Influence of Talker-Specific Phonetic Detail on Word Segmentation

Rachel Smith
University of Cambridge, UK
E-mail: rhs20@cam.ac.uk

ABSTRACT

Experiments show that individual talker characteristics affect performance in many language tasks, but most research has focused on isolated words. The present work investigated whether memory for talkers is also drawn upon in word segmentation. In a training phase, listeners were familiarized with 2 talkers' productions of pairs of phonemically-identical sentences. Training was followed by a word-monitoring test using novel tokens of the training sentences, which were spoken by either a familiar or an unfamiliar talker, and were cross-spliced to consist of allophonically matched or mismatched halves. It was predicted that when the talker was familiar, allophonic match would facilitate word-monitoring and mismatch disrupt it, but that allophonic match/mismatch would have less or no effect when the talker was unfamiliar. This pattern was found for reaction times in one experimental condition, suggesting that episodic memory for talkers critically includes linguistically-relevant phonetic detail.

1. INTRODUCTION

Experiments show that individual talker characteristics can affect speech processing in many tasks (e.g. serial recall, explicit recognition memory, word recognition in noise, shadowing, spontaneous imitation; [1,2]). The 'episodic' approach of Goldinger, Pisoni and colleagues holds that talker-specific information is remembered along with other details of perceptual episodes, and there is no obligatory unit into which the signal must be decomposed before meaning can be understood. Episodic memory could provide a framework for exploiting systematic variation in fine phonetic detail that is relevant to linguistic distinctions, e.g. [3]. But experiments have not manipulated linguistic-phonetic differences among talkers, so the data may reflect just a richer non-linguistic contribution: that is, the more a given task resembles an earlier task, the better subjects perform in the later one.

Talkers sound different from one another for many reasons. Some factors are arguably linguistically-irrelevant (e.g. vocal tract size, voice quality). Others are undoubtedly linguistically relevant inasmuch as they signal allophonic distinctions (cf. [4,5,6]). I aimed to investigate whether individual talker characteristics affect linguistic processing when the task critically depends on appropriate use of linguistically-relevant fine phonetic detail. To do this, one must manipulate particular types of fine phonetic detail: it must be detail that differentiates between linguistic meanings and that can differ somewhat between talkers, and it must be detail that is not retained in a linguistic description whose basic units are phonological feature-bundles or phonemes.

In general, a token of a previously-heard word is likely to be recognized better if spoken by a familiar talker than an unfamiliar talker. The hypothesis was that this generalisation would not apply if allophonic detail is inconsistent with what the familiar talker normally produces. Instead, mismatch in the allophonic detail around a token of a previously-heard word should disrupt recognition *more* if the talker's voice is familiar than if it is unfamiliar, because the mismatch will violate a more precise expectation on the part of the listener. Figure 1 is a schematic illustration of this idea: When the talker is familiar, allophonic match is expected to be facilitative, and allophonic mismatch is expected to be disruptive. When the talker is unfamiliar, allophonic match/mismatch is expected to have less or no effect.

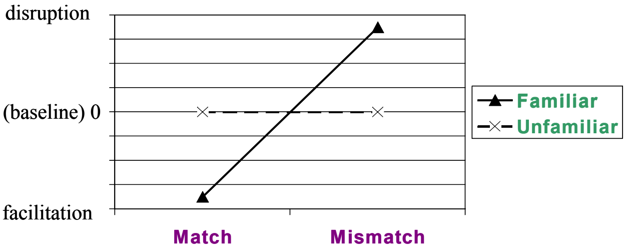

Figure 1. Experimental predictions for allophonic match *vs.* mismatch in a familiar *vs.* unfamiliar voice. Ordinate shows performance on a relevant task.

2. METHOD

The experiment involved a training period where subjects were familiarized with 2 talkers' tokens of pairs of phonemically-identical sentences (e.g. *So he diced them / So he'd iced them*) presented in a disambiguating context. Training was followed by a word-monitoring test, where subjects heard novel tokens of the training sentences without their context, and monitored for word targets (e.g. *iced* or *diced*). Test sentences were spoken either by a *familiar* or an *unfamiliar* talker, and were cross-spliced to consist of allophonically *matched* or *mismatched* halves.

2.1. Materials Core materials were those that listeners heard both in training and in the word-monitoring test. They were 24 pairs of sentences that were phonemically identical but differed (usually grammatically) in a critical part (underlined), e.g. *So <u>he diced</u> them* vs. *So <u>he'd iced</u>*

ISBN 1-876346-48-5 © 2003 UAB

them, or *They also offer* Mick stability vs. *They also offer* mixed ability. The prosodic pattern did not differ between the two members of a pair. The word-monitoring target was always the last word of the critical phrase (e.g. *diced* or *iced* in the first example pair).

For training, a disambiguating precursor sentence was made for each core sentence, e.g. a) *He wanted the carrots to cook fast. So he diced them;* b) *The top of the cakes had come out looking uneven. So he'd iced them.*

For word-monitoring, fillers were made to accompany the core sentences. Each trial consisted of one core sentence, and 1-6 fillers. The variable number of fillers meant that the core sentence could always be last in the trial, without listeners being able to predict which sentence the target would be in. Because the core sentences were heard in both training and test, it was also important to ensure that listeners could not predict the location of the target just by recognizing the sentence from training. Fillers therefore resembled the core sentences in systematic ways, e.g. they began in the same way (*So he dialed from a payphone; So he'd iron your clothes if you asked him to*), or rhymed with it (*You denied having enticed them*).

2.2. Talkers and Recording Familiar talkers were heard both in training and word-monitoring. Unfamiliar talkers were heard only in word-monitoring. All had similar accents (SSBE) but differed in age and gender. Talkers who were well differentiated on these dimensions were used to prevent a voice's familiarity being confounded with its similarity to other voices in the experiment. The 2 Familiar talkers (1 male, 1 female) were aged 27 and 25. The 4 Unfamiliar talkers (2 male, 2 female) were aged 35-54. Talkers read from a randomized list of sentences, as naturally as possible in an informal style.

2.3. Assignment of sentences to talkers 8 tokens of each precursor + core sentence were recorded from each Familiar and Unfamiliar talker. For training, the male Familiar talker's tokens were used for half the sentences, and the female Familiar talker's for the other half. 6 tokens of each sentence were chosen for use in training; the other 2 were used to make spliced stimuli (Section 2.4). For word-monitoring, each of the 4 Unfamiliar talkers contributed 3 sentences (12 altogether). Each Familiar talker provided 6 sentences (12 altogether). Talker familiarity was not varied *within* a sentence, i.e. all subjects heard the same talker's tokens of a given sentence (though different groups of subjects heard different spliced versions: Section 2.6.2).

Fillers for word-monitoring trials were assigned to talkers so that the odds of hearing a target word *vs.* not hearing a target word in a given voice were the same for all talkers. (Otherwise, listeners might have used strategies to predict the location of targets, e.g. that sentences in MJ's voice were likely to contain a target.)

2.4. Splicing Four versions of each sentence were created, by cross-splicing just before the start of the target word. Splicing was always to a token by the same talker. The segments of the target word never included spliced material, as the splice point always immediately preceded the target word. Table 1 illustrates the splicing principles for the sentence pair *So he diced them / So he'd iced them*. When the target was *diced* (Early Splice), an allophonic match was created by splicing two tokens of the same sentence, *So he* from one token and *diced them* from another. An allophonic mismatch was created by splicing *So he* from a token of *So he'd iced them,* to *diced them* from *So he diced them.* Likewise, when the target was *iced* (Late Splice), an allophonic match was created by splicing *So he'd* from one token of *So he'd iced them* to *iced them* from another; a mismatch involved splicing *So he d* from *So he diced them* to *iced them* from *So he'd iced them.*

	Early Splice: target *diced*	Late Splice: target *iced*
✓	**[so he]**$_{\text{diced}}$ **[diced them]**$_{\text{diced}}$	[so he'd]$_{\text{iced}}$ [iced them]$_{\text{iced}}$
✗	[so he]$_{\text{iced}}$ **[diced them]**$_{\text{diced}}$	**[so he d]**$_{\text{diced}}$ [iced them]$_{\text{iced}}$

Table 1. Splicing principles.
✓ denotes allophonic match, and ✗ denotes allophonic mismatch.
Bold type / subscript $_{\text{diced}}$: speech from a token containing *diced.*
Normal type / subscript $_{\text{iced}}$: speech from a token containing *iced.*

Obstruent-vowel and vowel-obstruent splice points used standard criteria. Obstruent-obstruent splice points coincided with regions of greatest perceptual change between segments. Tokens were chosen at random for splicing from the 8 recorded per talker. They were spliced automatically to create sets of 4 versions as in Table 1, and were evaluated by listening. If any version in a set did not sound acceptable due to large pitch, amplitude or formant discontinuities, the set was discarded and a new pair of tokens randomly chosen and spliced, until an acceptable set of stimuli had been obtained for all sentences.

2.5. Subjects The 40 naïve subjects (19 male), all monolingual British English speakers, were aged 18-29. None of the subjects knew personally any of the talkers.

2.6. Procedure Subjects were tested individually in an IAC booth. They received 90 minutes of training before the 20-minute word-monitoring test. A PC controlled stimulus presentation via a Tucker-Davis DD1 system, which recorded button presses and reaction times.

2.6.1. Training Subjects heard (over high-quality headphones) 8 practice items, then a list containing 2 repetitions of 6 tokens of the 48 2-sentence utterances, in pseudo-random order (576 utterances total). Immediately after each utterance, a comprehension question of the form *Does the event involve X?* appeared on the computer screen for 3 sec. Subjects were told that they would hear descriptions of situations or events, and should judge whether it was *likely* or *unlikely* that an event involved a given object, idea etc. They responded by pressing one of two labelled buttons on a response box. Brief breaks were after every 20 items, with a longer break after 45 minutes.

2.6.2. Word-monitoring Subjects heard 3 practice trials, and then one of 4 lists. Each list contained one version (Early-Splice Match, Early-Splice Mismatch, Late-Splice Match, or Late-Splice Mismatch) of each experimental sentence, with equal numbers of items in each condition. The order of trials, and of fillers within trials, was identical in the 4 lists. The target-bearing sentence always appeared last in the trial. Subjects were told that on each trial they would see a target word and should listen for that word in a series of sentences, pressing a button as fast as possible if they heard it. At the start of each trial, the target word (e.g. *iced*) appeared in orthographic representation in the centre of the computer screen for 1 sec before auditory presentation of the first sentence began. It remained visible for the duration of the trial. The ISI between sentences was 2 sec.

3. RESULTS AND DISCUSSION

Reaction times (RTs) were measured from target word onset, and were log-transformed. Missing responses to target-bearing sentences were treated as errors. Errors were analysed using logistic regression. RTs were submitted to 2 repeated-measures ANOVAs (Allophone x Familiarity x Splice Point), one with Subjects and the other with Sentences as the repeated factor. For errors and RTs, there were unpredicted interactions of the splice point location (Early/Late) with one or both of allophone and familiarity. Separate analyses (Allophone x Familiarity) were therefore carried out for Early and Late Splice conditions respectively; these are reported below.

Figure 2a shows that when the splice was Late (target e.g. *iced*), responses were more accurate if the talker's allophones matched than if they mismatched (χ^2 (1) = 40.59, *p*=0.0001). Talker familiarity improved accuracy too (χ^2 (1) = 20.06, *p*=0.0001). But the predicted interaction between allophone and familiarity was absent (χ^2 (1) = 0.63, *p*<1). In contrast, Figure 2b shows that when the splice was Early (target e.g. *diced*), allophonic match did not affect accuracy (χ^2 (1) = 0.77, *p*<1). Contrary to prediction, responses to familiar talkers were less accurate than to unfamiliar talkers (χ^2 (1) = 9.23, *p*=0.0024), but this may be due to a ceiling effect: logistic regression gives less robust results when event probabilities are very high or low. The predicted interaction between allophone and talker familiarity was absent (χ^2 (1) = 2.08, *p*<1).

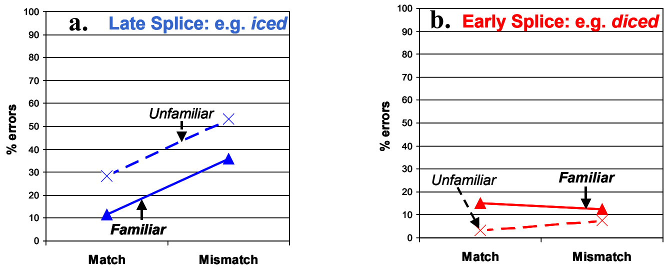

Figure 2. Mean error rates (%), by allophone (Match / Mismatch) and talker (Familiar / Unfamiliar) **a.** where the splice was Late (target e.g. *iced*), **b.** where the splice was Early (target e.g. *diced*).

Figure 3a shows that when the splice was Late (target e.g. *iced*), allophonic match interacted with talker familiarity as predicted (F_1 (1,282) = 4.72, *p*<0.05; F_2 (1,299) = 4.94, *p*<0.05). Planned comparisons showed that RT to familiar talkers was faster than to unfamiliar talkers when the allophones matched (t_1 (282) = 2.78, *p*<0.01; t_2 (299) = 1.71, *p*<0.1), but not when the allophones mismatched (t_1 (282) = 0.54, *p*<1; t_2 (299) = 0.95, *p*<1). For familiar talkers, RT was faster to allophonically matching stimuli than to allophonically mismatching stimuli (t_1 (282) = 2.68, *p*<0.01; t_2 (299) = 2.17, *p*<0.05) but allophonic match did not affect RT to unfamiliar talkers (t_1 (282) = 0.55, *p*<1; t_2 (299) = 1.08, *p*<1). In contrast, when the splice was Early (Figure 3b; target e.g. *diced*), neither allophonic match nor speaker familiarity had any detectable effect on reaction time, nor did they interact.

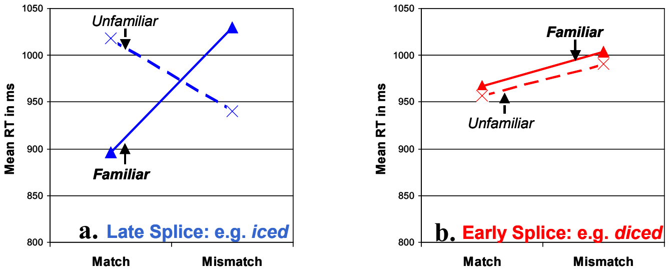

Figure 3. Mean RTs (ms), by allophone (Match / Mismatch) and talker (Familiar / Unfamiliar), **a.** where the splice was Late (target e.g. *iced*), **b.** where the splice was Early (target e.g. *diced*).

To summarise the difference between responses in Early *vs.* Late Splice conditions: When the splice was Early (e.g. [**so he**]$_{\text{diced}}$ [**diced them**]$_{\text{diced}}$ *vs.* [so he]$_{\text{iced}}$ [**diced them**]$_{\text{diced}}$), allophonic match or mismatch around the word boundary did not affect responses. Targets were recognized with high accuracy, and RT was not affected by match/ mismatch. In contrast, when the splice was Late (e.g. [so he'd]$_{\text{iced}}$ [iced them]$_{\text{iced}}$ *vs.* [**so he d**]$_{\text{diced}}$ [iced them]$_{\text{iced}}$) allophonic match/mismatch had a large effect, and interacted with talker familiarity in the predicted direction. The effects of an allophonic mismatch for Late Splice stimuli are as follows: if the talker is unfamiliar, the target (*iced*) is relatively unlikely to be recognized (accuracy below 50%). If the talker is familiar, the target may be identified, but recovery from mismatch takes some time (long RT).

The discrepancy between conditions has no simple explanation in terms of factors that would cause Early Splice stimuli to be better recognised overall. Early Splice targets did not sound systematically longer or louder than Late Splice targets. They were not higher-frequency words, nor were they more likely as parses of the ambiguous phrases according to a probabilistic prosodic grammar [7]. They could have been easier to segment because most of them were obstruent-initial, whereas most Late Splice targets were vowel-initial (*diced* vs. *iced, stability* vs. *ability*, etc), in that a syllable- or word-initial onset consonant might have special status as a perceptual cue [8]; and infants learn to use allophonic cues in segmenting CVC words several months before they can do the same for VC words [9]. But this explanation is insufficient because the discrepancy between conditions was found even for a subgroup of 3 sentences where both Early Splice and Late Splice targets contained initial obstruents (e.g. *That surprise* vs. *That's a prize*).

The best explanation seems to be that Early and Late Splice stimuli differ in where the allophonic mismatch is located relative to word boundaries. For Early Splice mismatching targets, which are recognized with high accuracy (e.g. [so he]$_{\text{iced}}$ **[diced them]**$_{\text{diced}}$) the portion of the signal corresponding to the target itself is phonetically coherent. The phonetic detail provides only one set of cues to the syllable affiliation of the segments of the target (consistently *diced*), and allophonic mismatch affects primarily the identity of the preceding word (*he'd* or *he*). In contrast, for Late Splice mismatching targets (e.g. [**so he d**]$_{\text{diced}}$ iced them$_{\text{iced}}$), considerable evidence has accumulated by the time the splice point is reached that /d/ forms a syllable (or word) onset, but the phonetic quality of the /aɪ/ then suggests that the nucleus is syllable- (or word-) initial too. Thus, the phonetic detail provides conflicting evidence as to the word affiliation of the segments of the target itself (which might belong to *diced,* or *iced*). Thus there is greater competition between multiple word candidates for the segments of the target word in Late Splice than Early Splice stimuli, and this probably underlies the difference between conditions. One way to test this explanation is to investigate whether allophonic mismatch interacts with talker familiarity in the Early Splice condition if the task causes the listener to focus on the word most affected by mismatch (*he'd* / *he*), e.g. using phrase monitoring or presentation in noise.

4. GENERAL DISCUSSION

The results provided partial support for the hypothesis that memory for talker characteristics critically includes linguistic-phonetic fine detail. The predicted interaction between talker familiarity and allophonic coherence was found in the RT data for the Late Splice condition. When a talker was familiar, word-monitoring was *facilitated* when the allophones matched, and *disrupted* when they mismatched. When the talker was unfamiliar, this pattern of facilitation and disruption did not occur. This interaction is difficult to explain without assuming that listeners were bringing to bear knowledge about details of speech that are a) specific to individual talkers, and b) relevant to linguistic distinctions in syllable structure that relate to word boundary location. If talker-specific knowledge were not involved, then the same pattern should have been found for familiar and unfamiliar talkers. If the knowledge were not also linguistically relevant, responses to familiar talkers' tokens should have been facilitated regardless of allophonic match (since both matched and mismatched stimuli had been spliced). Knowledge of these talker-specific allophonic details is acquired fast (1½ hours), and they are not captured by a linguistic description that consists of phonemes, or feature clusters corresponding to phonemes. The implication is that memory for speech encodes talker-specific fine phonetic detail that is relevant to linguistic distinctions at levels beyond the phoneme. Future work should focus on the effects of amount of exposure, and on how well learning transfers to novel sentences.

The results are consistent with the approach of Firthian phonetics, in which systematic variation in phonetic fine detail conveys information about many linguistic levels: 'phonemes', position-in-syllable, word boundaries, function *vs.* content words, morphology, information structure, etc. [10]. Memory for these linguistic structures also includes talker-specific details, which could have predictive and communicative value in perception. The data support the claim of [3] that memory for speech and language is neurally organised such that the same sensory information simultaneously feeds multiple strands in perception, e.g. segmental, prosodic, talker-identity-related and attitudinal.

ACKNOWLEDGEMENTS

Thanks to Sarah Hawkins for her support and guidance in all aspects of this work, and to John Coleman for allowing me to use the probabilistic prosodic grammar. Financial support was provided by the AHRB.

REFERENCES

[1] D.B. Pisoni, "Some thoughts on 'normalization' in speech perception," in *Talker Variability in Speech Processing,* K. Johnson and J.W. Mullennix, Eds., pp. 9-32. San Diego, London: Academic Press, 1997.

[2] S.D. Goldinger, "Echoes of echoes? An episodic theory of lexical access," *Psychological Review,* vol. 105, pp. 251-579, 1998.

[3] S. Hawkins and R. Smith, "Polysp: a polysystemic, phonetically-rich approach to speech understanding," *Italian Journal of Linguistics—Rivista di Linguistica,* vol. 13, pp. 99-188, 2001.

[4] R.E. Remez, J.M. Fellowes, and P.E. Rubin, "Talker identification based on phonetic information", *JEP:HPP,* vol. 23, pp. 651-666, 1997.

[5] S.M. Sheffert, D.B. Pisoni, J.M. Fellowes and R.E. Remez, "Learning to recognize talkers from natural, sinewave and reversed speech samples", *JEP:HPP,* vol. 28, pp. 1447-1469, 2003.

[6] F. Nolan and T. Oh, "Identical twins, different voices", *Forensic Linguistics,* vol. 3, pp. 39-49, 1996.

[7] J. Coleman, "Candidate selection", *The Linguistic Review,* vol. 17, pp. 167-179, 2000.

[8] J.J. Ohala, "Speech perception is hearing sounds, not tongues", *JASA*, vol. 99, pp. 1718-1725, 1996.

[9] S. Mattys and P.W. Jusczyk, "Do infants segment words or recurring contiguous patterns?" *JEP:HPP,* vol. 27, pp. 644-655, 2001.

[10] J.K. Local, "Modelling assimilation in a non-segmental rule-free phonology", in *Papers in Laboratory Phonology II,* G.J. Docherty and D.R. Ladd, Eds., pp. 190-223. Cambridge: CUP, 1992.