ASSIMILATION AND ANTICIPATION IN WORD PERCEPTION

Hugo Quené, Maya van Rossum, Mieke van Wijck

Utrecht institute of Linguistics OTS
Utrecht University, The Netherlands

ABSTRACT

Words in connected speech are often assimilated to subsequent words. Some property of that upcoming word may then be determined in advance; these advance assimilatory cues may facilitate perception of that word. A gating experiment was conducted in Dutch, studying anticipatory voice assimilation between plosives, in 24 two-word combinations. In Dutch, voicing in a word-final plosive can only be caused by anticipatory assimilation to the next, voiced initial plosive, e.g. “rie[db]lint”. Voiced and unvoiced variants of final and initial plosives were cross-spliced.

Responses for assimilated, voiced-final stimuli show a strong bias to voiced-initial responses, as predicted. Even at longer gates in the hybrid condition “rie[dp]lint”, after hearing the unvoiced initial plosive, listeners often came up with a voiced-initial response, with high confidence. Hence, advance phonological ‘voiced-initial’ cues were often stronger than acoustic ‘unvoiced-initial’ cues. These gating results suggest that listeners use advance assimilatory cues in word perception.

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 spreading of a phonological (i.e. distinctive) feature from one speech segment to its neighbour(s). Some well-known examples are: vowel nasalisation before a nasal coda consonant (e.g. in French, Hindi, Bengali); place assimilation of /n/ (e.g. English *pho*[m] *booth*, Dutch *i*[ŋ]*kopen*); place assimilation between plosives (e.g. English *swee*[k] *girl*); and voice assimilation (e.g. Dutch *stro*[b]*das*, *huis*[f]*uil*, for details see below). Assimilation is usually partial rather than complete, in that the resulting assimilated segment is phonetically distinct from the corresponding ‘underlying’ segment (cf. *swee*[k] *girl* and *meek girl*, Koster 1987, Passy 1890).

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 (Marslen-Wilson, Nix & Gaskell 1995; Gaskell & Marslen-Wilson 1996). 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 place of articulation of the final plosive in *sweet* is unspecified in the lexicon, then [swik] is a good match to that underspecified lexical representation. The sound form could also be specified in multiple variants (e.g. without and with assimilation, *sweet* and *sweek*) 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 *swee*[k] is mapped onto the underlying representation *sweet*, (taking in account the following velar consonant), which then matches the lexical representation.

Recently, several experiments have indicated that assimilation has an effect on the recognition of the assimilated word. Gaskell & Marslen-Wilson (1995) found that English listeners responded faster in a cross-modal priming task if the phonetic realisation of the stimulus word matched its phonological context (“viable” assimilation, e.g. *lea*[m] *bacon*), as compared to non-matching contexts (e.g. *lea*[m] *gammon*). Similarly, Otake, Yoneyama, Cutler & Van der Lugt (1996) found that Japanese listeners responded slightly faster in a monitoring task if the phonetic realisation of the target, a moraic nasal, matched its phonological context (e.g. *to*[m]*bo*), as compared to non-matching contexts (e.g. *ko*[m]*to*).

Both experiments indicate that appropriate assimilation facilitates recognition of the affected word somewhat. These and similar experiments crucially depend on stimulus material in which assimilated and unassimilated segments are cross-spliced into viable and unviable contexts for assimilation. The results could also be explained, as Lahiri (1995) has pointed out, as a ‘surprise effect’ if segment and context mis-match (as in English *lea*[m] *gammon*, Japanese *ko*[m]*to*, German *Wei*[m] *keltern*). After hearing the assimilated segment in these examples, listeners would expect a labial consonant as the following segment. The absence of such a consonant (i.e. the presence of a non-labial) is confusing, and this confusion somehow interferes with the response. According to Lahiri (1995), these experiments show primarily that inappropriate assimilation hampers word recognition, due to mis-match between stimulus and underspecified representation.

Hence, there is only weak and indirect evidence for a positive effect of appropriate assimilation on recognition of the assimilated word itself. However, if listeners are confused by unexpected speech sounds (as Lahiri 1995 suggests), then this confusion can only arise if listeners make phonological expectations about upcoming speech sounds. Hence, appropriate assimilation (where such expectations are met) may facilitate recognition of the **next** word, which provides the assimilatory context.

In case of anticipatory assimilation, some properties of the sound form of the next word are present in advance, i.e. before that next word. That is, the assimilated word contains phonetic cues about the subsequent word. (The artificial mis-match between these phonetic cues and the subsequent word would presumably surprise listeners, thus slowing down responses to the first word.) In the case of complete assimilation in *lea*[m] *bacon*, the realisation of the first word indicates that the second word begins with a bilabial. Even in the case of partial assimilation, a realisation as [lin^w] would indicate the same. Hence, some properties of the second word may be anticipated, on the basis of phonetic assimilatory cues. Our hypothesis is that such advance information, resulting from anticipatory assimilation, facilitates recognition of the second word.

This hypothesis was investigated in Dutch, where Regressive Voice Assimilation (RVA) provides a relevant assimilation process. In Dutch, obstruents in coda position are always devoiced. Hence, if two adjacent plosives differ with respect to phonological voicing, then their phonological pattern is always Unvoiced-Voiced (and never V-U), as in *zak+doek, op+drinken, riet blazen, sleep dragen*. In these contexts, anticipatory assimilation of voice (RVA) changes the voicing feature of the first, coda consonant, yielding *za*[g]*doek*, *o*[b]*drinken*, *rie*[d] *blazen*, *slee*[b] *dragen*. Hence, voicing in the final plosive of *riet* can only be caused by the voicing of the following initial plosive of *blazen*. This anticipatory information about the voiced onset of the second word may facilitate perception of that word.

This hypothesis was tested in a gating study (Grosjean 1980). Stimuli were nonsensical combinations of two Dutch monosyllabic words, with plosives as word-final and word-initial consonants. Voicing in each of the 2 plosives was varied independently, yielding 4 voicing conditions (Table 1):

Condition	Example	Assimilation	Context
U#U	*rie*[tp]*lint*	no	unviable
U#V	*rie*[tb]*lind*	no	viable
V#U	*rie*[dp]*lint*	yes	unviable
V#V	*rie*[db]*lind*	yes	viable

Table 1: Summary of voicing conditions.

2. METHOD

2.1. Stimuli

Initially, 36 two-word combinations of Dutch monosyllabic words were used. The 2 plosives in the assimilation context had to be heterorganic, hence only /tp/ and /pt/ were used, with all their voicing variants given in Table 1 (Dutch has no voiced velar plosive phoneme). In this study, it is essential that the second word constitutes an existing word, irrespective of the voicing of its initial consonant. Since *plint* and *blind* are both existing Dutch words (differing only in the voicing of their initial consonant), the distribution of responses over these two possibilities provides information about the perceptual use of anticipatory voice assimilation. The word pairs were placed at the end of neutral carrier sentences (of which there were 4).

In order to construct the 4 voicing conditions, both all-unvoiced [tp, pt] and all-voiced [db, bd] realisations were necessary. These were elicited as follows. For each two-word combination, two sentences were constructed with either the voiced-initial or unvoiced-initial word combination (e.g. *riet plint* and *riet blind*). All 36×2 sentences were put in random order, mixed with 64 filler sentences. A female native speaker of Dutch read the sentences, seated in a sound-treated booth. She was instructed to read the sentences as naturally as possible; she was unaware of the purpose of the experiment. Presumably, this procedure should yield no voice assimilation in *riet plint* (unviable context) and complete voice assimilation in *riet blind*, (viable context, realised as *rie*[db]*lind*). The sentences were recorded on DAT, and later downsampled to 22.05 kHz and stored on computer disk.

2.2. Pre-test

In order to verify whether the stimuli were realised with either no voice assimilation or with complete voice assimilation, a pre-test was conducted. Ten judges, phonetically trained, classified each stimulus with respect to the voice feature of the first plosive. To avoid a bias towards ‘unvoiced’ judgments (caused by the obligatory word-final devoicing of such plosives in Dutch), the lexical structure of the stimulus was destroyed by presenting only the VC#CCV part of each realisation.

A realisation was considered all-voiced if 8 (out of 10) judges classified the first plosive as voiced, and mutatis mutandis for all-unvoiced realisations. A two-word stimulus combination was discarded if either the all-voiced or all-unvoiced realisation failed to meet this criterion value. After this selection, there remained 24 non-ambiguous stimulus combinations, which could be used for the gating study. Of the 12 discarded stimulus combinations, 11 were judged to be ambiguous in the resulting degree of voicing, while 1 contained mispronunciations which were not detected during recording.

2.3. Manipulations

The all-unvoiced (U#U) and all-voiced (V#V) conditions were obtained from the un-manipulated, natural realisations described above. The hybrid conditions (U#V and V#U) were obtained by cross-splicing the separate words of the all-unvoiced and all-voiced realisations. Special care was taken to ensure that the acoustic correlates of voicing were left intact (e.g. duration of closure, voice bar during closure, intensity of release burst, etc.) while also yielding smooth transitions at the paste point. After this manipulation, each plosive in each stimulus was clearly unvoiced or clearly voiced, as indicated in Table 1.

Next, segment-size gates were made. The ‘zero’ gate was terminated before the first consonant of the second word, immediately after the release burst of the pivotal voiced or unvoiced plosive consonant. This gate also contained the preceding carrier sentence. Every subsequent gate consisted of the previous gate, plus the next neighbouring speech sound. Segment boundaries were determined with oscillographic, spectrographic and auditory feedback. All cuts were made at negative zero crossings, without amplitude smoothing. Each gate was stored in a separate file.

2.4. Subjects and Procedure

48 Listeners (aged 19 to 40) participated in the experiment. They were all native speakers of Dutch, with normal hearing. They were divided over 4 groups of 12 subjects each. The 24 combinations × 4 conditions were divided over the 4 subject groups in a Latin-square design, so that each subject heard only one condition of a given two-word combination.

Subjects were asked to write down which (second) word they had heard, as well as a confidence rating of their written response (1= complete guess, 10= totally confident). They were informed that the word combinations were nonsensical, and they were urged to give a response for each gate, even if uncertain. Gates of a stimulus combination were presented in incremental fashion, without further blocking. Subjects had 6 s response time after each gate; a short 500 Hz tone indicated the next word combination. An experimental session started with 5 practice combinations, after which feedback on the instructions was possible. The first 4 items of the real test were filler combinations for 'warming-up', which were excluded from further analysis.

3. RESULTS

According to our hypothesis, listeners use the perceived voicing in the pivotal consonant to anticipate the voicing of the following, word-initial consonant. This was investigated by classifying each response as voiced-initial or unvoiced-initial. The **percentage of voiced-initial responses** is plotted in Figure 1, as a function of gate length.

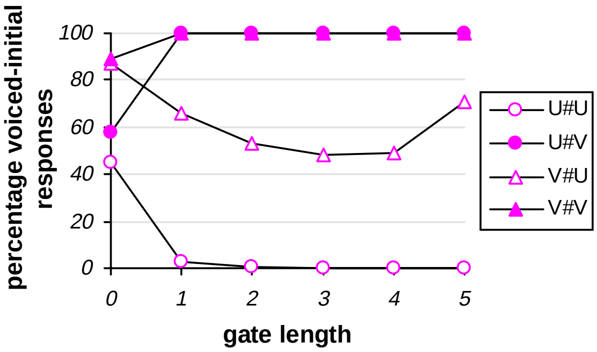

Figure 1: Percentage of voiced-initial responses, as a function of gate length (in speech segments), broken down by voicing condition.

These results show that voicing of the pivotal plosive affects identification of the following initial plosive, even before that initial plosive is presented (at gate 0). If the first word ends with an unvoiced plosive, e.g. *rie*[t] (circles), then listeners come up with guesses for the second word with either voiced or unvoiced initial consonants. The absence of voice assimilation in these conditions is compatible with both an unvoiced initial plosive (unviable context, no assimilation) and with a voiced initial plosive (viable context, no assimilation). If the first word ends with a voiced final consonant, e.g. *rie*[d] (triangles), listeners mainly report words containing voiced initial consonants. The presence of voice assimilation in these conditions is more compatible with a voiced initial plosive (viable context, assimilation).

Interestingly, the hybrid condition V#U (with unvoiced initial plosive) yields a substantial percentage of voiced-initial responses. This is seen even at longer gates containing this consonant, where listeners had sufficient acoustic information to classify the initial plosive correctly (as they did in the U#U condition). This indicates that listeners' responses were based not only on acoustic cues related to the initial plosive itself, but also on the preceding phonological context.

The hypothesis that appropriate assimilation facilitates recognition of the second word, may be tested by inspecting the **isolation points**. After an unvoiced plosive (no assimilation), there should be no facilitation. After a voiced plosive however (assimilation, e.g. *rie*[d]), voiced-initial words (*blind*) should be facilitated relative to unvoiced-initial words (*plint*). This was investigated by calculating the isolation point for each stimulus in each condition. The isolation point is defined as the gate number where 80% of the listeners come up with a correct guess for the second word, without changing their response at subsequent gates (Grosjean 1980). Average isolation points are given in Table 2:

Condition	Example	average	sd, n
U#U	*rie*[tp]*lint*	3.2	0.8, 23
U#V	*rie*[tb]*lind*	3.5	0.7, 22
V#U	*rie*[dp]*lint*	3.7	0.5, 6
V#V	*rie*[db]*lind*	3.4	0.7, 23

Table 2: Average isolation point (in gate number, or number of speech segments of the second word), with standard deviation and number of stimuli, broken down by voicing condition.

These differences between conditions were not significant in a one-way analysis of variance [F(3,70)<1]. The three conditions that are phonologically viable (U#U, U#V, V#V) all yield identical isolation points. Perception of a voiced-initial second word (*blind*) does not require less stimulus information in the 'assimilated' V#V condition (*rie*[db]*lind*), than in the 'unassimilated' U#V condition (*rie*[tb]*lind*), contrary to our prediction.

In the unviable V#U condition (*rie*[dp]*lint*), however, listeners often persisted in their initial unvoiced-to-voiced confusion described above. Because of these incorrect responses, the percentage of correct responses was thus often below the 80% criterion value for isolation points, and no isolation point could be calculated for these stimuli. This explains the low number of isolation points in the V#U condition in Table 2. This may be illustrated by the percentage of correctly recognised second words at the longest gate, where the whole second word was presented. In the V#U condition, listeners eventually came up with the correct word in only 55% of all stimuli, as opposed to 98% in the other conditions.

Closer inspection revealed that most incorrect word responses were voicing confusions in the initial plosive, e.g. responding *blind* rather than *plint* for the hybrid V#U stimulus *rie*[dp]*lint*. This type of error occurred in 43% of all stimuli in the V#U condition.

Remarkably, listeners' **confidence ratings** for these incorrect word responses were the same as for correct responses, viz. on average 6.5 and 6.9, respectively [$t(269)=1.1$, n.s.]. This indicates that listeners were not aware that there was a conflict in this condition between the preceding phonological context and the phonetic voicing of the initial consonant. Their incorrect responses were only based on the preceding phonological context. Although these incorrect responses were not based on acoustic-phonetic cues, this did not decrease subjects' confidence in their response.

4. DISCUSSION

The results of this gating study indicate that the identification of the second, word-initial plosive is affected by its preceding phonological context. Hence, listeners use anticipatory information to identify phonemes in connected speech, in accordance with our general hypothesis.

The aim of human speech processing is not to identify phonemes, but to recognise words. We predicted that less acoustic information would be required in case of appropriate assimilation in a viable context, but this was not observed.

Why was the predicted facilitation due to appropriate assimilation not observed? First, any positive effect is probably rather small: Gaskell & Marslen-Wilson (1995) and Otake et al. (1996) report effects with magnitudes roughly between 10 and 50 ms (in reaction times). These reported effects are even smaller than the segment-sized gate increments in our study. It could well be that the crucial dependent variable, viz. isolation point or gate number, was not sensitive enough to measure such a small effect. Secondly, in our study, anticipatory assimilation may help listeners to determine that the next, word-initial plosive is voiced, and hence to recognise the next word on the basis of less acoustic information. But Figure 1 shows that the relevant difference between U#V and V#V conditions has already disappeared at gate 1. In other words, the [+voice] attribute of the initial plosive is always correctly perceived immediately after that plosive has been presented. At that point in time, anticipatory assimilation has already contributed its positive effect, if any, even though the word is not yet isolated. Hence, the isolation point provides measurements at a point in stimulus time that is irrelevant for our hypothesis. What is needed, clearly, is a different experimental task, which allows on-line measurements (e.g. latencies) of word recognition. Such experiments, preferably with the same stimulus material, are now in preparation.

Perhaps most interesting are the many incorrect responses in the unviable V#U condition. Our subjects were not surprised by the (unviable, hence) unexpected unvoiced initial plosive. Instead, they just ignored the acoustic-phonetic cues to its [-voice] attribute, and responded in accordance with the preceding phonological context, in about half of the stimuli. This behaviour may be deduced from listeners' unvoiced-to-voiced confusions, in combination with their confidence ratings.

This pattern of results indicates that listeners do indeed use advance phonological cues to anticipate on certain attributes of the upcoming speech sounds. When the speech sound turns out to be different, however, listeners are not surprised nor confused, but they often fail to hear the mis-matching acoustic cues. This unexpected insensitivity to acoustic information is in fact even stronger support for our general hypothesis than we had anticipated.

In conclusion, this gating study indicates that listeners use anticipatory phonological information, resulting from assimilation, for the recognition of upcoming words in connected speech. For more compelling evidence, however, further experiments using on-line measurements are required.

5. REFERENCES

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

2. Grosjean, F. "Spoken word recognition processes and the gating paradigm", *Perception and Psychophysics, vol. 28*, 1980, 267-283.

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

4. Lahiri, A. "Undoing Place Assimilation", paper presented at the 129th Meeting of the Acoustical Society of America, Washington DC, 1995.

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

6. Otake, T., Yoneyama, K., Cutler, A. & Van der Lugt, A., "The representation of Japanese moraic nasals", *J. Acoust. Soc. Amer., vol. 100,* 1996, 3831-3842.

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