THE EFFECT OF TALKER FAMILIARITY ON WORD SEGMENTATION IN NOISE

Rachel Smith

Department of English Language, University of Glasgow, 12 University Gardens,
Glasgow G12 8QQ, UK.

R.Smith@englang.arts.gla.ac.uk

ABSTRACT

Perceptual learning about voices is known to facilitate speech perception, but it is unclear exactly which phonetic representations are altered to cause this facilitation. This study examines perceptual learning for a non-segmental phonetic property, talker-specific cues to word boundaries. An experiment tested intelligibility in noise of sentences that contained hard-to-segment sequences (e.g. /patsɔːd/, which can correspond to *Pat sawed* or *Pat's awed*). Testing occurred before and after training with a voice; improvement in performance after training was measured. Subjects who heard the same voice during training as during testing showed more improvement than those who heard a different voice. Implications for exemplar theories of speech perception are discussed.

Keywords: perceptual learning, exemplar theories, allophonic detail, familiarity, word segmentation

1. INTRODUCTION

Experiments show that perceptual learning occurs for speech and affects performance in many tasks. For example, speech in familiar voices is recalled better and understood better in noise [2, 9]; phonetic category boundaries can shift in a direction appropriate to a voice or accent, after exposure or even mere suggestion [1, 8, 5]; and in shadowing, phonetic characteristics of the shadower's speech are influenced by those of the talker being shadowed [3, 12].

Recent work seeks to use perceptual learning to understand the units of speech perception and storage. Non-analytic approaches [e.g. 3] propose that perceptual learning reflects memory for holistic episodes or exemplars, which may not need to be broken down into smaller abstract units. Others [8] suggest that perceptual learning operates over abstract phonemic representations. A third, non-segmental, approach proposes that the units over which perceptual learning operates are sensitive to syllabic, prosodic and/or grammatical structure [4, 11]. In this approach, both abstract structure and exemplars play a role, but the abstract structure is richer and more linguistically complex than that assumed by [8].

Smith [13] tested the non-segmental approach in a word-monitoring experiment that investigated perceptual learning for talker-specific patterns of allophonic detail at word boundaries. The experiment manipulated talker familiarity (established in a training session) and allophonic match or mismatch at word boundaries (created by cross-splicing). It was predicted that if perceptual learning about a voice is sensitive to position in syllable, allophonic mismatch should disrupt word-monitoring more in a familiar than unfamiliar voice. Though this pattern of results was found numerically, it was not statistically robust.

Several factors might explain why the experiment in [13] was inconclusive. It used many voices (2 familiar, 8 unfamiliar), which may have distracted subjects from applying their knowledge of the familiar voices (cf [2]). The training used intact speech but the test included cross-spliced speech, some of it allophonically inconsistent; the inconsistency may have led subjects to ignore any knowledge of talker-specific allophonic detail that they had acquired in training, as such knowledge was not always useful. The word-monitoring task required fast responses, but voice-specific properties may play a larger role in tasks where responses are slow [6]. Finally, subjects had the same accent as the familiar talkers, which perhaps meant they needed to learn little about the talkers.

The present study investigates learning of talker-specific word boundary cues in a simpler experiment, testing speech intelligibility in noise. It uses the same materials as [13], but fewer talkers, only unspliced speech, and an off-line, natural task. (It uses the same accent as [13], but ongoing work varies accent.) Talker familiarity helps word identification in noise [9] while correct fine phonetic detail improves intelligibility of synthetic speech in noise [10]. Therefore, familiarity with a talker's allophonic patterns at word boundaries is also expected to improve segmentation in noise.

2. METHOD

2.1. Overview and Design

The experiment consisted of a pre-test, training phase and post-test. In the pre- and post-test, subjects transcribed sentences in background noise; in the training phase they heard the same sentences without noise, and answered questions about their meaning. Improvement in performance between pre-test and post-test was measured.

The experiment manipulated two crossed factors, Test Voice (speaker PF or MJ) and Training Voice (Same or Different to the Test Voice). For any given subject, the same Test Voice was heard in the pre- and post-test.

2.2. Materials

Materials were 48 sentences: 24 pairs of phonemically-identical sentences, matched for prosodic structure, that differed in the placement of a critical word boundary. Two example pairs are:
1) *But* {*Pat sawed* / *Pat's awed*} *them;*
2) *They also offer* {*Mick stability* / *mixed ability*}.
Details are in [13]. This study used recordings of the sentences by two male speakers of Standard Southern British English (SSBE), PF, aged 53, and MJ, aged 27. Each talker read each sentence 8 times in a disambiguating context (e.g. *The fallen trees had blocked access. But Pat sawed them*).

Stimuli for the pre- and post-test were made from tokens of the 48 sentences, without their disambiguating context. For each talker, one token of each sentence was selected at random and mixed with randomly-varying cafeteria noise at an average signal-to-noise (S/N) ratio of +2 dB (average amplitude of sentence: average amplitude of noise). The noise had a gradual onset and offset, increasing from zero to its maximum amplitude over 5 s before the sentence began, then continuing for 15 s before decreasing to zero over 5 s.

Stimuli for the training session were tokens of the same 48 phonemically-identical sentences, in their disambiguating context. For each talker, six different tokens of each context+sentence were selected. Though the sentences were the same, the specific tokens used to make the pre- and post-test stimuli were never heard in training.

In each phase (pre-test, training and post-test), sentences were presented in quasi-random order with the constraint that members of a sentence pair did not appear adjacent to one another.

2.3. Subjects

Subjects were 68 SSBE speakers (17 in each condition), aged between 18 and 35, all students or staff of the Universities of Cambridge or Glasgow. Subjects tested in Glasgow had lived there for less than a year prior to the experiment.

2.4. Procedure

63 subjects were tested at the University of Cambridge in a sound-treated room, and 5 at the University of Glasgow in a quiet room. Subjects were tested individually using a PC laptop running DMDX, and high-quality Sennheiser headphones.

Subjects did a practice item, then the pre-test (48 items, 25 minutes), where their task was to type in all the words they had understood. They then did two practice items followed by the training session (288 items, 40 minutes). Their task during training was to answer questions about the events described in the sentences. All questions had the form *Does the event involve X?* where X was a person, object, idea etc. Subjects responded *Likely* or *Unlikely* by pressing buttons. Finally, they did a further practice item, then the post-test (48 items, 25 minutes), where the task was as in the pre-test.

3. RESULTS

3.1. Pre-processing

For each subject, the percentage of words correctly reported (Words) was scored for each of pre- and post-test. To be scored as correct, words had to be in the same order as in the actual sentence. Obvious mis-spellings and homophones were scored correct, morphological variants as incorrect.

The number of words correct does not necessarily tell us much about listeners' use of phonetic detail to segment words. Familiarity with a voice might allow subjects to identify more words, but might not help them to segment difficult sequences. Therefore, two other measures were used (Word1End and Word2Start), based on the percentage of syllable constituents at the critical word boundary correctly reported. For each sentence, the syllable constituents (onset, nucleus or coda) immediately on either side of the critical word boundary were scored as correct if the subject had reported the correct segment(s) in the correct position relative to the intended word boundary. It did not matter whether the rest of the word was correct. For example, for *They also offer* *Mick stability* the correct Word1End was the coda /k/ followed by a word boundary (#); the correct Word2Start was # followed by the onset /st/. (Thus, the response *He also had neck sterility* is correct for both Word1End and Word2Start, whereas *We all suffer Mick's ability* is wrong on both measures.)

For each subject, an Improvement score was calculated for each measure (Words, Word1End and Word2Start), by subtracting percentage correct responses in the pre-test from percentage correct in post-test. The Improvement scores were submitted to 3 between-subjects ANOVAs, one on each variable, each with two factors, Training Voice (Same vs. Different) and Test Voice (MJ vs. PF).

3.2. Main results

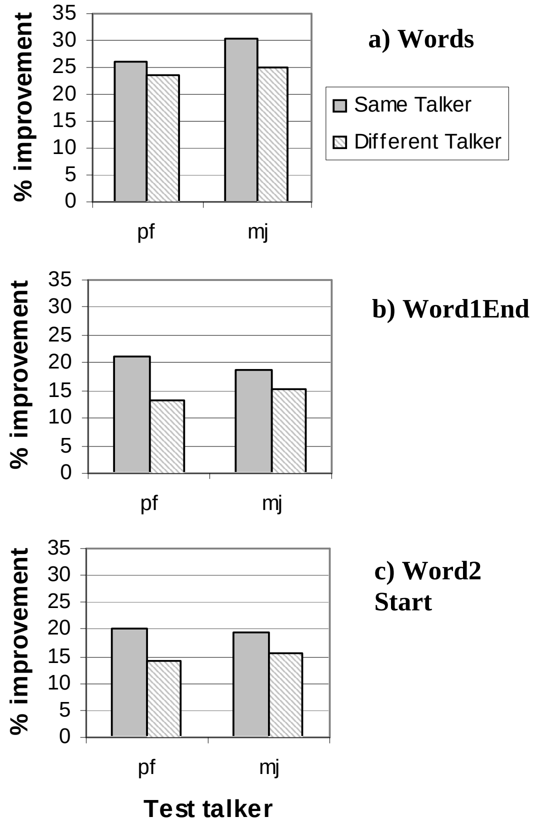
Figure 1a shows the amount of improvement from pre- to post-test in percentage of Words correctly reported. Figures 1b and 1c show improvement in percentage of correct syllable constituents reported at Word1End and Word2Start respectively.

Training with *any* voice (Same or Different as tests) improved performance by ~20-30% for Words and ~10-20% for syllable constituents. This large improvement is presumably because training gave all subjects useful experience of the test sentences: they heard both members of each sentence pair, each in a meaningful context.

Crucially, for all three measures, subjects who heard the Same voice at training as in the tests improved significantly more than those who heard a Different voice (Words: *F* (1, 64) = 4.8, $p < 0.05$, partial eta^2 = 0.07*;* Word1End: *F* (1, 64) = 6.0, $p < 0.025$, partial eta^2 = 0.09, Word2Start, *F* (1, 64) = 4.6, $p < 0.05$, partial eta^2 = 0.07). The Same-Voice advantage was 4% for Words, 6% for syllable constituents at Word1End, and 5% for Word2Start.

The amount of improvement was not affected by whether the test talker was MJ or PF ($p > 0.1$ for all variables). Nor did the size of the difference between Same and Different voice training depend on the test talker (n/s interaction between Training Voice and Test Voice, $p > 0.3$ for all three measures).

Figure 1: Improvement from pre- to post-test in correct reporting of: **a)** Words, **b)** syllable constituents at Word1End, **c)** syllable constituents at Word2Start.



4. DISCUSSION

40 minutes' exposure to sentences in a talker's voice leads to more improvement in understanding novel tokens of those sentences in background noise than does exposure to the identical sentences, in a different voice. This Same-Voice advantage is significant not only for identifying words, but also for segmenting difficult, phonemically-ambiguous sequences, as reflected in identification of syllable constituents adjacent to word boundaries. The results for word identification parallel those of [9] (though the present experiment used a much shorter training period), while those for segmentation are novel.

Could the improvement in segmentation have arisen merely because familiarization with a voice makes listeners better at segregating that voice from background noise [cf. 7]? If so, the results might have no interesting implications for the units involved in perceptual learning; they might just reflect the fact that linguistic processing is easier when a voice is more easily separated from noise. This possibility seems unlikely, however. The kinds of properties that may help segregate a voice from noise include pitch, intonation, rhythm, and amplitude envelope [7]. As the fine detail of these properties can cue word boundaries, it is unlikely that familiarity with them would contribute to segregation, but not at all to word segmentation. One way to distinguish a segregation-based account of the data from the perceptual learning account favoured here is to use as the pre-test and post-test task a forced-choice segmentation task, without background noise. If more improvement is again found after training with the same than a different talker, this will support the view that perceptual learning involves modifying linguistic-phonetic representations.

What are the implications of the results for the nature of perceptual learning? They must reflect some degree of abstraction over experience: test and training never used identical *tokens* of the sentences. However, the perceptual learning listeners carried out cannot have involved units as abstract as phonemes. Learning only about how a talker realizes particular phonemes would not have improved listeners' ability to locate word boundaries in phonemically-ambiguous sequences such as /patsɔːd/. It would be far-fetched to assume that altering phonemic representations (which are by definition insensitive to position in syllable) should help in the assignment of sounds to the correct position in syllables and words. Since the listeners did improve in their identification of syllable constituents on both sides of word boundaries, the natural conclusion is that the perceptual learning that took place was sensitive to position in syllable.

The data are broadly consistent with non-analytic approaches to speech perception [3], in which learning involves storing exemplars of words (or even larger units) heard. They are also consistent with approaches like [4, 11], in which phonetic representations are richly prosodically-structured and context-sensitive, and phonemes need not play an important role. Inspection of the patterns of errors offers some support for the prosodically-structured view over a purely non-analytic view. In the post-test, subjects sometimes correctly identified a syllable constituent, while failing to correctly identify the word or word sequence that contained it, suggesting that they were not relying exclusively on lexical exemplars.

To conclude, familiarity with a voice helps segmentation of difficult phonemically-ambiguous sequences in poor listening conditions. This finding is inconsistent with accounts of perceptual learning that are based solely on phonemic representations [8]. To distinguish between the other two major accounts of perceptual learning for speech, the non-analytic [3] and prosodically-structured [4, 11] views, will be best done by testing how well learning generalizes to novel sentences that have similar syllabic and prosodic structures. For the present, it is noted that the prosodically-structured view most naturally accommodates the data. It accounts not only for how information about systematic allophonic detail at word boundaries is used perceptually, but also for why such patterns exist in speech production in the first place.

5. REFERENCES

[1] Allen, J.S., Miller, J.L. 2004. Listener sensitivity to individual talker differences in voice-onset-time. *J. Acoust. Soc. Am.* 116, 3171-3183.

[2] Goldinger, S.D., Pisoni, D.B., Logan, J.S. 1991. On the nature of talker variability effects on recall of spoken word lists. *J. Exp. Psychol.: LMC* 17, 152-162.

[3] Goldinger, S.D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psych. Rev.* 105, 251-279.

[4] Hawkins, S., Smith, R.H. 2001. Polysp: a polysystemic, phonetically-rich approach to speech understanding. *Ital. J. of Linguistics-Rivista di Linguistica* 13, 99-188.

[5] Hay, J., Nolan, A., Drager, K. 2006. From *fush* to *feesh*: exemplar priming in speech perception. *The Linguistic Review* 23, 351-379.

[6] McLennan, C.T., Luce, P.A. 2005. Examining the time course of indexical specificity effects in spoken word recognition. *J. Exp. Psychol.: LMC* 31, 306-321.

[7] Newman, R.S., Evers, S. 2007. The effect of talker familiarity on stream segregation. *J.Phon.* 35: 85-103.

[8] Norris, D., McQueen, J.M., Cutler, A. 2003. Perceptual learning in speech. *Cognitive Psychology* 47, 204-238.

[9] Nygaard, L.C, Pisoni, D.B. 1998. Talker-specific learning in speech perception. *Percept. Psychophys.* 60, 355-376.

[10] Ogden, R., Hawkins, S., House, J., Huckvale, M., Local, J., Carter, P, Dankovičová, J., Heid, S., 2000. ProSynth: an integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Comp. Sp. & Lang.* 14: 177-210.

[11] Pierrehumbert, J. 2002. Word-specific phonetics. In: Gussenhoven, C., Warner, N. (eds), *Laboratory Phonology 7*. Berlin: Mouton de Gruyter, 101-139.

[12] Shockley, K., Sabadini, L., Fowler, C.A. 2004. Imitation in shadowing words. *Percept. Psychophys.* 66, 422-429.

[13] Smith, R. The role of fine phonetic detail in word segmentation. Unpublished Ph.D. dissertation, University of Cambridge.

ACKNOWLEDGEMENTS

Financial assistance was provided by the John Robertson Bequest Fund of the University of Glasgow.