. doi:10.1016/j.langsci.2009.06.001

Remez, R. E. (1987). Neural models of speech perception: a case history. In S. Harnad (Ed.), *Categorical Perception: The groundwork of cognition* (pp. 199–225). Cambridge, Mass.: Cambridge University Press.

Savin, H. B., & Bever, T. G. (1970). The nonperceptual reality of the phoneme. *Journal of Vebal Learning and Verbal Behavior*, *9*, 295–302. doi:10.1016/S0022-5371(70)80064-0

Scharenborg, O., Norris, D., ten Bosch, L., & McQueen, J. M. (2005). How should a speech recognizer work? *Cognitive Science*, *29*, 867–918. doi:10.1207/s15516709cog0000_37

Sjerps, M. J., & McQueen, J. M. (2010). The bounds on flexibility in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 195–211. doi:10.1037/a0016803

Van Bezooijen, R. (2005). Approximant /r/ in Dutch: routes and feelings. *Speech Communication*, *47*, 15–31. doi:10.1016/j.specom.2005.04.010

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, *76*, 1–15.