RECOGNITION OF ASSIMILATED WORDS IN NORMAL AND FAST SPEECH

Hugo Quené and Joyce Krull
Utrecht institute of Linguistics OTS, Utrecht University, the Netherlands

ABSTRACT

Even when words are normally assimilated in connected speech, listeners can recognise them easily. Word recognition may be so robust against assimilation, because listeners *expect* certain assimilation phenomena (given speech rate and style). From this hypothesis, we predict that listeners have more difficulty in recognising (unexpectedly) un-assimilated words than assimilated ones. This difficulty would even be greater when listening to fast speech. These predictions were tested in a word detection experiment, with speech of normal and fast rates. For normal speech, RTs were indeed slower for un-assimilated target words than for assimilated targets, as predicted. Listeners expect certain assimilations in the input speech, and they expect to undo these in word recognition. For fast speech, however, the effect of assimilation was reversed. Under adverse conditions, word recognition seems to be facilitated by a more faithful (un-assimilated) phonetic representation of the target word, even if such a realisation is unnatural.

1. INTRODUCTION

Words in connected speech deviate from their canonical form, partly because of sandhi phenomena such as assimilation. Assimilation may be considered as the modification of a phonological (i.e. distinctive) feature from one speech segment, under the influence of its neighbour(s). Some well-known examples are: place assimilation of /n/ (e.g. English *pho*[m]*e booth*, Dutch *i*[ŋ]*kopen*); place assimilation between plosives (e.g. English *swee*[k] *girl*); and deletion of /t/ between consonants (e.g. English *los*/t/ *property*, Dutch *pos*/t/ *brengen*). Assimilation is usually partial rather than complete, in that the resulting assimilated segment sequence is phonetically distinct from the corresponding segment sequence without assimilation (cf. in English *swee*[k] *girl* and *meek girl*, in Dutch *po*[sb]*rengen* and *mos brengen*) [4, 8].

In everyday connected speech, however, human listeners have no difficulty in perceiving assimilated word forms. Various means have been proposed to explain this observed robustness of human word recognition. The robustness against acoustic variation can be modelled in two ways [5, 2]. First, the way in which the sound form of a word is represented in the mental lexicon may allow for variation: the lexical form could be underspecified (i.e. without redundant phonological information). For example, if the final plosive in *post* is unspecified in the lexicon, then [pɔs] is a good match to that underspecified lexical representation. The sound form could also be specified in multiple variants (e.g. without and with deletion, *post* and *pos*) which are pre-compiled and stored permanently in the mental lexicon. Secondly, the auditory recognition of assimilated word forms may be modelled by means of phonological processing. Phonological inference rules mediate between the acoustic input and the single, fully specified representation of a word. For example, the input form [pɔsb...] is mapped onto the underlying representation [pɔstb...], (taking into account the following obstruent), which then matches the lexical representation.

In both accounts of assimilation in word recognition, it is conceivable that listeners *expect* certain assimilation phenomena to occur, given the style and rate of the input speech. These expectations may be modelled either as a preference for assimilated variants in the lexicon, or as an inclination to undo assimilation in the phonological parse. If an assimilation phenomenon occurs in normal style and rate, then listeners are biased towards undoing those assimilation rules in normal speech. Hence we hypothesize that listeners would have more difficulty in processing such (unexpectedly) un-assimilated word realisations than assimilated ones.

Recently, several experiments have indicated that assimilation has an effect on the recognition of the assimilated word. If assimilation occurs in *in*appropriate phonological contexts, then recognition of the target word is hampered. In appropriate contexts, however, there seems to be little perceptual difference between assimilated and unassimilated realisations. English listeners responded equally fast in a cross-modal priming task, whether the target word in the appropriate phonological context was assimilated or unassimilated (e.g. *lea*[m] *bacon*) vs. *lea*[n] *bacon*) [2]. Similarly, Japanese listeners responded somewhat faster in a monitoring task, if the phonetic realisation of the target, a moraic nasal, was assimilated to its phonological context (e.g. *to*[m]*bo* vs. *to*[n]*bo*) [7]. In Japanese, this phoneme is unspecified for place of articulation, and specified only as Nasal. The assimilated realisation in *to*[m]*bo* is still a perfect phonetic correlate of the corresponding phoneme (it does not "change into a different phoneme"), since the result is still a Nasal with unspecified place. By contrast, English listeners in a phoneme monitoring task were somewhat slower in responding to the target phoneme /t/ in the assimilated realisation (e.g. *freigh*[p] *bearer* vs. *freigh*[t] *bearer* [3]. Note that phonemic changes occur only in the former, and not in the latter realisation. The assimilated realisations in [3] do not contain simple phonetic correlates of the target phonemes, as opposed to the Japanese study [7].

These three studies suggest that there is no perceptual advantage for assimilated realisations over un-assimilated ones, in appropriate phonological contexts. In our opinion, this may

have been due to the combination of rate and style of the stimulus speech (and to the various experimental tasks). Perhaps rate and style were just so intermediate, that the relevant assimilation phenomenon was truly optional in the stimulus speech. In that case, listeners would have been unbiased towards assimilated or un-assimilated realisations. If the rate had been faster, or the style less formal, then listeners would have expected assimilation to occur, and they would have been biased toward undoing that assimilation.

The present experiment further investigates the hypothesis that listeners may expect assimilation to occur. To this end, we concentrate on an assimilation phenomenon which occurs regularly in Dutch speech of intermediate rate and formality, viz. deletion of /t/ between consonants [1]. The realisation of *post brengen* as [pɔsb...] is assimilated (with deletion), while [pɔstb...] is unassimilated (without deletion). Since listeners would expèct assimilation in this type of stimulus speech, we predict that assimilated words are recognised faster and with better accuracy than un-assimilated ones. We also predict that this advantage for assimilated words is even stronger in fast-rate speech (or in informal-style speech). For practical purposes, rate rather than style was chosen as the relevant factor to manipulate.

2. METHOD

The hypothesis was tested in a word detection study. Target words were embedded in meaningful sentences. Deletion of final /t/ of the target word, accentuation, and speech rate were varied independently, yielding 8 conditions (Table 1):

nr.	mnemonic	Deletion	Accent	Speed
1	+D+A+S	yes	yes	fast
2	+D+A-S	yes	yes	normal
3	+D-A-S	yes	no	normal
4	+D-A+S	yes	no	fast
5	-D-A-S	no	no	normal
6	-D-A+S	no	no	fast
7	-D+A+S	no	yes	fast
8	-D+A-S	no	yes	normal

Table 1. Summary of main conditions.

Deletion was varied by cut-and-paste of target realisations from assimilating and non-assimilating contexts. Accent was varied in order to obtain meaningful results even in the case of floor or ceiling effects, for unaccented and accented target words, respectively.

2.1. Stimulus Materials

Target words were 16 monosyllabic nouns, ending in /t/ preceded by an obstruent, and matched for lexical frequency. Examples are *post* [pɔst] 'mail', *vrucht* [vrœxt] 'fruit'. These target words were embedded in penultimate position in a meaningful sentence.

In order to vary the Deletion factor, both assimilated and unassimilated realisations were necessary. These were elicited as follows. For each target sentence, two additional and matching sentences were constructed, which had the target word in penultimate position, followed either by an assimilating context (starting with obstruent, e.g. *post bezorgen*) or by a non-assimilating context (starting with vowel, e.g. *post openen*). A female native speaker of Dutch read these 16×3 sentences several times in one session, in 2 accent versions and at 2 speech rates. Recordings were made in a sound-treated booth. The speaker was instructed to read the sentences as naturally as possible. The sentences were recorded on DAT, and later downsampled to 22.05 kHz and stored on computer disk. The deleted and undeleted conditions of the stimulus sentence were obtained by replacing the target word in the stimulus sentence, with its realisation from its matching assimilating and non-assimilating sentence, respectively. Special care was taken to avoid clicks at the paste point. After this manipulation, each target word in each stimulus sentence was clearly realised with or without its final /t/, as indicated in Table 1.

In addition, the speaker also read 12 filler sentences with a target word earlier in the sentence, and 28 catch sentences without a target word[1]. These filler and catch materials were read at two speech rates, during the same recording session.

2.2. Subjects and Procedure

80 Listeners (aged 18 to 55) participated in the experiment. They were all native speakers of Dutch, with normal hearing. They were not paid, but all knew that the best-performing three listeners would receive a monetary reward. They were divided over 8 groups of 10 subjects each. The 16 sentences × 8 conditions were divided over the 8 subject groups in a Latin-square design, so that each subject heard only one condition of a particular target word.

The experiment was conducted under computer control. The target word was presented visually for 750 ms before each item; it disappeared when the spoken sentence began. Subjects were instructed to press a button as soon as they heard the cue word in the stimulus sentence. Normal and fast speech items were tested in separate blocks; all items within these two blocks were randomized.

Reaction times of listeners' responses were measured from sentence onset. Responses later than 2 s after sentence offset were discarded. The remaining RTs were aligned to the offset of the target word.

3. RESULTS

Listeners failed to detect the target word in 2% of all presentations, with misses evenly distributed over listeners, sentences and conditions. All non-missing reaction times were included for further analysis. The average reaction times (in seconds) for each main condition are given in Figure 1 on the next page. At a first glance, these averages indicate that listeners found the task rather easy. This should come as no surprise, given the nature of the task. In effect, the spoken target word is always heavily primed by the preceding visual presentation (and recognition) of the target word itself. One could even argue that the observed reaction times do not reflect normal auditory word recognition at all: under normal circumstances, a target word is not primed by itself just before it

has to be recognised auditorily [6]. However, we assume that the auditory presentation of the target word results in a normal recognition attempt for the spoken target word; thus the observed reaction times reflect processing difficulties in auditory word recognition.

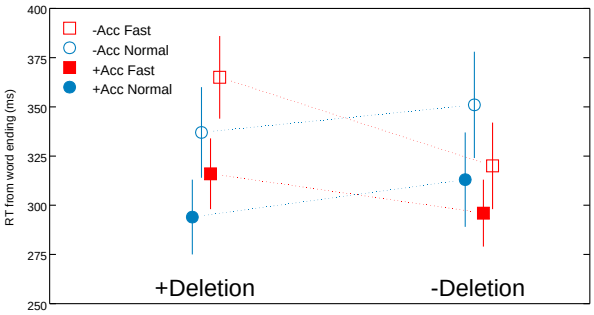

Figure 1. Average RTs in word detection (in ms, from the end of the target word) for each main condition. Error bars indicate 1 standard error above and below the average.

The results in Figure 1 suggest that there may be a main effect of Accent, but not of Deletion. In addition, there seems to be an interesting two-way interaction of Deletion:Speed which differs from the predicted interaction.

All reaction times were fed into two repeated measures ANOVAs, with Listener or Sentence as random factor, yielding F_1 and F_2 ratios, respectively. Fixed factors in both ANOVAs were Deletion, Accent, and Speed. Results of these ANOVAs suggest a main effect of Accent ($F_1(1,72)=3.83$, $p=.054$; $F_2(1,15)=2.37$, $p=.144$), but not of Deletion ($F_1(1,72)=2.39$, $p=.126$; $F_2(1,15)<1$, n.s.). The interaction of Deletion:Speed is marginally significant ($F_1(1,72)= 2.94$, $p=.091$; $F_2(1,15)=4.13$, $p=.060$).

In our opinion, this latter interaction should not be ignored. Its non-significance may well have been due to the very low number of items, viz. 2 sentences per condition per listener. This low number of sentences greatly reduces the power of our experiment. In this case, using the standard criterion of $\alpha=.05$ may well yield a Type-II error. A less strict significance level would be more appropriate here. Consequently, we assume that the observed interaction effect does reflect real behaviour of our listeners.

Other fixed effects and interactions are not significant, neither with a conventional $\alpha=.05$ nor with a more relaxed significance level.

4. DISCUSSION

Average RTs in Figure 1 show that there is no perceptual advantage for assimilated realisations of the target words, contrary to our general hypothesis. Our second prediction stated that the advantage of assimilated word forms would be greater in fast-rate speech. Results suggest that there is indeed an (insignificant) interaction between Deletion and Speed — but in the opposite direction, with a *dis*advantage for assimilated word forms in fast speech. As argued above, we assume that this interaction reflects real behaviour of our listeners.

For speech of **normal** rate, then, average RTs were faster for assimilated words than for unassimilated ones, as predicted by our hypothesis. This suggests that there is a perceptual advantage for assimilated word realisations. In the normal-rate stimulus sentences, listeners expected deletion of /t/ to occur, and hence to be undone in word recognition. Hence, realisations with deletion of /t/ were easier to recognize than canonical realisations with /t/ intact. It should be remembered, however, that this inclination of listeners to undo assimilation in the phonological parse can also be modelled as a preference for (pre-compiled) assimilated variants in the lexicon. On the basis of the normal speech data, we could conclude that their varying expectations help listeners to perceive spoken words correctly under varying circumstances.

For **fast** speech, however, average RTs were slower for assimilated words than for unassimilated ones, contrary to our hypothesis. Listeners seem to prefer a more faithful representation of the stimulus sentences. Obviously, it is more difficult to detect targets in fast sentences. Under these circumstances, a realisation with /t/ intact could in effect be more advantageous for word recognition. First, because such a realisation requires less phonological inference. Second, because a realisation with /t/ intact gives listeners somewhat more processing time, and it increases the redundancy of the speech signal. These perceptual advantages are apparently more important than the abnormality of encountering unassimilated word realisations in fast speech.

In conclusion, we see that with phonologically appropriate assimilation in normal-rate speech, assimilated word forms are recognised faster than unassimilated words. In fast-rate speech, assimilated word forms are recognised slower than unassimilated words, probably due to the higher processing load for fast speech.

ACKNOWLEDGMENTS

Our sincere thanks are due to Huub van den Bergh, for his advice and assistance in statistical matters (although this paper does no justice to his contributions), to Mieke Trommelen, for her advice on phonological matters, and to Theo Veenker, for technical assistance in running the experiments. Valuable comments throughout this work were provided by Huub van den Bergh, Sieb Nooteboom and Bert Schouten.

NOTE

1. Of the 28 catch sentences, 16 contained a word which initially overlaps with the specified target word, e.g. *schilder* 'painter' with target word *schil* 'rind'. This was done to avoid premature responses before the offset of the target word, where assimilation might occur.

REFERENCES

[1] Booij, G. 1995. *The Phonology of Dutch*. Oxford, Clarendon.

[2] Gaskell, M.G. and Marslen-Wilson, W.D. 1996. Phonological variation and inference in lexical access. *J. Exp. Psychology: Human Perception and Performance*, 22, 144-158.

[3] Gaskell, M.G. and Marslen-Wilson, W.D. 1998. Mechanisms of phonological inference in speech perception. *J. Exp. Psychology: Human Perception and Performance*, 24, 380-396.

[4] Koster, C.J. 1987. *Word recognition in foreign and native language: Effects of context and assimilation*. Dordrecht, Foris.

[5] Marslen-Wilson, W.D., Nix, A. and Gaskell, G. 1995. Phonological variation in lexical access: Abstractness, inference and English place assimilation. *Language and Cognitive Processes*, 10, 285-308.

[6] Nooteboom, S.G. 1998. personal communication.

[7] Otake, T., Yoneyama, K., Cutler, A. and Van der Lugt, A. 1996. The representation of Japanese moraic nasals. *J. Acoust. Soc. America*, 100, 3831-3842.

[8] Passy, P. 1890. *Étude sur les changements phonétiques et leurs caractères généraux*. Paris, Firmin-Didot.