WORD INTELLIGIBILITY AND PLACE ASSIMILATION IN SPONTANEOUS SPEECH

C. Sotillo[1], J. McAllister[1], E. G. Bard[1], G. Doherty-Sneddon[2], and A. Newlands[2]

Human Communication Research Centre,
[1]Dept. of Linguistics, Edinburgh University, [2]Dept. of Psychology, Glasgow University

ABSTRACT

Tokens of words involving word final place of articulation assimilation are less intelligible than canonical forms when excerpted from context and presented to a group of listeners. While the process of assimilation is able to explain certain differences in intelligibility, it is clear that there are additional factors, such as the frequency of productive morphological suffixes, which influence the ease with which a word is recognised.

INTRODUCTION

One way in which tokens of the same spoken word vary is the extent to which their acoustic form supports recognition of the word in the absence of the word's natural context. The ease with which subjects are able to recognise an excised word can be taken as a measure of that token's intelligibility. This definition of intelligibility thus represents the bottom-up processes involved in lexical access, reflecting the amount of information that is available in the acoustic signal. The poorer the quality of input, the lower the intelligibility.

Sources of information other than the auditory input also aid a listener in recognising the incoming message. For example, the syntactic and semantic context of an utterance will restrict the set of appropriate lexical choices [1]. Such top-down processes have been shown to have an effect on the duration and intelligibility of word tokens. Speakers adjust their pronunciation of words in running speech to complement the information available to listeners in the remainder of the discourse. So, for example, the more predictable a word is from its sentential context, the less clear the token will be [2].

Having established that such reductions in intelligibility occur, the question arises as to how such differences in clarity might be realised. Well-known phonological processes such as place assimilation are prevalent in running speech [3], and it seems reasonable to assume that the application of these processes may have an effect on intelligibility.

Opinions vary on how assimilation should be modelled in current theories of phonetics/phonology. It is not entirely clear, for example, whether assimilation should be accounted for at the phonetic or phonological level of representation. What cannot be disputed is that the sandhi phenomena of connected speech have implications for theories of lexical access.

This paper addresses the relationship between assimilation and the ability to recognise words, as reflected in measures of intelligibility. The question is whether regular and predictable variation, like that involved in place assimilation, imposes an additional *cost*, or leaves the process of word recognition unaffected.

Cross-modal priming experiments using isolated words [4, 5] show that even very small deviations from the canonical form of a word (i.e. changes of just one phonological feature) will result in a loss of priming. From this, one would predict that assimilatory processes would have an inhibitory rather than facilitatory effect on word recognition. An assimilated token would be harder to recognise when excerpted.

However, recent work on repetition priming of words *within sentences* using place of articulation assimilation found no loss of priming for assimilated words, except when followed by an unviable context [6]. This result could be accounted for by a conservative lexical matching process which, though intolerant of mismatch, only rejects candidates when there is unambiguous mismatching information. In this case, an assimilated token excerpted from context is no more likely to be rejected than an unambiguous token.

We report a study exploring the relation between intelligibility and place of articulation assimilation and consider whether assimilation necessarily involves an increase in processing load. We discuss the implications of the result for theories of lexical access.

METHOD

Materials

Data was selected from the HCRC Map Task Corpus [7]. The 128 unscripted conversations involve pairs of participants who collaborate to replicate on one's schematic map a route drawn on the other's. Task success requires discussion of the various landmarks along the route, the names of which were carefully chosen to provide the appropriate environment for certain phonological reduction processes in English. In particular, certain names involved possible place assimilation of word-final alveolar nasal stop consonants. Both labial (e.g. *caravan park*) and velar (e.g. *Indian country*) contexts were used. After completing their set of map tasks all speakers were required to read carefully a list of landmark names to provide clear 'citation' forms against which other tokens could be compared. Materials were recorded on DAT (Sony DTC1000ES) using one Shure SM10A close-talking microphone and one DAT channel per participant.

Single word tokens from fluent first and second mentions of landmark names were then used in a series of intelligibility experiments to explore the effects on intelligibility of information status within dialogue [8]. This set of studies established an intelligibility score (in terms of percent correct recognition) for each word token when it was excerpted from context and presented in noise to a panel of listeners from the same language community as the Corpus participants.

We selected from this pool those words relating to landmark names involving possible place assimilation of word-final nasals. There were 21 usable examples of nasals preceding labial stops (e.g. *telephone* box), and 13 usable examples of nasals preceding velar stops (e.g. *lemon* grove). For each of these landmarks there were two running speech tokens: the first, introductory, mention and the second mention. In some cases the second mention was by the same speaker who introduced the word, in other cases the second mention was by a different speaker. Each running speech token had a corresponding citation form against which it could be compared. This gave us a total of 68 running speech/citation form pairs.

Procedure

Each utterance containing a required token was digitised at a rate of 16KHz using the Entropic Signal Processing System through the XWAVES speech analysis program on a Sparc station. The start and end points of each word were located using a combination of audio and visual information provided by time-amplitude waveforms and broad band spectrograms. Cuts were made at zero-crossings.

Excerpted tokens were used for two kinds of studies. For the experiments on intelligibility tokens were overlaid with noise by multiplying, sample by sample, the original speech file by a 16KHz file of random noise (where all sample values were in the range 0.5 to 1.5). For each resulting stimulus the amplitude was related to that of the original speech data file, and the data points had the same sign as the original data values they replaced. The tokens presented to subjects in the perceptual task were not masked by noise.

Intelligibility Task

In each experiment, either 48 or 60 word types were used, with four or three tokens per type depending on the experimental design. Tokens were allocated to different presentation tapes according to Latin square designs and played to groups of listeners. Only one token of any word type was presented to each subject, and each token was heard by at least 9 subjects. Subjects were asked to write down the identity of each word they heard. For the subset of word types used in the perceptual task, mean scores for correct recognition were calculated for all tokens which appeared in more than one intelligibility experiment.

Perceptual Task

The unscripted nature of our running speech material, the use of non-linguistically trained subjects, and the ratio of tokens to speakers rendered a detailed acoustic analysis (for example in terms of pole/zero decomposition [9]) impossible.

We opted therefore to explore the perceptual evidence for assimilation by presenting tokens to a group of nine phoneticians who were asked to make a set of judgements about the place of articulation of each word-final nasal consonant. Experts rated each nasal on three scales: labial, alveolar, and velar. A rating of 0 indicated that no evidence was perceived to suggest the consonant was produced at this place, while a rating of 5 indicated that the perceptual evidence was fully consistent with an articulation at this place. The three options were not mutually exclusive, so that in principle it was possible to assign a rating of 5 on more than one scale for any one token.

RESULTS

Scores for correct recognition for this subset of intelligibility data were submitted to an ANOVA by materials (by subjects analysis was not possible since the items were gathered from a series of different experiments). Raw intelligibility scores for citation forms and spontaneous mentions can be seen in Table 1.

Citation forms are significantly more intelligible than their corresponding running speech tokens [*Form:* F_2 (1,31) = 33.74, $p < 0.0001$].

In addition, an ANOVA run on *loss of intelligibility*, that is, the difference in rate of correct identifications between citation forms and running speech tokens, revealed a significant effect of mention, with greater loss of intelligibility for second mentions [*Mention:* F_2 (1,31) = 7.27, $p = 0.01$].

Thus the repetition effects on intelligibility reported elsewhere [10, 8], hold for this subset of data.

Table 1. Intelligibility of citation and running speech tokens for introductory and repeated mentions

Form	Mention	
	First	Second
Citation	.70	.76
Running speech	.48	.41

When the experts' overall judgements of assimilation were examined, just over one third of all responses indicated no assimilation had taken place (35.2%); nearly one fifth of all tokens involved a clear assimilation (17.85%), while the remainder involved percepts of an alveolar with a a varying degree of labial or velar quality.

ANOVAs on experts' mean place judgements showed strong Form effects with citation forms being judged as significantly more [n]-like and less [m]- or [ŋ]-like than corresponding running speech tokens [*[n]:* $F(1,32) = 20.09$, $p < 0.0001$; *[ŋ]:* $F(1,32) = 8.79$, $p < 0.01$; *[m]:* $F(1,32) = 3.62$, $p = 0.066$].

The difference in [ŋ]-ness judgement between running speech tokens and citation forms was greater for second mentions than for first mentions, with second mentions being perceived as more assimilated. [*Form X Mention:* $F(1,32) = 4.22$, $p < 0.05$]. This was true regardless of following context, though [ŋ]-ness judgements preceding labials were significantly lower than those preceding velars [*Place:* $F(1,32) = 15.14$, $p < 0.0005$].

No effect of mention was found for [m]- or [n]-ness judgements of tokens preceding either labials or velars. [*Form X Mention for [n]:* $F(1,32) < 1$, n.s.; *for [m]:* $F(1,32) = 1.33$, n.s.].

It appears that we have evidence to suggest assimilation was indeed taking place, and we also have an intelligibility effect to explain. What, then, is the relation between the two?

Assimilation and intelligibility

A series of correlations showed that although judged assimilation was related to intelligibility it did not account for all of the intelligibility differences in the data.

Significant correlations between intelligibility and place judgements were found only for words preceding velar stops: the more [ŋ]-like (i.e., assimilated) were less intelligible [$r = -.409$, $p < 0.005$], the more [n]-like (un-assimilated) more intelligible [$r = .491$, $p < 0.001$]. For words preceding labial stops, however, analogous correlations were not significant [$r = -.105$, n.s., and $r = .184$, n.s.].

In addition, non-assimilatory non-target pronunciation ([m]-like character in a velar context) was also found to correspond with decreased clarity [$r = -.361$, $p < 0.009$] for words preceding velar stop consonants.

Intelligibility subjects' responses

In an attempt to account for the lack of correlation between intelligibility and judgements of assimilation for words preceding labials, we analysed the alternative responses of the original subjects in the intelligibility studies.

The alternative words offered in cases of incorrect recognition were classed according to their word-final segment, and these subjects' responses were compared with the responses of the experts.

Words judged by experts as [m]-like elicited more incorrect identifications ending in [m]. This was true both of assimilated tokens preceding labials [$r = .212$, $p = 0.05$] and of tokens preceding velars which were judged by experts as sounding (inappropriately) [m]-like [$r = .313$, $p = 0.02$].

Words preceding velars and judged by experts to have assimilated towards [ŋ] correlated with subjects' incorrect identifications ending in [ŋ] [$r = .418$, $p = 0.002$]. However, words preceding labials and judged by experts as sounding inappropriately [ŋ]-like showed no relation to subjects' responses [$r = -.076$, n.s.]. A closer examination of this set of data revealed that subjects were offering words ending in [ŋ] regardless of the experts' judgements.

We suggest that this result can be explained by the structure of the lexicon in English. The productive -ING affix leads to subjects responding with lots of [ŋ] ending words, whether or not there is auditory evidence for velar articulation.

CONCLUSIONS

The general conclusion is that there is a relation between intelligibility and assimilation: tokens of a word which are perceived to have been assimilated result in poorer recognition when excerpted from context. We infer from this, that there is indeed a cost involved in the processing of assimilated tokens. It is necessary to exert effort in recognising the context in which an assimilation occurs in order for it to be successfully recognised as an appropriate change. Without supporting context, an assimilated token is harder to recognise than its canonical counterpart. These results are in line with experiments on cross-modal priming of isolated words, where a single feature mismatch reduces the priming effect.

We must also conclude that the relation between intelligibility and assimilation is complex. Firstly, the effect of assimilation on intelligibility varies according to the place of articulation of the assimilatory environment (e.g. labial or velar). Secondly, assimilation appears to be one of several factors which make tokens harder to recognise. The failure of perceived assimilation to account for the repetition effect on intelligibility indicates that there are other factors at play. We argue that these factors include not just the phonetic and phonological, but also the lexical. The structure of the lexicon, and the frequency of occurrence of particular morphological structures need also to be considered in any full account of what makes words easy or difficult to recognise.

This work was supported by the ESRC(UK) via the HCRC. Dr. McAllister is currently at the Department of Psychology, University of Auckland, New Zealand. This work was conducted while Dr. McAllister was in Edinburgh as a Visiting Reseach Fellow. Dr. Doherty-Sneddon is now at the Department of Psychology, University of Stirling. Address for correspondence: C. Sotillo, HCRC, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK

REFERENCES

[1] Bard, E.G., Shillcock, R.C. and Altmann, G.T.M. (1988), "The recognition of words after their acoustic offsets in spontaneous speech: effects of subsequent context", *Perception and Psychophysics*, 44, 395-408.

[2] Lieberman, P. (1963), "Some effects of semantic and grammatical context on the production and perception of speech", *Language and Speech*, 6, 172-5.

[3] Dalby, J (1984), *Phonetic structure of fast speech in American English*, Unpublished PhD thesis, University of Indiana.

[4] Marslen-Wilson, W.D. and Zwitserlood, P. (1989), "Accessing spoken words: On the importance of word onsets", *JEP:HPP*, 15, 576-585.

[5] Marslen-Wilson, W.D. and Gaskell, G. (1992), "Match and mismatch in lexical access" [abstract] *IJP*, 27, 61.

[6] Gaskell, G. and Marslen-Wilson, W.D. (in press), "Phonological variation and inference in lexical access", *JEP:HPP*.

[7] Anderson, A.H., et al. (1991), "The HCRC Map Task Corpus", *Language and Speech*, 34, 351-366.

[8] Bard, E.G., Sotillo, C., Anderson, A.H., Doherty-Sneddon, G., & Newlands, A. (forthcoming) "The control of intelligibility in running speech", *Proceedings of XIII ICPhS*, Stockholm.

[9] Yegnanarayana, B. (1981), "Speech analysis by pole-zero decomposition of short-time spectra", *Signal Processing*, 3, 5-17.

[10] Fowler, C.A. & Housum, J. (1987), "Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction", *JML*, 26, 489-504.