Coping with Speaker-related Variation Via Abstract Phonemic Categories

The ability to adjust rapidly to new speakers is an important component of language competence. Listeners adjust the bounds of phonemic categories after only very limited exposure to deviant realisation of a given phoneme. Lexical information allows the intended phoneme to be identified, and the adjustment is in consequence applied to the phonemic category as a whole.

Norris, McQueen and Cutler (2003) developed a two-phase paradigm for demonstrating this adjustment. In the first (training) phase of the experiment, listeners performed auditory lexical decision (i.e., decided whether spoken forms – say, *brand, flonken, parallel, milp* – were real words of their language). Twenty items contained a phoneme which was ambiguous ([f/s], halfway between /f/ and /s/). Some listeners heard [f/s] replacing /s/ in words such as *muis* or *paradijs* ('mouse, paradise'; the experiment was conducted in Dutch); others heard it replacing /f/ in words like *rif, biograaf* ('reef, biographer'); others heard it in forms which would be nonwords with either /f/ or /s/. In this phase, listeners overwhelmingly responded yes to the words which were slightly mispronounced in this way. In the second (test) phase, listeners performed phonetic categorisation on a /s/-/f/ continuum. Compared with untrained listeners, listeners who had heard [f/s] as /f/ judged more of the continuum as /f/, and listeners who had heard [f/s] as /s/ judged more as /s/. Categorisation of the whole continuum shifted, not just responses to the single midpoint token heard in the first part. Only listeners who could interpret the sound via lexical information showed this shift – there was no learning effect for the listeners who heard the sound in nonwords.

Norris et al. argued that this rapid adjustment allowed listeners to adapt speech perception to cross-speaker variation. As would be expected in this case, the adjustment proved to be speaker-specific (Eisner & McQueen, 2005), and indeed to vary with the degree to which types of phoneme encode speaker-specific information (Kraljic & Samuel, 2005). Another prediction is that it should be stable across time - the next time we hear the same talker, we should be able to apply the learned adjustment. We here report, first, a test of this issue. Thirty-six Dutch participants heard a 644-word story in which every /f/ or every /s/ (78 in each case) had been changed to the ambiguous [f/s]. Twelve hours after this they performed /s/-/f/ categorisation. For half the subjects, the 12 hours spanned a night in which they slept, for the other half, the 12 hours were daytime. Both groups showed adjustment of phonetic categorisation towards whichever training they had heard 12 hours before (/s/-, /f/-biased). Thus the speaker-specific adaptation is, as predicted, stable across time.

Further, the usefulness of speaker-specific adaptation depends on its generalisability across words (beyond speeded recognition of words heard in training, e.g., *biograaf*). We also report a test of this generalisability. A lexical decision training phase identical to that described above preceded a cross-modal priming experiment in which the critical prime items were minimal pairs of words differing only in final /s/ versus /f/, e.g., *doof-doos* ('deaf, box'). In cross-modal priming, lexical decisions are made to visually presented words preceded by spoken prime words. If the spoken and the visual word are the same word, responses are faster ("priming") than if the words differ. We measured responses to visual presentation of DOOF or DOOS preceded by a control prime or by spoken *doo*[f/s] with the ambiguous [f/s] heard in the training phase. We predicted that if hearing the ambiguous sound in *biograa*[f/s] or *paradij*[f/s] generalises across the lexicon, then doo[f/s] should be heard as *doof* after training on *biograa*[f/s] etc. (giving priming for *doo*[f/s] - DOOF only), and as *doos* after training on *paradij*[f/s] etc. (giving priming for *doo*[f/s] - DOOS only). Exactly this pattern of results was observed. Thus an initial training based on 20 words generalises to all words in the lexicon containing the affected phoneme.

This speaker-specific learning cannot, therefore, be based on word-level traces, but must involve adjustment of abstract phonemic representations for the generalisation to occur. Some word recognition models (e.g., Goldinger, 1998) involve episodic word-level traces and no abstract phonemic representation. To check whether such a model could account for our data, we performed simulations in which the model was trained on one or other of two sets of 20 ambiguous versions of words (each 400-element vectors) in the model's 500-word lexicon. The ambiguity was midway between two possible endings (on analogy to the midpoint of an /s/-/f/ continuum) but it was applied to 20 words for which the lexicon contained only one ending (cf. *biograa-*, *paradij-*). After training, the model was given ambiguous forms for which both possible endings existed in the lexicon (cf. *doo*[f/s]), and the content of the resulting echo trace was determined. If the model had learned from training, then the ambiguous forms should be interpreted as the forms ending with the training-consistent phoneme (cf. *doof*, *doos*), i.e., the echo content should be more like that of the corresponding forms in the lexicon. If the model learned nothing, echo content should match both forms equally well. The latter result was found (and persisted even with ten times as much training). Thus a model with no abstract phonemic representations cannot account for the way in which listeners cope with speaker-related variation in speech by rapidly adjusting phonemic categories. [*Theme*: Variation, phonetic detail and phonological modeling. *Presentation preference*: oral only]