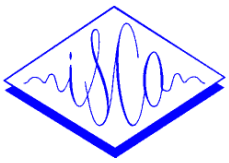

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

PERCEPTUAL EFFECTS OF PLACE AND VOICING ASSIMILATION IN DUTCH CONSONANTS

Vincent J. van Heuven & Willy Jongenburger

Dept. Linguistics/Phonetics Laboratory/Leiden University
and
Holland Institute of Generative Linguistics
Leiden/Amsterdam, The Netherlands

ABSTRACT

In this study we develop and test the hypothesis that spoken word recognition in context is facilitated by anticipatory assimilation but inhibited by perseveratory assimilation. Three experiments are presented. Experiment I shows that homorganic final nasals provide sufficient perceptual cues to reconstruct the place of articulation of the following stop which was electronically eliminated from the stimulus. In experiment II the target stops were not eliminated from the stimulus but presented - through digital cross-splicing - after homorganic versus heterorganic nasals. Removing anticipatory assimilation proved detrimental to the recognition of the target stops, thereby showing a positive effect of anticipatory assimilation on word recognition. In experiment III cross-splicing was used to simulate blocking of perseveratory assimilation of voicing. Here the results show that targets were recognised better when assimilation was blocked, showing that, indeed, perseveratory assimilation inhibits spoken word recognition.

1. INTRODUCTION

Assimilation is the phonological process by which a property of one sound spreads to a neighbouring (typically adjacent) sound so that the two sounds become more alike. It seems safe to assume that assimilation primarily serves the needs of the talker, since it avoids awkward sound sequences. For instance, it has often been observed that it is virtually impossible to pronounce two adjacent obstruents in connected speech that have opposed voicing characteristics. Under such circumstances, either both obstruents become voiced, or both become voiceless.

The listener pays the price for the increased articulatory ease on the part of the talker. Due to assimilation the initial or final sounds of words and morphemes lose some of their distinctive qualities. Consider the Dutch minimal pair *sein* vs. *zijn* /sEIn - zEIn/ ('signal - his'), the members of which are differentiated through a voiced-voiceless contrast in the initial fricative. Dutch has an assimilation rule that obligatorily devoices a fricative after a voiceless obstruent [1]. Thus after the preposition *met* /mEt/ 'with' it is no longer possible to make a lexical distinction between the two words, which are pronounced as [mEt sEIn] in both cases. This puts a heavier processing burden on the listener, since he has to do more work in order to reconstruct the speaker's intended message from a more ambiguous signal than would have been the case when no assimilation had been applied.

Assimilation is an asymmetrical process. One sound spreads a characteristic onto another sound. As a result the former sound keeps its original shape, whilst the latter loses some of its identity. When a later sound spreads a property onto a preceding sound, the assimilation process is **anticipatory**. In the reverse case, when an earlier sound spreads a feature to a following sound, the assimilation is called **perseveratory**.

When assimilation is applied across a morpheme or word boundary one of two things happen:

- in the case of anticipatory assimilation the final sound of an early morpheme or word is changed, and the initial sound of the next word remains unaffected.
- in the case of perseveratory assimilation the final sound of the earlier morpheme is unaltered whereas the initial sound of the next word is changed.

We know from studies of auditory word recognition that the word onsets are important but that the final sounds of a word generally add little information to the recognition process. Polysyllabic words can be (and in fact are) uniquely distinguished from their competitors before their final sounds are heard. In connected speech, due to the extra information supplied by the preceding context, this is also true of monosyllabic words [2]. Therefore we predict that a word's recognition is not impeded when its final sound is changed through assimilation. Assimilation often changes the word into a non-word, or into a word that blatantly does

not fit in with the preceding context [3]. In such cases the listener may realise that the change could only have been caused by anticipatory assimilation. One might hypothesize, therefore, that anticipatory assimilation heralds certain distinctive properties of the next word's onset, which would facilitate the next word's recognition, while at the same time the recognition of the assimilated word is not endangered. Anticipatory assimilation therefore serves both the interests of the talker and of the listener.

When assimilation is perseveratory, the predictions are different. Now the important, informative onset of a word is damaged, so that the recognition of the target word suffers. The final sound of the preceding word or morpheme is left intact, but since the word-ending is redundant anyway, the listener stands to gain little by this. Perseveratory assimilation may serve the talker's needs, but is undesirable from the listener's point of view.

We know that listeners are quite capable of undoing the effects of assimilation. They know that an initial /s/ after a preceding obstruent in connected (fast) speech may either be a true /s/ or an assimilated /z/. However, the listener's preferred interpretation of such stimuli is the one that is closest to the phonetic surface [4], i.e. [s]. It takes time and effort to realize that such a stimulus is compatible with /z/.

The predictions that were developed above, are not easily tested. The effects of assimilation will be determined by measuring phoneme identification of target sounds in stimuli where assimilation is or is not blocked (digitally removed). Positive effects of assimilation will then show up as a deterioration of the subject's performance when assimilation is blocked; negative effects of assimilation will be visible as superior performance for stimuli without assimilation. Clearly, we need pairs of utterances that are identical in all respects except for the fact that assimilation is applied to one member of a pair but blocked in the other. Human speakers would never be able to provide such utterance pairs, since, obviously, blocking an assimilation process can only be executed at the expense of a different temporal organisation of the sounds across the target word boundary. We therefore decided to take recourse to assembling utterance pairs through digital splicing of LPC-parametrised speech, which allows us to present the same word tokens in different spoken environments without introducing discontinuities in pitch contour or spectral energy distribution at the splice.

Concretely, the questions that we address in this study, are the following:

(i) is word recognition inhibited when, in fluent speech, the effects of anticipatory assimilation are removed,

(ii) is word recognition facilitated when, in fluent speech, the effects of perseveratory assimilation are removed from the speech signal?

(iii) alternatively, does the listener expect (perseveratory) assimilation to occur in a given environment and is he confused when the assimilation effects are removed?

These questions were taken up in three experiments. The first two experiments examine the potentially detrimental effect of blocking (i.e. electronically undoing) the effects of anticipatory (homorganic nasal) assimilation on the perception of word-initial stops. The third experiment addresses the complementary case, examining the effects of blocking perseveratory assimilation (of voicing) on the recognition of word-initial obstruents.

2. EXPERIMENT I

In experiment I we merely wished to determine the contribution of anticipatory assimilation of a word-final nasal to the identification of the place of articulation of the following word-initial stop. Although studies on the perceptual effects of anticipatory coarticulation in Dutch abound (for references see [4]), no comparable studies exist that simply measure the extent to which an upcoming sound (plosive) can be identified on the basis of information in the immediately preceding nasal.

Method. A male speaker of standard Dutch recorded five minimal triplets, whose members differed in the articulation place of their initial stop consonants: [p,t,k]. Targets were embedded in a fixed carrier sentence as in:

dat is een paard	*dat is een taart*	*dat is een kaart*
[dAt Is @m pa:rt]	[dAt Is @n ta:rt]	[dAt Is @N ka:rt]
that is a horse	that is a pie	that is a card

Under visual and auditory control of a high-resolution digital waveform editor (10 kHz, 12 bits, 4.5 kHz LP) the portion of the waveform from the end of the homorganic nasal until 50 ms into the target vowel was eliminated (i.e. removing the target-initial stop as well as the complete CV-transition), and replaced by one of two pink noise bursts. In the "variable noise" condition, the noise burst was given exactly the duration of the portion of the waveform that had been eliminated, so as to maintain any cues that could be provided by the original temporal organisation of the utterance. In the "fixed noise" condition the inserted noise burst was given a constant duration of 165 ms, i.e. the mean duration of the 15 noise bursts of the earlier condition, thereby obliterating any remaining temporal cues. The noise bursts' intensity was sufficient to create the auditory illusion that the entire speech signal was still present [5,6]. As a control condition the stimuli in the "fixed noise" condition were included, but with the portion of the utterance preceding the inserted noise burst deleted.

The set of 45 stimuli were presented twice, the second time in reversed order, to 20 adult native Dutch listeners, with 3s ISI (offset to onset) over headphones, who identified the consonant that was spoken during the noise bursts on the tape as /p/, /t/ or /k/, with forced choice.

Results and conclusion. Table I presents mean identification scores (in percent) for /p/, /t/, and /k/ broken down by the three presentation conditions.

On average, deleted stops are recognised from the preceding assimilated homorganic nasal at 76% correct when the target was replaced by variable noise, and at 75% correct when the noise burst had fixed duration. Notice that /t/ is

Table I: Correct identification (in percent) of deleted plosives with preceding homorganic nasal present or absent

information present in context preceding target	target consonant		
	p	t	k
with nasal, fix. noise	87	60	81
with nasal, var. noise	83	62	80
no context, fix. noise	60	56	25

identified poorly (61%) whilst /p/ and /k/ are identified above 80% correct. This indicates that the preceding labial or velar nasal is picked up as a clear place cue (for /p/ and /k/, respectively), but that the alveolar nasal (the unmarked articulation place) is relatively compatible with all places. When the preceding context with the assimilated nasal was deleted, percent correct stop identification dropped to 47. The latter percentage is still above chance (33%), which shows that some information on the deleted stop was left in the target word's vowel. Crucially, stop identification is 28% better, on average, when the homorganic nasal is present. The contribution of assimilation is larger (56%) for /k/ than for /p/ (25%). The improvement is negligible (a mere 5%) for /t/, which is mainly a consequence of the poorer cue value of the alveolar nasal in the contexted conditions (see above). The effect of context is significant, $F(2,1797)=95.0$ ($p<.001$), as is the effect of target articulation place, $F(2, 1797)=29.1$ ($p<.001$) and the context x place interaction, $F(4,1795)= 25.3$ ($p<.001$). The two noise contexts do not differ from each other (Newman-Keuls procedure, $p<.05$) but both differ from the no-context condition.
We conclude that anticipatory homorganic nasal assimilation provides significant perceptual cues to the articulation place of a following word-initial stop.

3. EXPERIMENT II

Given that a preceding homorganic nasal contains perceptually useful information on the identity of a following word-initial consonant, can we show that listeners actually use this information in running speech, when the identity of the target word-initial consonant is not obscured by some experimental trick (such as replacing the target by a noise burst)? This question is taken up in the second experiment.

Method. A male Dutch standard speaker recorded (for details cf. experiment I) 8 minimal monosyllabic triplets differing only in the place of articulation of the initial voiceless stop, in a fixed carrier *Ik herken* ... [Ik hErkEn ...] 'I recognise ...'. The speaker (first author) consistently applied assimilation across word boundaries, so that the final nasal of *herken* was pronounced as /m/, /n/ or /N/ depending on the target word. The recordings were A/D converted (10 kHz, 12 bit, 4.5 kHz LP) and parametrised (F1/B1 through F5/B5, Amplitude, F0, 10 ms frames). Targets and carriers were cross-spliced (using parameter files) such that there were three correct (homorganic) nasal-plosive combinations [m-p, n-t, N-k], and two non-assimilated combinations [n-p, n-k] (as in deliberate speech where assimilation is blocked). Any discontinuities in parameter tracks were smoothed across the splice over a 30 ms window. Notice that homorganic combinations were created through cross-splicing (by exchanging homorganic carriers between triplets), so as to ensure a fair comparison between homorganic and heterorganic versions of the same targets.
The resulting 40 stimuli were presented in random order, as part of a larger experiment. Fifteen native Dutch listeners identified the target plosives with forced choice from [p,t,k]). On a second presentation listeners rated the stimuli for naturalness along a five-point scale (1: highly unnatural, 5: highly natural) with special instruction to attend to the transition between the last two words of each utterance.

Results and discussion. The results of this experiment are presented in table II.

Table II: Identification error (percent) and judged naturalness (1: highly natural, 5: highly unnatural) for /p,t,k/ when homorganic nasal assimilation was applied and (artificially) blocked.

homorganic nas. assim. applied?	error (%)			naturalness		
	p	t	k	p	t	k
yes: homorganic nas.	1	6	0	4.8	4.5	4.9
no: neutral nas.	6	-	7	4.5	---	4.6

Our listeners identified the target-initial stops almost perfectly: 2% error on average. As in experiment I, the alveolar stop was identified less adequately. The naturalness ratings are good, and in perfect agreement with the identification scores (/t<p<k/). When the effects of anticipatory assimilation are blocked, the error rates for /p/ and /k/ increase considerably (from 1% to 7%), and naturalness drops by .3). Note that /t/ is no part of this comparison since the preceding nasal is the same (/n/) whether assimilation is blocked or not.
We conclude that anticipatory homorganic nasal assimilation provides a measurable and useful cue to the identity of an onset consonant, even when the target is left fully intact.

4. EXPERIMENT III

In this experiment we test the prediction that perseveratory assimilation negatively affects the identity of the word-initial consonant.

Method. Twelve minimal pairs were selected that differed only in their onset fricatives, 8 labial /f,v/ pairs (e.g. *fel* /fEl/ 'fierce' vs. *vel* /vEl/ 'skin'), and 4 alveolar /s,z/ pairs (e.g. *saai* /sa:j/ 'boring' vs. *zaai* /za:j/ 'sow vb.' These were pronounced in preceding carriers that ended in either a vowel (V- or voicing context: *Hij zei* ... /hEI zEI .../ 'He said ...') or a plosive (C- or devoicing context: *Hij zegt* ...

/hEI ZExt/ 'He says ...'.
Using the same speaker and procedures as in experiment II, the minimal word pairs were cross-spliced between carriers such that the acoustic effects of perseveratory assimilation were maintained or eliminated. The same 15 listeners that took part in experiment II identified the target words, and rated their naturalness.

Results. Table III summarises the results of this experiment.

Table III: Identification error (percent) and naturalness (italics; 1: highly unnatural, 5: highly natural) for fricatives /f,s,v,z/ that were originally spoken in a voiced (V-) versus devoicing (C-) context and cross-spliced into a voiced versus devoicing context.

target's original context	context spliced into	intended target			
		/f/	/s/	/v/	/z/
V-	V-	84 *4.3*	97 *4.5*	97 *4.3*	100 *4.8*
V-	C-	81 *4.0*	88 *4.4*	97 *4.1*	100 *4.3*
C-	V-	30 *2.6*	88 *4.2*	70 *3.5*	43 *3.0*
C-	C-	64 *3.4*	33 *4.4*	88 *3.0*	60 *3.4*

The results show that the original context in which a target is produced, it the overriding factor for the target's correct (intended) identification. Targets taken from and spliced back into a preceding context that is conducive to voicing (V-), are identified best, with scores between 84 and 100% correct and naturalness ratings of 4.3 and up. Crucially, when the same targets are spliced into a C-context, i.e. an obligatory devoicing context, the correct identification of voiced fricatives is not impeded at all. Yet listeners do consider these utterances slightly less natural, since the ratings drop by .4 on average. Apparently, blocking perseveratory assimilation of voice yields highly intelligible voiced fricatives, even though they sound slightly odd in a context where devoicing is obligatory.
Interestingly, the identification of underlying voiceless fricatives deteriorates somewhat: /f/ and /s/ sustain a loss of 3 and 11 percentage points, respectively, and lose .2 point on the naturalness scale. Presumably, voiceless fricatives are produced less forcefully in a voicing context, so that they are not voiceless enough to make truly convincing voiceless fricatives when they are spliced into a devoicing context.
Targets that were originally produced in a devoicing context are identified relatively poorly, with scores between 30 and 88% correct. When targets are spliced back into a similar devoicing context, the voiced fricatives are identified better than the cognate voiceless phonemes. The identification of underlying voiceless fricatives oscillates more or less randomly between the voiced and voiceless cognates, indicating that listeners know that both voiced and voiceless responses are compatible with a voiceless fricative in a devoicing context. However, when a moderate cue to voicing is picked up, as must have been the case for the underlying voiced fricatives, the balance tips towards a voiced fricative.
Identification scores are poorer still when these fricatives are spliced into voicing contexts. The moderate cue favouring voiced responses for intended voiced fricatives loses some of its strength, so that voiced responses lose 18% when the targets are spliced into a voicing context. Naturalness ratings as well as identification scores behave unsystematically for the underlying voiceless fricatives, which we shall not discuss any further.
The overall result that emerges is that fricatives are better identified when perseveratory voicing assimilation is blocked. This indicates that perseveratory assimilation impedes the recognition of words that are affected by it.

5. CONCLUSION

Experiments I and II show that anticipatory assimilation is conducive to word recognition (question 1, introduction), particularly to the identification of the place feature of word-initial stops. Experiment III bears out that perseveratory assimilation inhibits the correct identification of word-initial fricatives, so that word recognition will suffer (question 2, introduction). Moreover, listeners realise that ambiguities remain as a result of perseveratory assimilation: both voiceless and voiced responses are viable options in a devoicing context (question 3, introduction), so that the correct decision may eventually be taken, but only at the cost of extra processing.
Finally, our results have implications for speech technology. If human speakers must apply assimilation rules in fluent speech, machines can be programmed not to. Our data show that intelligibility of speech is superior when perseveratory assimilation processes are blocked throughout, even though a marginal loss of naturalness is incurred.

ACKNOWLEDGEMENT

Experiment I was run by Caroline Vaneveld, as part of a Master's thesis written under the supervision of the first author.

REFERENCES

[1] Jongenburger, W., Heuven, V.J. van (1993). Sandhi processes in natural and synthetic speech, in V.J. van Heuven, L.C.W. Pols (eds.) *Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation*, Mouton de Gruyter, Berlin, 261-277.
[2] Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition, Cognition, 25, 71-102.
[3] Koster, C.J. (1987). *Word recognition in foreign and native language*, Foris, Dordrecht.
[4] Dupuis, M.Ch. (1988). Perceptual effects of phonetic and phonological accommodation, an experimental study on effects of coarticulation and assimilation on perception of words and word beginnings, Doct. diss. Leiden University.
[5] Warren, R.M. (1970). Perceptual restoration of missing speech sounds. *Science*, 167, 292-293.
[6] Warren, R.M., Obusek, C.J. (1971). Speech perception and phonemic restorations, *Perception and Psychophysics*, 9, 358-362.