Phonology and Phonetics
4-4

Editor
Aditi Lahiri

De Gruyter Mouton

Laboratory Phonology 10

edited by
Cécile Fougeron
Barbara Kühnert
Mariapaola D'Imperio
Nathalie Vallée

De Gruyter Mouton

How abstract phonemic categories are necessary for coping with speaker-related variation

Anne Cutler, Frank Eisner, James M. McQueen and Dennis Norris

Listeners can cope with considerable variation in the way that different speakers talk. We argue here that they can do so because of a process of phonological abstraction in the speech-recognition system. We review evidence that listeners adjust the bounds of phonemic categories after only very limited exposure to a deviant realisation of a given phoneme. This learning can be talker-specific and is stable over time; further, the learning generalizes to previously unheard words containing the deviant phoneme. Together these results suggest that the learning involves adjustment of prelexical phonemic representations which mediate between the speech signal and the mental lexicon during word recognition. We argue that such an abstraction process is inconsistent with claims made by some recent models of language processing that the mental lexicon consists solely of multiple detailed traces of acoustic episodes. Simulations with a purely episodic model without functional prelexical abstraction confirm that such a model cannot account for the evidence on lexical generalization of perceptual learning. We conclude that abstract phonemic categories form a necessary part of lexical access, and that the ability to store talker-specific knowledge about those categories provides listeners with the means to deal with cross-talker variation.

1. The problem of variability

Listeners are able to perceive speech sounds and spoken words reliably despite considerable variability in the acoustic signal. The factors underlying this variability are numerous. In addition to talker-specific components such as individual differences among talkers' vocal tract shapes and differences in dialect and affect, these factors include differences in rate, ambient noise, position (of sounds in words, and of words in sentences), and so on. No complete set of invariant physical attributes has been found that could be used to identify speech sounds reliably. The problem is that two utterances of the same speech sound are extremely unlikely to ever be physically identical, not even when produced by the same talker, and certainly not when produced by different talkers. Worse, physically identical sounds can elicit different phonemic percepts depending on context (Repp and Liberman 1987). In models of spoken-word recognition (e.g., Shortlist, Norris 1994) it is traditionally assumed that the perceptual system deals with such variability by extracting relevant information from the signal in a prelexical abstraction process. The products of this process are relatively simple abstract representations (e.g., phonemes) that can be mapped onto representations of words in the lexicon containing the same abstract symbolic sublexical units. According to an extreme version of this view, information about voice, affect, etc. is not used in the computations leading to lexical access.

Support for the view that perception of words and voices are independent processes is provided by findings suggesting that one function can be isolated from the other. In whispered speech or noiseband-vocoded speech, information about the identity of the talker is largely lost while comprehension remains fairly effortless (Shannon et al. 1995). Accordingly, different acoustic properties of the signal are said to carry information about one or the other perceptual function. Further evidence for the functional independence of voice processing and lexical access comes from double dissociations in neuropsychological investigations. In receptive types of aphasia, typically after left temporal lobe damage, speech comprehension is often impaired while voice recognition remains intact. Right temporal lobe infarctions, in contrast, can produce the reverse: impaired talker recognition without comprehension deficits (e.g., Peretz et al. 1994). Amnesic patients have been shown to display impaired voice-specific priming while maintaining intact repetition priming for words (Schacter, Church, and Bolton 1995).

The speech signal carries multiple acoustic cues to a particular speech sound in parallel, and the perceptual system can tolerate the absence of one or more such cues. Nevertheless, all cues are potentially informative, including indexical properties dependent upon talker identity. Nygaard, Sommers, and Pisoni (1994) showed this, for instance, by training listeners (for nine days) to identify previously unfamiliar voices and associate each with a name. Participants then heard new sets of words, in noise. They recognised significantly more words than listeners for whom the talkers were unfamiliar. Since exposure to talkers' voices facilitated later recognition of new words uttered by the same talkers, talker-specific information must have been encoded. This suggests that indexical and linguistic properties of the speech signal are closely interrelated and not independent (Pisoni 1997; Mullennix and Pisoni 1990).

Other studies on word or phoneme identification suggest that compensating for changes in talkers slows processing. Lists spoken by multiple talkers produce slower and less accurate identification than lists spoken by a single talker (Mullennix, Pisoni, and Martin 1989; Nusbaum and Morin 1992). Listeners in the multiple-talker conditions have to make perceptual adjustments to various voices, with greater processing demands. Pisoni and Lively (1995) suggest that, when a talker is heard, perceptual knowledge is obtained and retained in procedural memory; this may enhance processing efficiency of other utterances by this talker, as re-analysis of idiosyncratic voice properties becomes unnecessary. They report experiments in which native Japanese speakers were taught the English [r]/[l] contrast. Training with multiple speakers led to robust generalization of the newly-learned phonetic contrast to new talkers. This advantage was still present three months later.

There is thus strong evidence that talker-specific information plays a role in speech perception. An extreme abstractionist view, in which talker-specific information is discarded during lexical access, is therefore untenable. The opposite extreme view, that all talker-specific detail is stored in the mental lexicon, has been proposed as a radical alternative. Goldinger (1996, 1997, 1998), for example, suggests that the lexicon consists of specific instances of words which, among other attributes, include information about the talker's voice. The listener compares these representations with incoming acoustic information. In such an episodic lexicon, memory traces for words would be complex and detailed, and, as a byproduct, prelexical normalization procedures would be redundant (see also Klatt 1979; Johnson 1997b).

We argue here that extreme episodic models in which the lexicon consists only of detailed acoustic traces are just as untenable as extreme abstractionist models. While the evidence on talker specificity shows that knowledge about specific voices is stored in long-term memory, it does not show that this knowledge is stored in the mental lexicon. It could, for example, be stored prelexically. This would facilitate word recognition: If talker-specific knowledge influenced a relatively small set of abstract prelexical perceptual units, then that knowledge could be used in the recognition of all words containing those units. Once the prelexical system had learned about a talker idiosyncrasy which affected, for example, a single phoneme, that learning would automatically generalize across the vocabulary and thereby benefit the recognition of any word containing that sound which was spoken by that talker.

On this view, abstraction is an efficient way to deal with the variability problem. Through prelexical abstraction, the listener would be able to recognise the words that were intended by a given talker, irrespective of that talker's idiosyncrasies. Prelexical abstraction would thus allow the listener to map different acoustic events onto the same underlying lexical representations. We show that prelexical phonemic categories are indeed an essential part of word recognisounds reliably. The problem is that two utterances of the same speech sound are extremely unlikely to ever be physically identical, not even when produced by the same talker, and certainly not when produced by different talkers. Worse, physically identical sounds can elicit different phonemic percepts depending on context (Repp and Liberman 1987). In models of spoken-word recognition (e.g., Shortlist, Norris 1994) it is traditionally assumed that the perceptual system deals with such variability by extracting relevant information from the signal in a prelexical abstraction process. The products of this process are relatively simple abstract representations (e.g., phonemes) that can be mapped onto representations of words in the lexicon containing the same abstract symbolic sublexical units. According to an extreme version of this view, information about voice, affect, etc. is not used in the computations leading to lexical access.

Support for the view that perception of words and voices are independent processes is provided by findings suggesting that one function can be isolated from the other. In whispered speech or noiseband-vocoded speech, information about the identity of the talker is largely lost while comprehension remains fairly effortless (Shannon et al. 1995). Accordingly, different acoustic properties of the signal are said to carry information about one or the other perceptual function. Further evidence for the functional independence of voice processing and lexical access comes from double dissociations in neuropsychological investigations. In receptive types of aphasia, typically after left temporal lobe damage, speech comprehension is often impaired while voice recognition remains intact. Right temporal lobe infarctions, in contrast, can produce the reverse: impaired talker recognition without comprehension deficits (e.g., Peretz et al. 1994). Amnesic patients have been shown to display impaired voice-specific priming while maintaining intact repetition priming for words (Schacter, Church, and Bolton 1995).

The speech signal carries multiple acoustic cues to a particular speech sound in parallel, and the perceptual system can tolerate the absence of one or more such cues. Nevertheless, all cues are potentially informative, including indexical properties dependent upon talker identity. Nygaard, Sommers, and Pisoni (1994) showed this, for instance, by training listeners (for nine days) to identify previously unfamiliar voices and associate each with a name. Participants then heard new sets of words, in noise. They recognised significantly more words than listeners for whom the talkers were unfamiliar. Since exposure to talkers' voices facilitated later recognition of new words uttered by the same talkers, talker-specific information must have been encoded. This suggests that indexical and linguistic properties of the speech signal are closely interrelated and not independent (Pisoni 1997; Mullennix and Pisoni 1990).

Other studies on word or phoneme identification suggest that compensating for changes in talkers slows processing. Lists spoken by multiple talkers produce slower and less accurate identification than lists spoken by a single talker (Mullennix, Pisoni, and Martin 1989; Nusbaum and Morin 1992). Listeners in the multiple-talker conditions have to make perceptual adjustments to various voices, with greater processing demands. Pisoni and Lively (1995) suggest that, when a talker is heard, perceptual knowledge is obtained and retained in procedural memory; this may enhance processing efficiency of other utterances by this talker, as re-analysis of idiosyncratic voice properties becomes unnecessary. They report experiments in which native Japanese speakers were taught the English [r]/[l] contrast. Training with multiple speakers led to robust generalization of the newly-learned phonetic contrast to new talkers. This advantage was still present three months later.

There is thus strong evidence that talker-specific information plays a role in speech perception. An extreme abstractionist view, in which talker-specific information is discarded during lexical access, is therefore untenable. The opposite extreme view, that all talker-specific detail is stored in the mental lexicon, has been proposed as a radical alternative. Goldinger (1996, 1997, 1998), for example, suggests that the lexicon consists of specific instances of words which, among other attributes, include information about the talker's voice. The listener compares these representations with incoming acoustic information. In such an episodic lexicon, memory traces for words would be complex and detailed, and, as a byproduct, prelexical normalization procedures would be redundant (see also Klatt 1979; Johnson 1997b).

We argue here that extreme episodic models in which the lexicon consists only of detailed acoustic traces are just as untenable as extreme abstractionist models. While the evidence on talker specificity shows that knowledge about specific voices is stored in long-term memory, it does not show that this knowledge is stored in the mental lexicon. It could, for example, be stored prelexically. This would facilitate word recognition: If talker-specific knowledge influenced a relatively small set of abstract prelexical perceptual units, then that knowledge could be used in the recognition of all words containing those units. Once the prelexical system had learned about a talker idiosyncrasy which affected, for example, a single phoneme, that learning would automatically generalize across the vocabulary and thereby benefit the recognition of any word containing that sound which was spoken by that talker.

On this view, abstraction is an efficient way to deal with the variability problem. Through prelexical abstraction, the listener would be able to recognise the words that were intended by a given talker, irrespective of that talker's idiosyncrasies. Prelexical abstraction would thus allow the listener to map different acoustic events onto the same underlying lexical representations. We show that prelexical phonemic categories are indeed an essential part of word recognition, and hence that the mental lexicon cannot consist only of detailed episodic traces.

2. Lexically-guided perceptual learning

Recent findings on perceptual learning in speech perception demonstrate that the perceptual system adjusts rapidly to idiosyncratic articulation of a phonemic contrast by a particular talker. Norris, McQueen, and Cutler (2003) conducted a two-phase experiment. In an initial training phase, Dutch participants listened to words that ended in [f] or [s]. For one group of subjects, the final [f] in these words (e.g., *olijf*, 'olive') were replaced with an ambiguous fricative midway between [f] and [s], but the [s]-final words (e.g., *radijs*, 'radish') remained natural. A second group received words manipulated in the reverse pattern, with natural sounding [f]-final words and the ambiguous fricative [?] replacing [s]. A control group listened to a set of nonwords which ended with the ambiguous sound. These critical items were presented interspersed with other words and nonwords that contained neither [f] nor [s], in the context of a lexical decision task. In the experimental groups 90% of [?]-final items were accepted as real words. After this training, participants were asked to categorize sounds from an [ɛf]–[ɛs] continuum (the same series from which the ambiguous [?] had been selected). When compared to the control group, participants who had listened to the natural [s]-final words and ambiguous [f]-final words were more likely to categorize sounds on the continuum as [f], whereas those who had received the reversed training categorized more sounds as [s] (see Figure 1). Norris et al. argued that this result reflects a prelexical adjustment in how the acoustic signal is mapped onto a phonemic category.

Eisner and McQueen (2005) investigated this type of perceptual learning further by testing whether, under similar lexically-biased training conditions, a modulation of the [f]/[s] category boundary is specific to the talker whose ambiguous productions caused the adjustment, or generalizes to speech from others. For the adjustment to be useful to the listener, it should only be applied again when speech from the training talker is encountered. It is less likely to be beneficial if applied to a whole language community, in the absence of evidence that other talkers share the speech idiosyncrasy. The results suggested that learning was indeed highly talker-specific: Listeners applied the category boundary modulation only to fricative test sounds uttered by the training talker (see Figure 1). Effects of equal magnitude were observed even when these sounds were presented in the context of carrier vowels from other talkers which elicited the percept of a talker change. No effect was found with test fricatives produced by another talker. An effect did appear, however, when, under identical test conditions, this novel talker's ambiguous fricatives had been spliced into the original talker's utterances during training.

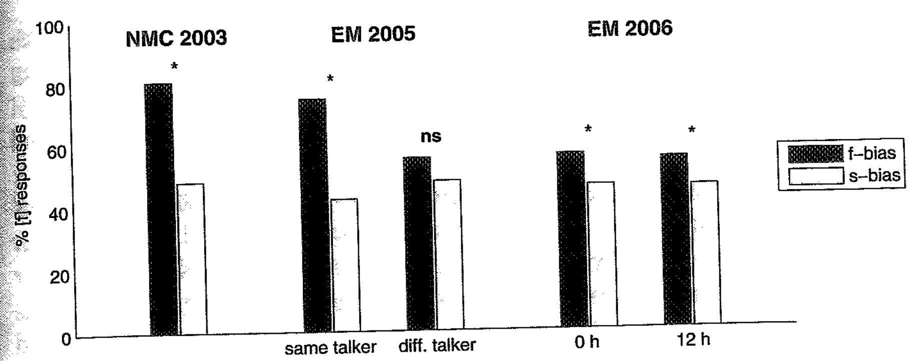

Figure 1. Mean percentages of [f] categorizations for groups with [f]- or [s]-biased training in the studies by Norris, McQueen, & Cutler, 2003 (NMC 2003), Eisner & McQueen, 2005 (EM 2005), and Eisner & McQueen, 2006 (EM 2006). Starred differences are statistically significant at $p < .05$.

Two recent studies with this training–test paradigm have further qualified the conditions under which talker-specific learning occurs. Kraljic and Samuel (2006) found that learning generalized to speech input from another talker in the case of the English [d]–[t] stop contrast, putatively because stops contain less information about the identity of the talker than fricatives, so that talker-specific learning is harder to achieve. In addition to generalization of learning to another talker, Kraljic and Samuel observed generalization to another place of articulation ([b]–[p]). Similar transfer effects across place of articulation have been observed for learning a novel VOT contrast (e.g., Tremblay et al. 1997) and for selective adaptation (Eimas and Corbit 1973). For the case of these stop consonants, the perceptual adjustment may thus mainly affect a voicing cue which is relatively abstract. One interpretation of the talker-generalization effect with stops is that the temporal VOT cue is adjusted at a higher level in the perceptual system than the low-level spectral manipulation employed in the studies with fricatives.

Kraljic and Samuel (2005) investigated the conditions under which perceptual learning might be reversed, using the English [s]/[ʃ] fricative contrast. They again employed a variant of the Norris et al. (2003) paradigm, but with a 25-minute delay between training and test. The delay by itself produced no decrease of the effect. Hearing either a talker other than the training talker produce unambiguous tokens of the critical trained sounds during the delay, or hearing the training talker produce speech that contained none of the critical sounds, also had no effect on the magnitude of perceptual learning. Hearing the training talker produce unambiguous versions of the trained sounds during the delay, however, did reduce the effect. This pattern was obtained both for male and female voices, and is in line with the findings of Eisner and McQueen (2005), as it suggests talker-specificity of perceptual learning. All of these conditions were also run with a talker change in the test phase. The results of the talker change conditions were asymmetrical, such that conditions with a male talker at training and a female talker at test showed no perceptual learning effect, suggesting talker-specificity, whereas hearing the female talker during training and the male talker in the test phase did show an effect, suggesting generalization. Kraljic and Samuel proposed that these results were caused by an asymmetry in the average spectral centre of gravity of the training and test stimuli, as revealed in an acoustic analysis. Generalization of learning was more likely to occur when the fricative sounds used at training and test were spectrally similar, whereas talker specificity was associated with spectrally more dissimilar sounds.

The current data on the specificity of lexically-guided perceptual learning in speech suggest that, while there may be situations in which generalization occurs (e.g., after multiple-talker exposure, or when learning adjusts a more abstract featural representation), there are clear cases of talker-specific learning. Talker-specific knowledge affects the processing of fine phonetic detail, which in turn affects the phonetic category boundary between two speech sounds, and must therefore be stored in some way by the perceptual system.

3. Stability of learning

If learning about a talker's idiosyncrasies is of real value to the listener, then that learning ought to be stable over time. Eisner and McQueen (2006) therefore investigated whether lexically-driven adjustments are short-lived, or remain stable and can be reapplied when the listener re-encounters the same talker later. Listeners either (a) were exposed to manipulated speech in the morning and tested 12 hours later, or (b) were trained in the evening and tested the following morning (again 12 hours later). All participants were also tested immediately after training.

The group which learned in the evening should have received less (potentially interfering) speech input from other talkers and they slept for at least six hours during the delay. For both reasons this group's learning may be more stable. Fenn, Nusbaum, and Margoliash (2003), who trained listeners on transcribing poorly synthesised speech, found that performance improvement due to such training decayed over a day but not over a night of sleep. The learning in the present experiments, however, took place without explicit training and generally without listeners' awareness. Accommodating an unusual pronunciation of a speech sound presumably reflects a process which listeners engage very frequently and which is thus, in contrast to dealing with synthetic speech, highly overlearned. Learning which is constantly useful to the listener should not require a lot of time to consolidate.

Eisner and McQueen (2006) found significant perceptual learning immediately after training. After 12 hours, the learning had not decreased (see Figure 1). Furthermore, the effect was just as stable for the group who had been awake during the delay as for the group who had slept. Thus perceptual learning remained very stable during the interval, with neither a decay during waking due to interference from other talkers, nor an additional benefit from the opportunity for consolidation of learning during sleep. One further difference between this study and the original Norris et al. (2003) study is noteworthy: instead of making lexical decisions in the training phase, listeners heard a short story (644 words, in which every [f] or [s], 78 in either case, had been changed to the ambiguous fricative). The adjustment in the fricative boundary caused by just listening to the story suggests that lexically-guided perceptual learning is automatic; it does not depend upon explicit judgements being made to words containing the ambiguous fricative during the training phase (see also McQueen, Norris, and Cutler 2006).

4. Lexical generalizability

The strongest evidence that lexically-driven perceptual learning has a prelexical locus would come from a demonstration of lexical generalizability. If the listener has learned that a talker produces an [f] sound in an unusual way, and that knowledge is coded at the prelexical level, then recognition of all words containing that talker's unusual [f] will be affected.

Testing for lexical generalization is also critical with respect to episodic models. Evidence of lexically-guided perceptual learning which comes from metalinguistic judgement tasks such as phoneme categorization could be accounted for by episodic models (e.g., Johnson 1997b) in which metalinguistic judgements about phonological categories are made postlexically. In such models, the categorizations can be based on abstractions made over lexical episodic traces. The adjustments to phonemic categories we have described so far are thus not necessarily inconsistent with episodic models. If it could be shown, however, that the learning generalizes to the processing of words which were not presented during training, this would indicate that the learning affected phonemic categories with a functional role in the lexical access process, namely, abstract prelexical representations. Such evidence would suggest that word recognition requires phonological abstraction, and therefore does not consist solely of the storage of detailed acoustic episodes.

McQueen, Cutler, and Norris (2006) found that the results of lexically-guided perceptual learning are indeed applied across the lexicon. Their training conditions were identical to those of Norris et al. (2003): an auditory lexical decision phase in which either the [f] in 20 [f]-final words or the [s] in 20 matched [s]-final words was replaced with an ambiguous fricative. This lexically-biased training, however, was followed by a cross-modal identity priming task, which measures lexical activation, and so assesses generalization of learning over the vocabulary. Two groups of listeners had either [f]-or [s]-biased training, followed by a test phase in which an auditory prime was presented before a visual target word or target nonword, on which they made lexical decisions. The critical materials consisted of 20 minimal pairs of Dutch words such as *doof-doos* ('deaf-box'). If training leads to prelexical adjustments to the [f]-[s] boundary, then [do:?] should be heard as *doof* by listeners with [f]-biased training, and as *doos* by listeners with [s]-biased training. This was measured in the priming task by comparing speed and accuracy of lexical decisions to visual DOOF or DOOS after hearing [do:?] versus after hearing a phonologically completely unrelated word (e.g., [krɔp], 'head of lettuce'). Previous research in Dutch (van Alphen and McQueen 2006) has shown that, relative to an unrelated condition, there is facilitation of responses to visual words when those words have just been heard, but not when the target and the preceding spoken word differ in one phoneme. Facilitation of responses in the related condition relative to the unrelated condition would therefore indicate that the listener had interpreted [do:?] as a token of the visually-presented word (DOOF or DOOS).

After a prime such as [do:?], listeners were faster and more accurate in their responses to a target containing a fricative consistent with their training. Responses to DOOS were facilitated for listeners with [s]-biased training, while responses to DOOF were facilitated for listeners with [f]-biased training (see Figure 2). In error rates there was also an inhibitory effect. There tended to be more "no" responses to visual target words containing a fricative inconsistent with training (e.g., more "no" responses to DOOS after [f]-biased training).

None of the words in the priming phase had been part of the training phase. These results therefore suggest that the perceptual adjustment induced during training affected a prelexical stage of processing, allowing learning to transfer to other words in the lexicon. Talker-specific adjustments to abstract prelexical categories are thus beneficial for the listener: They help with the recognition of other words spoken by that talker. In the case of minimal pairs, the adjustments have acted to resolve what would otherwise have been a lexical ambiguity; due to the training, the listener knows whether to interpret [do:?] as either *doof* or *doos*.

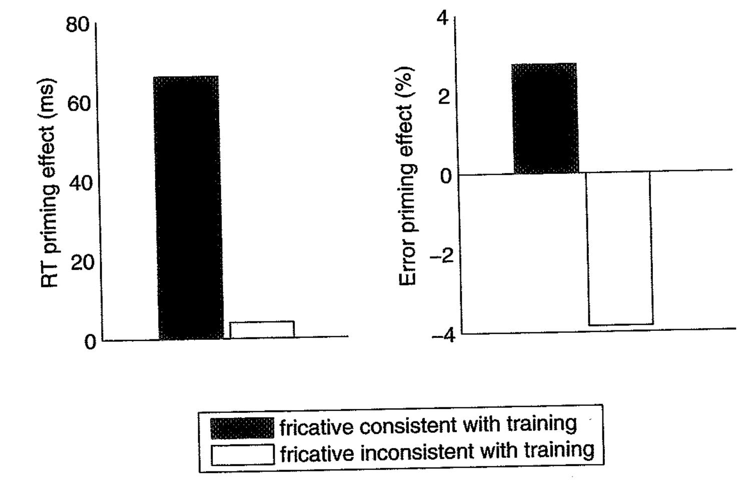

Figure 2. Mean priming effect (unrelated–related) in reaction times (RTs, in ms; higher values indicate faster responses) and error rates (in %; positive values indicate fewer errors) for visual targets containing a fricative either consistent or inconsistent with listeners' lexically-biased training. Data from McQueen et al. (2006).

5. Simulations with an episodic model

Abstract but flexible prelexical representations thus provide the means by which listeners can deal with phonetic variability. An extreme episodic model, in which word recognition entails a comparison of the current input, in all its detail, with previous lexical episodes, lacks this stage of abstraction, and thus lacks this flexibility. Without prelexical abstraction, such a model cannot capitalize upon sublexical regularities in the current talker's speech during word recognition. A model of this type should therefore be unable to explain the lexical generalization data. Acoustic traces corresponding to the training trials could be stored in the lexicon, but these traces should not affect the match between [do:?] and previous traces of *doof* and *doos*. We tested the validity of this claim directly, in simulations with MINERVA-2 (Hintzman 1986), following the example of

Goldinger (1998). Goldinger argued that to assess the capability of episodic models, it is best to begin with "pure" models, of which MINERVA-2 is by far the best-known example.

In MINERVA-2, each episode of previous experience lays down a trace in Long-Term Memory (LTM); applying this model to word recognition means that these traces are episodes of heard words. When a new input arrives, a probe activates all traces in LTM in proportion to their match to the probe's acoustic content. An aggregate echo of all activated traces is then returned to Working Memory. Echoes have two properties: echo intensity (the summed activation of all traces contributing to the echo) and echo content (the values of each element in the echo vector). For the equations used to compute these characteristics, see Appendix A of Goldinger (1998). The focus here will be on echo content.

If MINERVA-2 receives the 20 training trials of the lexical generalization experiment, the effect will be to add traces that are very similar to existing traces for those 20 words, except that their final portion does not correspond exactly to anything anywhere in the traces already in the lexicon. Every other trace for every other word will be unaffected. When a test stimulus is presented, it is matched against all traces in memory. The activation of each trace is a function of the overall similarity between it and the input. The content of the retrieved vector is then determined by summing over all traces multiplied by their activation. It is immediately clear that the very limited number of new episodes (one per word) added during the training is unlikely to be enough to change the model's performance. The 20 critical episodes would have to compete with the effect of many thousands of existing episodes.

Even infinitely many ambiguous training episodes, however, would not produce the experimental finding. Ambiguous training stimuli just add ambiguous traces to the lexicon; no benefit flows to other words. Whether or not a test word and the training words share the same ambiguous portion, the training phase will not affect the test word's interpretation. When a test episode is presented, its ambiguous phoneme will match the corresponding part of the traces of the critical training items, and this will increase the activation of those traces. Activation will increase for the entire trace, however, not just those elements corresponding to the ambiguous phoneme. The effect of these episodes will be to bias the content of the echo towards an ambiguous fricative interpretation, not towards [f] or [s] – the training episodes thus resonate with the test input, but bias the resulting echo in neither direction.

The previous episodes corresponding to each member of the minimal pairs will also have no biasing effect on echo content. A test item such as [do:?] will match equally poorly to both sets of episodes, and thus to both possible interpretations of the test item (e.g., *doof* or *doos*). Furthermore, the activation of these two sets of traces, and hence their effect on echo content, cannot be biased in either direction by the presence of the traces with ambiguous phonemes left by the training words. There is therefore nothing in the model that would make a test item more likely to retrieve a vector corresponding to the training-consistent interpretation of the ambiguous phoneme than one corresponding to the training-inconsistent interpretation.

The model's situation is actually even worse. As already noted, adding 20 new episodes to the lexicon may have no detectable effect. But consider what would happen if the training phase of the experiment were repeated many times. In the test phase, as we saw, the ambiguous training tokens would have little effect, but additional unambiguous training tokens would indeed contribute to echo content. Recall that recognition in this model involves activation of stored traces in proportion to their match to the current input. If training includes many words ending with unambiguous [s], but no words ending with unambiguous [f], [s]-final word traces would become stronger than [f]-final word traces, and thus would have more effect than [f]-final traces on echo content given a probe partially matching their content. The echo evoked by an ambiguous test word would thus be more strongly dominated by [s]-final word traces partially matching the test word than by [f]-final traces. As a result, ambiguous test words would more strongly activate the interpretation consistent with the unambiguous training phoneme. This of course is exactly the opposite pattern from that observed in the data.

To confirm that this description of the model is correct, we performed simulations of the McQueen et al. (2006) experiment in MINERVA-2. These necessitated certain assumptions about the model, differing from those made in Goldinger's (1998) simulations of shadowing data, but still preserving MINERVA-2's core storage and retrieval mechanisms.

5.1. Assumptions

5.1.1. *Form versus name vectors*

In Hintzman's (1986) formulation of MINERVA-2, 10 elements in the vector of each trace encode the category name and 13 elements encode the stimulus pattern. Each stimulus event can thus be associated with a category label represented by the pattern across the name elements. In Goldinger's (1998) implementation, each episodic trace was a vector of 200 elements: 100 name elements representing the acoustic-phonetic information in the word; 50 voice elements representing the voice-specific aspects of the word's acoustic form; and 50 context elements representing other components of the word's acoustic form (e.g., background noise). Goldinger's name fields thus seem to have represented the raw acoustic-phonetic traces. Our version of MINERVA-2 resembled that of Goldinger in that we had long multi-element vectors which easily enable sub-portions to be deemed to stand for "phonemes", but it resembled Hintzman's original version in preserving a distinction between name and form subvectors. Both types of element seemed necessary. Without category labels the model could not recognize input as an instance of any category, and thus could not comprehend speech (it could only reverberate varyingly to different stimuli). The other (form) elements then stand for the acoustic-phonetic patterns in the input, the primary material for storage. In our version of the model, therefore, form fields encoded stimulus properties and name fields coded category identity.

5.1.2. *Cross-modal priming*

In cross-modal priming, the physical forms of the auditory and visual stimuli are completely different, so that priming cannot arise from overlap between perceptual components of auditory and visual episodes. The most probable interpretation of cross-modal priming is that its origin is lexical. The experimental literature suggests that incomplete phonological overlap between auditory primes and visual targets is insufficient to produce priming (e.g., van Alphen and McQueen 2006). It thus seems reasonable to assume that priming involves modality-independent components of lexical representations.

One possibility for simulating cross-modal priming in MINERVA-2 is that priming effects could be mediated by modality-independent representations. This could be, for example, the prime and target words' semantic features. Alternatively, experience with reading and writing might have led to associations between orthographic and auditory forms. Presentation of an auditory form could then make available a representation of the corresponding orthographic form. In either case, the auditory form would retrieve another representation that would be more or less similar to some part of the representation elicited by the visual form. In the present simulations we simply assume that such representations correspond to the name components of the lexical episodes.

5.1.3. *Lexical decision*

Because episodic models have no single canonical lexical representation against which to match the input, lexical decision cannot be modeled by decisions concerning match between input and pre-existing unitary representations. However, as the critical data here concern the relative amount of priming in two conditions, it is unnecessary to model the lexical decision task itself; comparison of lexical activation in the two conditions is sufficient. The values we report are therefore the percentage of times that the echo content of the retrieved name field or form

field is more similar to the training-consistent than the training-inconsistent interpretation of the target. If the training phase had no effect at all, the score would be 50%. If perceptual learning were complete, so that the ambiguous sound was always interpreted in a training-consistent manner, probing with a word with the ambiguous sound should have a similar effect to probing with a word with the appropriate unambiguous sound, and the score would be 100%.

5.2. Method

Our initial lexicon comprised 20 traces of each of 500 words. Each word consisted of a 400-element vector with elements randomly set to 1 or −1; 200 elements represented the name field and 200 represented the form field. Eighty items in the lexicon were used to represent the stimuli in critical training and test trials. Forty items corresponded to words ending in [s], and 40 to words ending in [f]. Twenty elements of the form field stood for the critical phoneme. In this sub-vector, odd numbered elements set to 1 and even numbered elements set to −1 represented [f], and the complement of this pattern represented [s]. Ambiguous phonemes were then represented by setting the first 10 elements to 1 and the rest to −1. Forty of the words available as training items could not be changed to any other word simply by changing the critical phoneme from [f] to [s] or vice versa; the 40 items for the test phase formed 20 pairs differing only in the critical phoneme.

Training involved adding episodes for 20 ambiguous items from the training set, all of which originally ended with the same final phoneme, and 20 episodes of unambiguous items ending with the other final phoneme. During training the episodes added to the lexicon consisted of both the name and form components of the vectors corresponding to each item. In the test phase the form fields alone (of each of the 20 critical minimal pairs, with an ambiguous final phoneme) were used to retrieve echoes from the model. In each case, the content of the echo was compared, separately for the form and name fields of the echo, to the two possible interpretations of the ambiguous word to determine whether it was more similar to the trained than the untrained interpretation. Each separate simulation run consisted of generating a new lexicon, then adding episodes for the training, and then probing with the form field of each test stimulus. Only a single episode corresponded to each test word. These episodes corresponding to the test probes were not added to the lexicon; this allowed the test phase to be repeated without altering the lexicon content.

The test phase was repeated ten times in each simulation run. Two sets of simulation runs are reported. In one set each item in the training phase was presented once, followed by the test phase (repeated ten times). In the other set ten new episodes were added for each item in the training phase, followed by the test phase (again repeated ten times). This second set allowed us to control for the possibility that the effect of training was too weak (the single training episode per stimulus may have no discernible effect given the 20 traces for each word already in the lexicon). Note also that in the experiment listeners received additional training during the test phase (60 trials with the ambiguous fricative in lexically-biased contexts). The condition in which training was multiplied by ten thus also served to simulate even more training than that in the actual experiment's test phase. Simulations reported are averaged over 1000 runs, each starting with a new lexicon. To reflect variability in the episodes, random noise (reversal of the sign of each form element with a probability of 0.25) was applied to all training and test episodes in each simulation run.

5.3. Results

The model was 99.98% correct in recognising the trained items. Its performance on the test items was much poorer. Training made no difference to name retrieval at test. As Table 1 shows, scores were at chance whether the training phase occurred once or 10 times.

Table 1. MINERVA-2 Simulations. Percentage of test trials where echo content is more similar to the training-consistent interpretation of the target than the training-inconsistent interpretation.

Number of training phases	Name field	Form field
1	50	48
10	50	35

As predicted, and as also shown in Table 1, the results for form retrieval were quite different. When there was only a single pass through the training stimuli, the score was a little below chance, representing a slight reversal of the effect found in the human data. With 10 repetitions of the training phase, the score was 35%, almost identical to the effect that was obtained if the ambiguous stimuli were omitted from the training phase (36%). This indicates that the reversal in the test condition is due to strengthening of the echo that is caused by activation of traces corresponding to training words containing the unambiguous phoneme. Note that we did not distinguish between voice, context and acoustic-phonetic elements within the form field. Adding speaker-specific elements to the form fields would however only aggravate the problem further. If the training and test traces were made more similar to each other and less similar to the other traces in the lexicon (i.e., by coding those traces as all coming from the same novel speaker), the effect of the unambiguous training episodes on test echo content could only become stronger.

5.4. Discussion

A pure episodic model is therefore unable to simulate the data from McQueen et al. (2006). The model does not show generalization from the training words that biases recognition of the novel test words in the direction indicated by training. The reason for this is that the generalization process in MINERVA-2 (schema abstraction; Hintzman, 1986) is inadequate for this type of generalization. When episodic models are presented with a number of episodes embodying variation on some prototype representation, they can abstract a representation of the prototype. Presentation of one of these episodes will tend to retrieve a representation more similar to the prototype than the episode itself. In the present case, presentation of any particular unambiguous token of *doof* will result in an echo that is more similar to the prototypical pronunciation of the word than that token happens to be. But when the input is ambiguous [do:?] the resulting echo will be a balanced mixture of *doof* and *doos* (to the extent that the ambiguous phoneme is perfectly ambiguous and the two words are of equal frequency and thus representational strength).

The reason that the episodes corresponding to the ambiguous training words in one or the other training condition cannot tip the balance in either direction is that the model does not form abstract representations of components of these trace vectors. The effect of the critical component of the ambiguous training episodes on the resulting echo is to make the equivalent component of that echo more like the ambiguous phoneme, and not more like either [f] or [s]. Instead of a training effect consistent with the experimental data on the form fields of the test trial echoes, the model predicts, because of the effect of the form fields of the unambiguous training words on the test echoes, a pattern opposite to that observed in the experiment. The name fields of the ambiguous training episodes can have no effect on the name fields of these echoes either, because there is no relationship between the name fields of, for example, *olijf* and *doof*.

One of the main benefits claimed for episodic models is that they render normalization unnecessary. Abstract representations of any form require some degree of normalization. Nonetheless, one might ask whether MINERVA-2 could be adapted to contain abstract prelexical representations in addition to detailed perceptual traces, and might then capture the results. As the present simulations show, however, the model is unable to account for the lexical generalization effect irrespective of the content of the name fields (i.e., we did not stipulate whether these vectors contained raw acoustic or phonologically abstract elements). An "episodic" model with abstract representations in the name fields would therefore have to account for the perceptual learning effect by retuning the mapping between the speech input and those abstract categories – this would be the only way that the learning would generalize to the recognition of new words. The model would therefore be an abstract model in all but name. Furthermore, if episodes did contain abstract representations, then all instances of a word would contain the same abstract representations. Storing these representations with every episodic trace rather than in a single abstract lexical representation would be completely redundant.

6. Conclusions

Listeners need abstract prelexical representations of speech sounds in order to deal with variation in the speech signal. Although we have focussed here on speaker-related variability, the same argument applies to other sources of variability. In the case where a talker produces a particular speech sound in an unusual way, the use of abstract prelexical representations in decoding speech is both efficient and beneficial. It is efficient in that knowledge about talker idiosyncrasies can be coded for a single sublexical representation, rather than separately for all words in the lexicon containing the unusual sound. It benefits comprehension in that once such talker-specific knowledge has been acquired, it can assist in the recognition of all words containing the unusual sound.

McQueen et al. (2006) have provided experimental evidence of the benefits that prelexical abstraction has for word recognition: Adjustments to a fricative category do indeed generalize across the vocabulary. Other recent data also show that phonetic learning can generalize over words (Davis et al. 2005; Maye, Aslin, and Tanenhaus 2008). Although these demonstrations are central in showing the functional role that sublexical abstract representations play in lexical access, it is important to note that there is a considerable body of other evidence supporting prelexical abstraction (see McQueen et al. 2006).

The evidence from the perceptual learning studies suggests not only that there are prelexical abstract representations, but also that prelexical processing is quite flexible. There is a critical constraint on this flexibility, however. It is well known that listeners learning a second language have difficulty acquiring new phonemic categories (Strange 1995). In contrast to the rapid and apparently automatic learning in the Norris et al. (2003) paradigm, listeners in a second language need considerable training to acquire a phonemic distinction that was absent in their first language (e.g., Japanese listeners acquiring the /l/-/r/ distinction; Logan, Lively, and Pisoni 1991). This kind of evidence offers further support for prelexical abstraction. After children have learned their first language, and prelexical representations of the phonology of their language have been established, those representations strongly affect how speech both in the first and in any subsequent language will be perceived (Best 1994). This phonological sieve also influences word recognition. Even after extensive exposure to a second language, recognition of second-language words is still influenced by the phonological structure of a first language (Cutler, Weber, and Otake 2006; Pallier, Colomé, and Sebastián-Gallés 2001; Weber and Cutler 2004).

The prelexical processing stage is thus rather inflexible with respect to the acquisition of new phonemic categories, but at the same time very flexible with respect to adjustments to existing categories. Effects of first-language categories on word recognition in spite of almost a lifetime's exposure to second language episodes poses a challenge to extreme episodic models (Pallier et al. 2001). Our simulations with a version of MINERVA-2 show that the flexibility of first-language categories also poses a serious challenge to any extreme episodic model. Models in which lexical access consists only of the comparison of acoustically detailed traces to previously stored traces, with no assistance from abstract sublexical representations, are unable to explain the McQueen et al. (2006) data on lexical generalization. This is because any knowledge that such models may have about the compositionality of language (e.g., that *giraffe* ends with the same sound as *flamingo* begins with) is not used during lexical access.

The empirical data which motivated episodic models is described in section 1 and reviewed by Goldinger (1998): Talker-specific effects are found in memory for words, in shadowing, and in phoneme and word identification. This evidence refutes any purely abstractionist model in which talker- (and situation-)specific characteristics of a speech episode are normalized away and forgotten. But it does not refute an abstractionist model where episodes are retained but not as part of the lexical system. Some data on talker-specific effects indeed support this view. Luce and Lyons (1998) showed that listeners in an old/new memory task (judging whether words had been heard earlier) recognized words spoken by the same speaker as "old" more quickly than words spoken by a different speaker. But repetition priming in auditory lexical decision with the same materials was no larger for within- than for across-speaker repetitions. This suggests that talker-specificity effects may not reflect lexical-level processing. We are not aware of any evidence on episodic effects which requires episodes to be stored in the lexicon.

The data showing episodic influences across a range of tasks and abstraction in word recognition are explicable in terms of a hybrid model: a word-recognition system with abstract prelexical and lexical representations combined with an episodic memory system that is distinct from the mental lexicon (and from the prelexical processor). The prelexical normalization and abstraction process which deals with talker variability can cause processing costs when multiple speakers are encountered, but also results in benefits in word recognition given continued exposure to a talker. Critically, the claim that there is this kind of abstraction prior to lexical access does not require that the detail in specific episodes is forgotten, nor that this detail could not influence performance in, for example, recognition-memory tasks.

Many models have recently been proposed in which a central role in spoken-language processing is assigned to detailed acoustic traces of linguistic experience in memory (Bybee 2001; Goldinger 1998; Hawkins 2003; Johnson 1997a, 1997b; Klatt 1979, 1989; Pierrehumbert 2001, 2002). Speaker-related variation is one of the main motivating factors for models of this type. Our research indicates that episodic traces offer an insufficient account of how listeners cope with speaker-related variation. Listeners cope with such variation by using lexical information to retune abstract phonemic categories, thus allowing rapid generalization of the retuning to other lexical items. Any model of spoken-word recognition, with or without episodic representations, can only capture this speaker-related retuning if it includes a process of prelexical abstraction.

References

Best, Catherine T.
1994 The emergence of native-language phonological influences in infants: A perceptual assimilation model. In: Judith C. Goodman and Howard C. Nusbaum (eds.), *The development of speech perception: The transition from speech sounds to spoken words,* 167–224. Cambridge, MA: MIT Press.

Bybee, Joan
2001 *Phonology and language use.* Cambridge: Cambridge University Press.

Cutler, Anne, Andrea Weber and Takashi Otake
2006 Asymmetric mapping from phonetic to lexical representations in second language listening. *Journal of Phonetics* 34: 269–284.

Davis, Matthew H., Ingrid S. Johnsrude, Alexis Hervais-Adelman, Karen Taylor and Carolyn McGettigan
2005 Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General* 134: 222–241.

Eimas, Peter D. and John D. Corbit
1973 Selective adaptation of linguistic feature detectors. *Cognitive Psychology* 4: 99–109.

Eisner, Frank and James M. McQueen
2005 The specificity of perceptual learning in speech processing. *Perception & Psychophysics* 67: 224–238.

Eisner, Frank and James M. McQueen
2006 Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America* 119: 1950–1953.

Fenn, Kimberly M., Howard C. Nusbaum and Daniel Margoliash
2003 Consolidation during sleep of perceptual learning of spoken language. *Nature* 425: 614–616.

Goldinger, Stephen D.
1996 Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22: 1166–1183.

Goldinger, Stephen D.
1997 Words and voices: Perception and production in an episodic lexicon. In: Keith Johnson and John W. Mullennix (eds.), *Talker variability in speech perception,* 33–66. San Diego, CA: Academic Press.

Goldinger, Stephen D.
1998 Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105: 251–279.

Hawkins, Sarah
2003 Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics* 31: 373–405.

Hintzman, Douglas L.
1986 "Schema abstraction" in a multiple-trace memory model. *Psychological Review* 93: 411–428.

Johnson, Keith
1997a The auditory/perceptual basis for speech segmentation. *Ohio State University Working Papers in Linguistics* 50: 101–113.

Johnson, Keith
1997b Speech perception without speaker normalization: An exemplar model. In: Keith Johnson and John W. Mullennix (eds.), *Talker variability in speech processing,* 145–165. San Diego, CA: Academic Press.

Klatt, Dennis H.
1979 Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics* 7: 279–312.

Klatt, Dennis H.
1989 Review of selected models of speech perception. In: William D. Marslen-Wilson (ed.), *Lexical representation and process,* 169–226. Cambridge, MA: MIT Press.

Kraljic, Tanya and Arthur G. Samuel
2005 Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology* 51: 141–178.

Kraljic, Tanya and Arthur G. Samuel
2006 Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review* 13: 262–268.

Logan, John S., Scott E. Lively and David B. Pisoni
1991 Training Japanese listeners to identify English /r/ and /l/. *Journal of the Acoustical Society of America* 89: 874–886.

Luce, Paul A. and Emily A. Lyons
1998 Specificity of memory representations for spoken words. *Memory & Cognition* 26: 708–715.

Maye, Jessica, Richard Aslin and Michael Tanenhaus
2008 The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science* 32: 543–562.

McQueen, James M., Anne Cutler and Dennis Norris
2006 Phonological abstraction in the mental lexicon. *Cognitive Science* 30: 1113–1126.

McQueen, James M., Dennis Norris and Anne Cutler
2006 The dynamic nature of speech perception. *Language and Speech* 49: 101–112.

Mullennix, John W. and David B. Pisoni
1990 Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics* 47: 379–390.

Mullennix, John W., David B. Pisoni and Christopher S. Martin
1989 Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America* 85: 365–378.

Norris, Dennis
1994 Shortlist: A connectionist model of continuous speech recognition. *Cognition* 52: 189–234.

Norris, Dennis, James M. McQueen and Anne Cutler
2003 Perceptual learning in speech. *Cognitive Psychology* 47: 204–238.

Nusbaum, Howard C. and Todd M. Morin
1992 Paying attention to differences among talkers. In: Yoh'ichi Tohkura, Eric Vatikiotis-Bateson, and Yoshinori Sagisaka (eds.), *Speech perception, production, and linguistic structure,* 113–134. Tokyo: Ohmsha.

Nygaard, Lynne C., Mitchell S. Sommers and David B. Pisoni
1994 Speech perception as a talker-contingent process. *Psychological Science* 5: 42–46.

Pallier, Christophe, Angels Colomé and Núria Sebastián-Gallés
2001 The influence of native-language phonology on lexical access: Exemplar-based versus abstract lexical entries. *Psychological Science* 12: 445–449.

Peretz, Isabelle, Régine Kolinsky, Mark Tramo, Raymonde Labrecque, Claude Hublet, Guy Demeurisse and Sylvie Belleville
1994 Functional dissociations following bilateral lesions of auditory cortex. *Brain* 117: 1283–1301.

Pierrehumbert, Janet B.
2001 Exemplar dynamics: Word frequency, lenition and contrast. In: Joan Bybee and Paul Hopper (eds.), *Frequency and the emergence of linguistic structure,* 137–157. Amsterdam: John Benjamins.

Pierrehumbert, Janet B.
2002 Word-specific phonetics. In: Carlos Gussenhoven and Natasha Warner (eds.), *Laboratory phonology,* Volume 7, 101–139. Berlin: Mouton de Gruyter.

Pisoni, David B.
1997 Some thoughts on 'normalization' in speech perception. In: Keith Johnson and John W. Mullennix (eds.), *Talker variability in speech processing,* 9–30. San Diego, CA: Academic Press.

Pisoni, David B. and Scott Lively
1995 Variability and invariance in speech perception: A new look at some old problems in perceptual learning. In: Winifred Strange (ed.), *Speech perception and linguistic experience: Issues in cross-language research,* 433–459. Baltimore: York Press.

Repp, Bruno H. and Alvin M. Liberman
1987 Phonetic category boundaries are flexible. In: Stevan Harnad (ed.), *Categorical perception,* 89–112. Cambridge, UK: Cambridge University Press.

Schacter, Daniel L., Barbara A. Church and Elisa Bolton
1995 Implicit memory in amnesic patients: Impairment of voice-specific priming. *Psychological Science* 6: 20–25.

Shannon, Robert V., Fan-Gang Zeng, Vivek Kamath, John Wygonski and Michael Ekelid
1995 Speech perception with primarily temporal cues. *Science* 270: 303–304.

Strange, Winifred (ed.)
1995 *Speech perception and linguistic experience: Issues in cross-language research.* Baltimore: York Press.

Tremblay, Kelly, Nina Kraus, Thomas D. Carrell and Therese McGee
1997 Central auditory system plasticity: Generalisation to novel stimuli following listening training. *Journal of the Acoustical Society of America* 102: 3762–3773.

van Alphen, Petra M. and James M. McQueen
2006 The effect of voice onset time differences on lexical access in Dutch. *Journal of Experimental Psychology: Human Perception and Performance* 32: 178–196.

Weber, Andrea and Anne Cutler
2004 Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language* 50: 1–25.