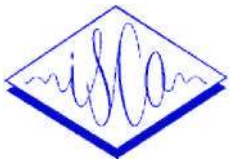
ISCA Archive
http://www.isca-speech.org/archive



ETRW on Sound Patterns of
Spontaneous Speech (SPoSS)
La Baume-les-Aix, France
September 24-26,1998

THE RECOGNITION OF SPOKEN WORDS WITH VARIABLE REPRESENTATIONS

Anne Cutler
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
E-mail: anne.cutler@mpi.nl

RESUME

Selon les modeles actuels de reconnaissance de la parole, l'identification des mots se deroule en deux temps: une phase d'activation des candidats lexicaux, et une phase de competition. Cependant, la plupart des recherches portant sur ce sujet ont ete realisees avec des stimuli moins naturels, enregistres dans des conditions acoustiques optimales. Depuis peu, les psycholinguistes ont commence a s'interesser aux caracteristiques acoustiques de la parole spontanee et a leurs consequences sur les processus d'activation du lexique. Cette presentation fournit un resume des recherches r6centes sur les effets de resyllabification, de reduction, d'assimilation et d'insertion des segments phonemiques. Les resultats de ces etudes nous amenent a conclure que ces processus ne compliquent que rarement la tache de l'auditeur; parfois meme, ils la rendent plus facile.

1. INTRODUCTION

"Most of the speech we listeners hear is spoken spontaneously. [...] Psycholinguistics has compiled an extensive body of research on human speech recognition. [...But] the speech mode most frequently encountered in psycholinguistic experiments is *not* the speech mode most frequently encountered outside the laboratory". (Mehta and Cutler, 1988 [41], 135-136).

Of course, nothing has changed in the past ten years to affect the general validity of this statement. But by the same token, nothing has changed in the mission of psycholinguistic research: to account for human language processing in *all* its aspects. The recognition of spontaneous speech must be considered subject to psycholinguistic explanation.

Yet the few experiments which have used spontaneous speech materials have shown that such materials can produce effects that differ from those observed with read-speech materials (of the kind usually employed as stimuli in psycholinguistic experiments). Two studies using the gating task, for instance, contrasted spontaneously produced with read materials. Bard, Shillcock and Altmann [1] found that words in spontaneous speech could often not be recognised until a syllable or more of following context was available. McAllister [38] found that stressed syllables could be recognised earlier than unstressed syllables in speech that had been produced spontaneously, but in read speech there was no difference in how early the two types of syllable could be identified.

A phoneme detection study by Mehta and Cutler [41] similarly contrasted detection of targets in utterances which had either been produced spontaneously, versus read from a text (by the same speaker who had earlier spoken the utterances in conversation). Two effects long known from the literature to appear in phoneme detection (with read-speech materials) disappeared with the spontaneous utterances: Phoneme targets occurring later in utterances were detected more rapidly than targets occurring earlier, and targets preceded by long words were detected more rapidly than targets preceded by short words, in the read speech but not in the spontaneous speech. In contrast, two effects appeared with spontaneous speech that were not observed with read speech: Targets were detected more rapidly in accented than in unaccented words, and in stressed as opposed to unstressed syllables.

Mehta and Cutler explained their findings in terms of differences in prosodic structure across speech modes. Read speech encourages long smooth prosodic contours; spontaneous speech is characterised by much greater prosodic variation, more frequent pauses, and shorter prosodic units. Thus the different modes of speech can encourage listeners to rely on different listening strategies. There is no suggestion that listening differs in principle as a function of speech mode, of course. Note that Cutler and Butterfield [11] showed that the same patterns could be found in perceptual errors irrespective of whether the errors occurred as spontaneous slips of the ear, or were induced by difficult listening conditions imposed in the laboratory. What listeners may at most do is select, from the total repertoire of such strategies which they command, certain listening strategies more likely to be useful with a particular type of input. Listening (to the native language) is always efficient; different modes of speech may vary in the scope which they offer for particular ways of achieving this efficiency [31].

This suggests that everything we know from decades of psycholinguistic experimentation on listening to spoken language will be applicable to listening to spontaneous speech, wherever spontaneous speech offers scope for the relevant effects to appear. But there may also exist listening strategies and effects which are as yet undiscovered because only spontaneous speech offers scope for them to appear. That is, the type of speech materials used in psycholinguistic experiments may have offered no opportunity for discovery of part of the range of ways in which listening is robust and efficient.

In recent years, however, psycholinguists have begun to turn their attention to some of the phenomena that occur in spontanous speech and how they affect listening. This contribution summarises recent experimental evidence on

four phenomena which have caught the concern of psycholinguists in a number of countries and laboratories: resyllabification, assimilation, deletion, epenthesis.

First, however, section 2 presents an overview of the current framework within which spoken-word recognition research is conducted in Psycholinguistics; in this field, a real revolution has occurred in the past (not much more than) ten years.

2. THE RECOGNITION OF SPOKEN WORDS

Two developments revolutionised psycholinguistic modelling of spoken-word recognition in recent years: the availability of machine-readable vocabulary databases, and the introduction of connectionist modelling techniques into the field. The first meant that theoretical claims could be tested against the actual makeup of the vocabulary. In particular, the extent of inter-word similarity and overlap could be fully appreciated; every natural human language has a vocabulary running into the tens of thousands, but no language has a phonemic inventory running into even the low hundreds, so that the words of a language inevitably resemble one another strongly, and are often found to be embedded within one another. Some well-known theoretical positions were quickly abandoned when it turned out that the models involved would not actually work efficiently with real vocabularies.

The second development produced models which were computationally implemented and which could generate precise simulations of existing experimental findings, and predictions of results, with reference to specified vocabularies. Models which were not implemented quickly became less attractive.

Current models of human spoken-word recognition are thus all computational, and all able to draw to a greater or lesser extent on realistic vocabularies. What is more, they are all based on the same assumptions (for which there is now a very large amount of accumulated evidence): that candidate words which are compatible with the input are automatically activated and compete with one another for recognition. Models of this kind include TRACE [39]; Shortlist [44]; the Neighborhood Activation Model [32]; or the latest version of the Cohort model [22]. Of course there are great differences between the models in terms of their architecture. For instance, Shortlist is the only one with an architecture which allows tractable simulations to be run with a truly realistic vocabulary of tens of thousands of words. The models also differ notably in how they instantiate competition between words; in TRACE and Shortlist, competition is instantiated in terms of lateral inhibition between units at the same level. The input may activate any word compatible with part of it, so that words which overlap compete with one another for the same parts of the input. However, the more a given word is activated by the incoming speech, the more it is able to compete with - inhibit - other activated words; as words are inhibited, they lose activation and therefore become less able to inhibit other words. This process will eventually lead to only one successful competitor for each part of the input, and recognition of the string will have been achieved.

There is now abundant experimental evidence for concurrent activation of multiple word candidates, and for competition between simultaneously activated words (see [10] and [20], for recent review articles). Among the more recent studies, a number have addressed the question of how exactly the input needs to match stored canonical representations in order for word candidates to be activated. Words may be activated if they differ from the input by just a single feature (e.g. *dopic* can activate *topic;* [3,6,7,42]), and the greater the similarity of a nonword to a real word, the greater the activation of that word (thus *dopic* activates *topic* to a greater extent than *gopic* does, and *gopic* to a greater extent than *nopic* [8]). Some segments seem more likely than others to activate imperfectly matching candidates; thus listeners are more ready to alter vowels than consonants to turn a nonword into a real word [45].

On the other hand, coarticulatory information can be efficiently used to identify upcoming phonetic segments and hence to guide word candidates [18, 30, 35, 36, 52]; and mismatching coarticulatory information can significantly hamper recognition [34,40, 53, 54]. Words embedded within other words (as *can* in *candle*, for example), may be less effectively activated by their embedded form than by their citation form [13, 43].

From the accumulated evidence one can provisionally conclude that inexact forms can produce activation of stored lexical forms, but that the better the match in the input, the greater the activation. However, as pointed out in section 1, most psycholinguistic laboratory studies are carried out on carefully produced speech. The forms that produce maximum activation in psycholinguistic studies may be of a kind hardly ever encountered in natural-speech situations. The following sections discuss in more detail some recent research dealing with various phenomena applying to phonetic sequences in speech. Of course, in order to study these phenomena under carefully controlled laboratory conditions, and to separate their effects from the effects of possible confounding factors, the experiments that will be reviewed have all used: carefully-controlled laboratory speech! Nevertheless, they provide useful evidence which can in the long run be generalised to listening situations which are more typical of real life.

3. RESYLLABIFICATION

The first line of research concerns where segments occur in syllables, and whether the alignment of segments with syllable structure affects lexical activation. Consider, for instance, the syllable [kip], which to an English listener presumably activates the word *keep.* But what if the input is *keep it,* with the [p] uttered as onset of the second syllable? Is the changed allophonic form of the [p] sufficient to activate competitors such as *patrol* and *potato,* which might be excluded from competition if the [p] were initially perceived as syllable-final?

Such resyllabification of words - i.e. construction of syllables which cross word boundaries - is extremely common in very many languages. But whether or not it could be seen as complicating the process of lexical activation depends on one's assumptions about the stages involved in speech processing. In earlier psycholinguistic times the debates relevant to this issue concerned "units of perception". If it is assumed that speech input has to be translated into a representation in terms of syllables, and that this is the representation used for contacting the lexicon (as assumed for instance in [50]), then syllables which do not map neatly onto lexical units pose a problem for recognition. On the other hand, if it is assumed that input is translated into an abstract string of phonemes, whereby syllable-initial and syllable-final allophones of a given phoneme would be represented in the same form (as assumed for instance in [48]), then resyllabification need cause listeners no problem at all.

A recent study by Whalen, Best and Irwin [55] examined listeners' perception of allophones of [p] in American English words such as *opaque* and *soapy;* in the former the [p] is syllable-initial and aspirated, in the latter it is unaspirated (and arguably ambisyllabic). Listeners presented with these words containing the wrong allophones of the [p] and asked to repeat them tended to correct the [p] to the allophonic form appropriate for the word; in contrast, when given nonwords with the same stress patterns and the same appropriate or inappropriate allophones, listeners were much more accurate at repeating exactly what they had heard. Whalen et al. argue that the lexical activation so efficiently captured the input that the fine detail of the phonetic structure was inaccessible to the listeners' awareness.

Matter [37] and Dejean de la Batie and Bradley [14] investigated the rather more dramatic case of liaison phonemes in French - e.g. the final [t] of *petit* which is silent in *petit chien* but articulated in *petit arbre.* When it is articulated, it is also resyllabified. Both studies used the phoneme detection task. In Matter's experiment, listeners were asked to respond to occurrences of word-initial [a] (as in *arbre).* Neither native speakers nor non-native speakers showed a significant effect of the liaison; the /a/ of *arbre* could be detected as rapidly in *petit arbre* as in *joli arbre.* Dejean de la Batie and Bradley asked listeners to respond to word-initial *IM.* They found that responses to the initial phoneme of, for instance, *talent* were slower in *petit talent* than in *vrai talent,* presumably because the [t] could possibly have been the resyllabified final phoneme of *petit* attached to a vowel-initial following word. In other words, the activation of *petit* caused competition with *talent* for the [t]. They also found that non-native learners of French sometimes erroneously made responses for word-initial [t] in sequences like *petit arbre,* suggesting that in their case the activated form of *petit* was not efficiently competing for its final phoneme.

Matter [14] also conducted a further test of sensitivity to resyllabification, using the fragment detection method; listeners were asked to respond to the targets LA or LAR in, for instance, *la rente, l'arabe, la revue, largeur.* Only in the latter case did he find a significant effect: LAR was detected more rapidly than LA. Unfortunately his experimental design did not allow a comparison between detection of the same target in different word types. Interesting for the question of resyllabification, however, is the finding that both LA and LAR could be detected equally rapidly in both *la rente* and *l'arbre.*

All of these studies suggest that resyllabification does not cause great problems for the listener, which is perhaps as well given its prevalence across languages! However, a recent study by Vroomen and De Gelder [51] reports significantly longer phoneme detection latencies for resyllabified word-final phonemes in Dutch (in this study, listeners were asked to detect the phoneme targets anywhere in the word, not just in word-initial position). Thus detection of the [t] in *boot* was faster in *de boot die...* than in *de boot is...* ; the [t] in the latter string is resyllabified. When the same strings were presented as nonwords *(oot die* versus *oot is,* for example), the response time advantage reversed: detection was faster for syllable-initial phonemes i.e. [t] in *oot is.* Vroomen and De Gelder argue that the disadvantage for detection of resyllabified phonemes in words is therefore due to difficulty at the word recognition stage. The most obvious interpretation is that the resyllabified phoneme, together with its folowing context, activates more competing lexical candidates than the non-resyllabified version (with its context) does - in Dutch, there are many words beginning *ti-* but none beginning *td-.*

In summary, these findings probably do not illuminate the older questions of levels of representation during speech processing; but they do give a hint that listeners can constrain activation by accurate perception of the assignment of segments within syllable structure.

4. ASSIMILATION

The human articulatory system apparently finds rapid alteration of manner of articulation preferable to rapid alteration of place of articulation or rapid alteration of voicing. Across the languages of the world, it is far more common for two successive consonants with different manner of articulation to share place of articulation than to differ in place of articulation, and (to a lesser extent) to share voicing than to differ in voicing. Consider the case of nasal consonants preceding stop consonants. In English, such a sequence can occur within a syllable, in the coda position, and common place of articulation is obligatory if the syllable is morphologically simple: thus *lint, limp* and *link* all have different nasal consonants, and in each case the nasal has the same place of articulation as the stop consonant which follows it. By contrast, a mismatch in place of articulation produces an illegal sequence: **limt, *linp, *limk* etc.

In some languages this regressive assimilation of place for a nasal and a following stop consonant is obligatory

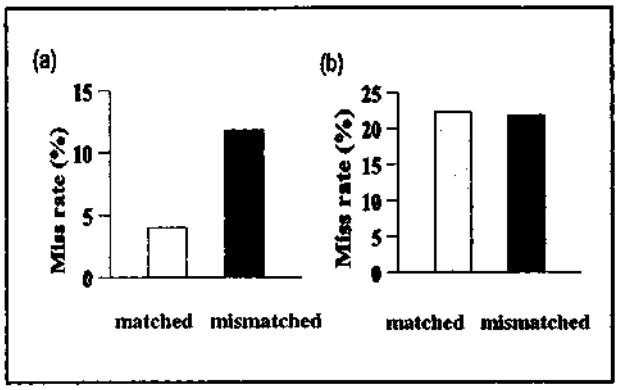


Figure 1. Mean percentage missed detections of phonemes in Japanese words with matching versus mismatching place of articulation of target phoneme with preceding nasal, for (a) Japanese listeners and (b) Dutch listeners.

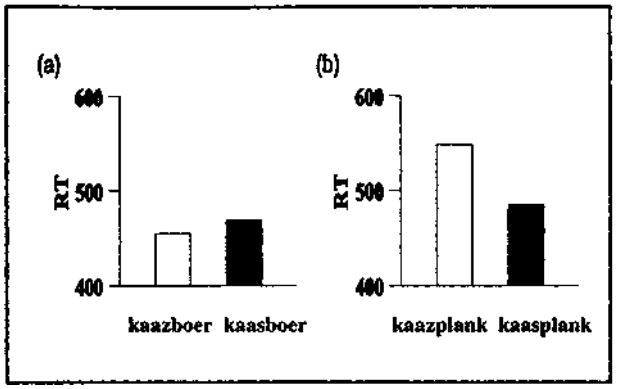


Figure 2. Dutch listeners' mean RT (ms) to detect target phoneme (a) with and without optional preceding voicing assimilation, and (b) with and without preceding voicing violating syllable-final devoicing constraint.

even across syllable boundaries. Japanese is one such language. Indeed, intrasyllabic environments do not even occur in Japanese, because the phonology does not allow nasal-stop clusters (or any other kind of cluster in syllable coda position, for that matter). Thus for your Japanese lunch you may have some *tempura* in your *bento* box, but *tenpura* in your *bemto* box is not possible. In the kana orthographies of Japanese, mere is only a single representation for any nasal consonant in syllabic coda position; the representation is unmarked for place, in other words, though it will differ in place as a function of the place of any following consonant. Consider, for example, the word *san,* three'. It occurs in many compound words, for example: *sangatsu* (March'); *sanban* (third'); *sanju* (thirty*). In the first of these the final nasal of the first syllable is pronounced as a velar, in the second as a bilabial, in the third as a dental-alveolar; yet of course the first part of each compound is the same element, and is identically represented in both Japanese orthographic forms, kanji (Chinese characters) and kana (mora-based phonological representation).

In some other languages, however, regressive assimilation of place in nasal-stop sequences is not obligatory. English allows failure of assimilation across some syllable boundaries, for instance in compounds such as *songbird, sometimes* and *sunglasses,* and even within morphologically complex syllables such as *jammed* and *longed.* The situation in German and Dutch is even less constrained. Morphologically simple words with unassimilated codas exist - e.g. *Hemd/hemd* ('shirt") or *Fremd/vreemd* ('strange'), along with unassimilated morphologically complex syllables such as *schwimmt/ zwemt* ('swims') or *singt/zingt* ('sings'). Similarly unassimilated sequences can occur across syllable boundaries in morphologically simple words *(Imker/imker,* 'beekeeper'), derived or inflected words *(Raumte/ruimte* 'space') or compound words *(Blumenkohl/bloemkool* 'cauliflower', *Rennbahn/renbaan* 'racecourse').

There have been a relatively large number of experimental studies in recent years in which the effects of assimilation processes on word processing have been investigated, via phoneme detection or word recognition tasks. These studies, of which some have been carried out in our laboratory, have shown a highly consistent pattern of results: spoken-language processing is neither facilitated nor interfered with by optional regressive assimilations, but is inhibited by violations of obligatory regressive assimilation.

Thus for speakers of English or Dutch, in which most assimilations are optional, speed of detection of a phoneme target is unaffected by whether or not an immediately preceding consonant is assimilated to the target. This was shown in quite similar phoneme detection studies by Koster [26], Otake, Yoneyama, Cutler and Van der Lugt [46], Gaskell and Marslen-Wilson [23], and Kuijpers and Van Donselaar [28]. Some results of the Otake et al. study are shown in Figure 1, and of Kuijpers and Van Donselaar's study in Figure 2. Dutch listeners, for whom assimilation of place in nasal-stop sequences is optional, as described above, showed no significant difference in detection rate for stop consonant targets preceded by nasals matched versus unmatched in place of articulation (Figure lb). Likewise, Dutch allows optional voicing assimilation across obstruent sequences; thus the word *kaas* normally ends with /s/, but in *kaasboer* ('cheesemonger") the following segment *Ibl* can cause the final segment of kaas to be voiced, giving *kaazboer.* Dutch listeners again show no effect of this optional assimilation in phoneme detection; responses to *Ibl* in *kaasboer* and *kaazboer* are not significantly different (Figure 2a).

On the other hand, when a target is preceded by a violation of assimilation - either by an unassimilated phoneme in Japanese, where assimilation is always obligatory, or by an "assimilated" phoneme in English or Dutch which does not match its following context - detection is significantly slowed. Again this result has been demonstrated in a number of separate studies [23,28, 46]. Otake et al. [46] found that detection of a nasal phoneme in coda position in Japanese is insensitive to assimilation, i.e. equally correct irrespective of its place of articulation realisation; but detection of the following stop

is highly sensitive to the assimilation, in that it is significantly worse if the place of articulation of the nasal does not match the place of articulation of the stop (Figure 1a). Dutch listeners likewise detect the /p/ in *kaasplank* more rapidly when the word is spoken *kaasplank* than when it is spoken *kaazplank* (Figure 2b); the latter form is not an assimilation environment, so that alteration of the voicing at the end of *kaas,* though it had no effect when it was a legal assimilation, causes a violation here and as a consequence interferes with word processing.

Recognition of spoken words is likewise unaffected by optional assimilation but impaired by violation of obligatory assimilation. Marslen-Wilson, Nix and Gaskell [33]), and Gaskell and Marslen-Wilson [21] found that assimilated word-final phonemes did not interfere with word activation and priming effects, as long as the assimilation was legal; exactly the same changes, however, when appled in an environment which was inappropriate for the relevant assimilation, led to significant inhibition of word recognition.

Effects on activation were also observed in the study of Koster [26], who found that detection of final consonants in English or Dutch words such as *line/lijn* was slower if they were assimilated to a following bilabial stop such that the spoken form of the word was identical to another existing word *(lime/lijm).*

In an attempt to investigate the effects of assimilations on the representations which listeners form of spoken words, Cutler and Otake [12] compared the processing of assimilated forms by speakers of Japanese and Dutch. They used a task in which listeners produced blended forms of pseudo-word pairs such as *ranga-serupa* (in the Japanese materials set) or *mengkoop-trabeek* (in the Dutch materials set). Assimilated blends of these pairs would be *rampa* and *membeek,* unassimilated blends *rangpa* and *mengbeek.* Although the actual response measure involves speech production, the produced response forms offer a window on listeners' representations of the perceived stimulus forms.

Both Japanese and Dutch listeners were asked to do the task, with their native-language materials and also with the foreign-language materials. The results are shown in Figure 3. The pattern was absolutely clear: the Japanese listeners produced many more assimilated than unassimilated forms, both with pseudo-Japanese and pseudo-Dutch materials, while Dutch listeners produced more unassimilated than assimilated forms in each materials set. In all cases statistical analysis showed that the differences were highly significant; indeed for the Japanese they were separately significant for every one, and for Dutch for all but one, of the six types of assimilation tested: bilabial to alveolar or to velar, alveolar to bilabial or velar, velar to bilabial or alveolar (the above examples involve velar-to-bilabial assimilation).

These results suggest that Japanese listeners, whose native-language phonology involves obligatory assimilation constraints, represent the assimilated nasals in nasal-stop sequences as unmarked for place of articulation,

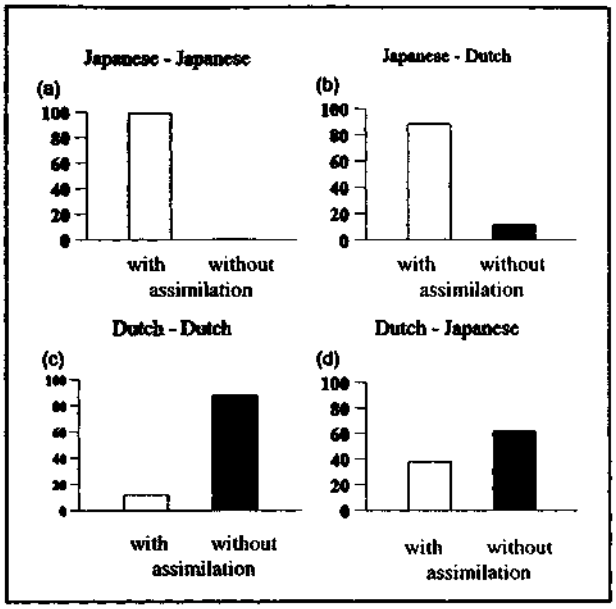


Figure 3. Mean percentage of (scorable) blend responses with and without assimilated nasal-stop sequences, for Japanese listeners listening to (a) Japanese and (b) Dutch, and for Dutch listeners listening to (c) Dutch and (d) Japanese.

while Dutch listeners, who are accustomed to hearing unassimilated forms represent the same nasal segments as marked for place of articulation. The full consequences of these differing forms of perceptual representation for lexical activation have, again, as yet to be investigated.

5. DELETION

Typical pronunciations of words in natural speech frequently appear to contain fewer segments, or even fewer syllables, than the careful pronunciations of the same words in isolation. Thus it is rare for English words like *family* or *government* to be pronounced with three syllables; instead, the medial weak syllable (the vowel in the case of *family,* the VC in the case of *government)* are essentially deleted, and the spoken word is bisyllabic. Likewise, French words like *galerie* and *calepin* are pronounced with two syllables; Dutch and German words like *referaat/Referat* and *veteraan/Veteran* can meet the same fate; in each case it is the medial syllable of the canonical form which is reduced out of metrical existence. Similarly, weak initial syllables (as in French *semaine,* English *support,* Dutch *beleid,* German *beraten)* can lose their vowel, reducing the surface form of the utterance by one syllable.

In such surface forms, there may well be residual effects of the canonical form in the surface pronunciation; Browman and Goldstein [5], for instance, argue that apparent deletions are in fact the effect of gestures overlapping rather than being omitted. Thus the medial consonantal sequences would be articulatorily different in English/am *'ly* versus *hamlet,* French *cal'pin* versus *palper,* Dutch and German *ref'raat/Ref'rat* versus *saffraan/Safran.*

Beckman [2]) discusses such syllable reduction phenomena in a number of phonologically differing languages: English and German, Montreal French, Tokyo Japanese, Korean. Beckman argues that a universal mechanism of gradient phonetic reduction is at work in all these languages; the human articulatory system treats all languages similarly. Nevertheless, Beckman argues, the effect of this process must be evaluated within the overall phonological structure of each language. In English and in German, syllabic reanalysis (e.g. of *support* in English, or of *beraten* in German) would yield words one syllable shorter in length, beginning with consonant clusters; such forms (as the existence in these cases of the minimally differing words *sport,* in English, and *braten,* in German, attests) are permissible for the languages in question.

On the other hand, the application of the same process to the Japanese word for *sport (supootu)* would produce the consonant cluster which is found in the English word upon which it is based; but such clusters are illegal in Japanese. Accordingly, phoneticians describe the process of vowel deletion in Japanese, as in the first syllable of *supootu,* as devoicing rather than as deletion leading to syllabic reanalysis. Beckman argues that it is indeed appropriate to view what is essentially the same process in different ways in different languages, because the process interacts with language-specific phonological structure. Thus in Japanese, devoiced syllables maintain a temporal contribution, preserving the mora-based rhythm of the language. In English and in German (bom stress-timed languages), deletion processes as observed in support and beraten do not affect stressed syllables, hence they leave the stress rhythm intact - the number of stress units does not change when a weak syllable is deleted. In syllable-timed languages like French and Korean, as Beckman shows, the effect of such processes depends on a given syllable's position in larger rhythmic groupings.

Certainly deletions and reductions are highly sensitive to the rhythmic context; slips of the tongue involving a deleted segment or syllable tend to result in an utterance which is rhythmically more regular than the intended utterance would have been [9]; and optional deletions in Dutch words such as *kinderen* (in which the medial vowel may delete, leading to a strong-weak bisyllabic form are more likely to occur in contexts which a strong-weak alternating rhythmic pattern [27].

From the point of view of word recognition, the question raised by such reduction phenomena is their effect on the process of activation of stored forms. There have been no experimental investigations so far of whether listeners are sensitive to details of the phonetic realisation of segment sequences (as for instance in *fam'ly* versus *hamlet,* or *s pport* versus *sport).* But there have been a number of simple investigations of whether deletion and reduction processes lead to word recognition difficulty.

Many of these experiments have been conducted with French-language materials. Thus Matter [37] included in his series of studies a number of deletion and reduction effects. For instance, he found that the personal pronoun TE was recognised more rapidly in a context such as *te partem,* in which it is spoken as a full syllable (and, as he pointed out, would also be written in full) than in *t'a parte,* in which its spoken (and written) form corresponds to a single consonant. He found that elision which deleted the final sound of a word had no effect on detection of the initial sound (thus responses to /p/ were equally rapid in *pretre a Lyon,* in which the final vowel of *pretre* is elided, than in *pretre du Sacre Coeur,* in which no elision occurs). Elision at word onset, however, had a significant effect on word-initial phoneme detection; thus responses to /r/ were significantly slower in *deux records,* where the vowel following /r/ is deleted, than in *sept records,* where the vowel is not deleted.

Similarly, Racine and Grosjean [49] found that both lexical decision and word repetition responses were significantly slower for words (again in French) with reduction of the initial syllable (e.g. *semaine* spoken as *s'maine).* Peretz, Lussier and Beland [47] studied French words like *calepin* in a memory retrieval task. In this task, retrieval prompts which correspond to a syllable of the word are more effective than prompts which are larger or smaller than a syllable. Thus CAL is a better prompt than CA for *calmant* (first syllable *cal-),* and CA is a better prompt than CA for *calorie* (first syllable *ca-).* Words like *calepin* behaved as if their first syllable was *cal-* not *ca-,* suggesting that the memory representations were based on the surface form pronunciation.

All of these studies suggest that deletion processes do affect the way words are recognised and represented in memory, especially if the deletion occurs near the onset of the word. Even in word-medial position, deletion certainly does not help recognition. Thus Kuijpers, Van Donselaar and Cutler [29] found that lexical decision responses to Dutch words like *referaat* were faster and more accurate when the words were presented in their canonical trisyllabic form than in the form with reduction of the medial syllable. Listeners presented with reduced forms such as *sport* are able to accept them as potentially either *sport* or *support* [19]. Presumably the set of activated word candidates differs as a function of the precise nature of the incoming signal; sensitive studies of how this process is affected by deletion and reduction are therefore called for.

6. EPENTHESIS

A different kind of adjustment to the phonological form of an uttered word consists in the *addition* of a phonetic segment. Again this can happen for a number of reasons.

Some languages rule out consonant clusters or obstruent codas, so that when foreign words are used in native speech any such illegal sequences will be made to conform to the native constraints, usually by having vowels inserted; to Japanese listeners then, there is no difference between the name *MacDonalds* uttered by an American speaker and by a Japanese speaker, although the latter will say *Makudonarudo.* Several recent experiments have converged on the conclusion that the representations

which Japanese listeners extract from a spoken signal are in conformity with the native phonology even if the spoken input itself is not. Thus Kashino and colleagues[24, 25] compared Dutch and Japanese listeners in a consonant (cluster) identification task. Dutch and Japanese listeners were equally accurate in identifying intervocalic consonants in VCV stimuli. Dutch listeners, however, were significantly more accurate in identifying the consonants of VC1C2V stimuli than Japanese listeners, suggesting that the latter found it difficult to perceive consecutive consonants.

Similarly, Dupoux, Kakehi, Hirose, Pallier and Mehler [17] showed that Japanese listeners perceived vowels between consonants in non-words, even if these vowels had been deleted from the speech signal and were not physically present; sequences such as *ebzo* and *ebuzo* were reported by Japanese listeners as identical. In contrast, French listeners had no difficulty in distinguishing between these two types of string. Dupoux et al. showed that it was not the case that the Japanese listeners simply listened less carefully; lengthening of the medial vowel in the same sequences (e.g. a contrast between *ebuzo* and *ebuuzo)* was easily discriminated by Japanese listeners but caused great difficulty for French listeners. In a more recent study by the same authors (personal communication), impossible words in Japanese such as *komgi* and *namda* were presented in a lexical decision task. There is a real Japanese word *komugi,* but no real word *natnuda.* YES responses to the real word komugi and to its impossible variant *komgi* were statistically indistinguishable; likewise, NO responses to the nonword *namuda* and to its impossible variant *namda* were equivalent. This finding suggests that the phonological knowledge is brought into play before word candidates are lexically activated.

Other cases of intruding segments, however, do not result from constraints of the phonology. Unassimilated nasal-stop sequences across syllable boundaries in languages such as English (see section 4 above) can trigger intrusive segments with the place of articulation of the nasal - an intrusive [p] in *something* or *dreamt,* for instance, or a [k] in *gangster* or *wingfeather.* There is as yet no research on the role of such intrusions in lexical activation. Nor do we know whether activation is affected by intrusive linking segments between vowels - for instance the [r] that can be inserted at the word boundary in *idea over,* for instance, of the [j] that can be inserted at the boundary in *high amplitude.* Do such segments result in unwanted lexical candidates becoming activated, even if only momentarily (e.g. *yam* in *high amplitude, row* and *rover* in *idea over)?* Or can English listeners compensate for these effects at an earlier stage of processing?

Other epenthetic segments have however been the focus of research attention in recent years. In our laboratory we have conducted a series of studies on the processing of Dutch words with epenthetic vowels. In Dutch, words like *tulp* 'tulip', *werk* 'work' and *film* 'film' are routinely pronounced *tulep, werek* and *filem* respectively.

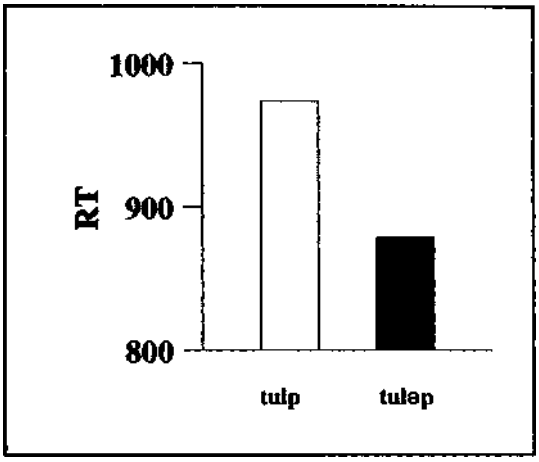


Figure 4. Dutch listeners' mean lexical decision RT (ms) to words without and with epenthetic vowel.

Such forms are as acceptable as the variants without epenthesis. The epenthesis most often affects coda clusters consisting of liquids followed by another consonant other than /s/ or /t/. The epenthesis is not forced by constraints of the phonology, since clusters are acceptable, indeed common, in Dutch words in both onset and coda position. There is no other obvious pressure to avoid clusters; for example, nicknames and other word formation processes in Dutch do not eschew them - thus someone named *Marcus* can be known as *Marc,* or *Nicolaas* can be *Klaas,* and someone with the function of *direkteur* ('director') may be referred to as the *dirk.*

Because of this, mere has been considerable discussion of why Dutch vowel epenthesis occurs. One obvious explanation is that it can facilitate articulatory ease. Liquids followed by non-coronal obstruents form heterogenic clusters since the members of the cluster do not share place of articulation. From a phonetic point of view it is quite natural that heterorganic consonant clusters are broken up since they require more articulatory effort than homorganic clusters [4]. The articulatory ease account also explains why schwa epenthesis in the context of a liquid and non-coronal obstruent seems to be practically standard in child Dutch [56]. An articulatory aspect is further suggested by the fact that the frequency of epenthetic insertion varies with rhythmic context. Thus just as slips of the tongue involving syllable insertion are more likely to result in an utterance which is rhythmically more regular than the intended utterance would have been [9], so is epenthesis more likely when it results in a rhythmically more regular output [27]. Against this argument however is still the general acceptability of non-coronal clusters in Dutch, and the fact that the optional variant with epenthesis simply co-exists with, rather than threatens, the non-epenthesised form. In simple production tasks with words which allow epenthesis, it is applied about 50% of the time [27].

In our own studies, however, we were concerned with the consequences of epenthesis for the listener. Surely,

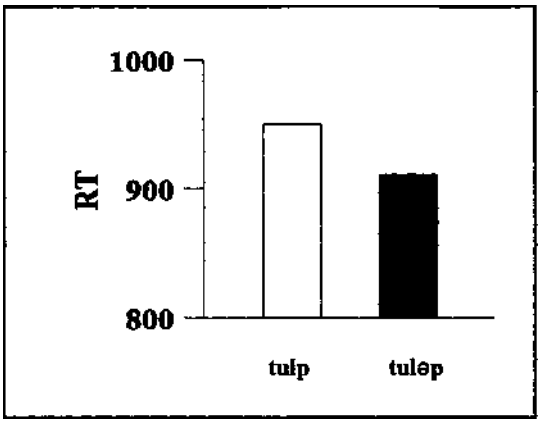


Figure 5. Dutch listeners' mean RT (ms) to spot embedded words without and with epenthetic vowel.

one might imagine, changing the pronunciation of a word, whether or not it makes life easier for the speaker, cannot be in the best interests of the listener. We first asked whether the two variant forms seemed to activate the same stored representation; after all, separate storage for the two optional forms could perhaps be a simple way to ease problems of processing. One way to address the question of whether access representations are separate is via the question of syllabicity. Obviously, adding a vowel between two consonants adds an extra syllable to the word; the optional form with epenthesis has one more syllable than the underlying form without. If access representations are separate, one will have one syllable more than the other.

In our first experiment on this topic [15], we used the fragment detection task. We reasoned that a difference in response time to detect, for instance, the target TUL in *tulp* versus *tulap* would indicate that different phonological representations were activated by the different inputs. In two detection experiments, we found no difference in RTs as a function of the input form, which strongly suggests no difference in the phonological representation accessed by the input *tulp* and the input *tulep.*

In a second study [16], we used a syllable reversal task. Listeners were asked to reverse words, either by saying the phonemes in reverse order (if the word was monosyllabic; thus *hot* 'cat' became *tak),* or by saying the syllables in reverse order (if the word was bisyllabic, so that *diner* would become *ner-di).* The question then became whether the input forms *tulp* and *tulap* would be treated equivalently; evidence for a single phonological representation for both would be provided by parity of subjects' responses given each input form. Indeed, this was again what happened; subjects overwhelmingly selected the monosyllabic response option, so that both *tulp* and *tulap* became *plut.*

The second issue we addressed was whether the variant forms, with the epenthetic vowel, differed from the canonical forms without epenthesis in how hard they were to process; i.e. given that both forms appeared to activate the same representation, did one form activate it more rapidly than the other? First, we simply compared the two forms in auditory lexical decision [16,29]. Figure 4 shows a representative result; in all our lexical decision experiments, we actually found that YES responses to the words like *tulp* were faster when these words were presented in the form with an epenthetic vowel *(tulep)* than in their canonical monosyllabic form *(tulp).*

We therefore followed this experiment up using the word-spotting task, which measures how rapidly a spoken word can be perceived in a (minimal) context. Van Donselaar et al.[16] presented words like *tulp* with a following context that made the whole utterance into a nonsense bisyllable: *tulpfoos* or *tulepfoos.* The listeners' task was simply to respond whenever they detected the occurrence of any real word in the input; they did not know in advance what words might occur, and there were many filler strings with no words in them. Once again, as Figure 5 shows, responses were significantly faster when the words had been presented with vowel epenthesis.

Why should this be so? We suggest that that the perceptual effect of vowel epenthesis in liquid-obstruent clusters is simply to make the liquid more easily perceptible. We tested this explanation by conducting a final experiment using the phoneme detection task (Van Donselaar et al., submitted). The materials used in the word-spotting experiment were presented to listeners who were asked to detect *HI* (e.g. in *tulp/tulep)* or /r/ (e.g. in *werk/werek).* As predicted, the targets followed by an epenthetic vowel were detected significantly more rapidly than the targets followed directly by the consonant. Whatever the advantages of vowel epenthesis may be for the speaker, therefore, epenthesis is also advantageous to the listener: it simply renders the perception of certain segments easier.

7. CONCLUSION

As studies of the perception of nonwords have shown, lexical activation does not necessarily require a fully specified, acoustically distinct input representation. This is just as well, given that spontaneous speech frequently presents listeners with indistinct and partially unspecified signals. The studies reviewed here have provided evidence that listening is rarely adversely affected by various types of phonological effects - resyllabification, assimilation, deletion, epenthesis - often encountered in natural speech. Adverse effects do arise from violations of phonological legality (e.g. assimilations in inappropriate environments, or failure of obligatory effects of one kind or another); but such violations of course do not occur in natural speech.

Only when phonological variation has the effect of producing a signal which maps onto unintended alternative words, and hence allows activation of spurious word candidates, can one talk of an adverse effect for the listener; but although such spurious activations are presumably responsible for the adverse effects which have been demonstrated in a number of experiments, e.g. on deletions near the beginnings of words, or in certain resyllabifications, explicit investigations of this suggestion

using laboratory tasks which assess lexical activation and competition have not yet been undertaken.

Far from variant phonological forms having as a rule an adverse effect for the listener, in fact, some such forms clearly make the listener's task easier; thus epenthetic vowels which break up consonant clusters render the consonant which precedes them more easily perceptible. The speech mode most frequently encountered outside the laboratory, in other words, may often cause the listener less trouble than the kinds of speech used in a typical psycholinguistic experiment.

8. ACKNOWLEDGEMENTS

Thanks to my collaborators in investigation of the perceptual effects of phonological variation: Takashi Otake, Cecile Kuijpers, Wilma van Donselaar.

9. REFERENCES

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

[2] M. Beckman, "When is a syllable not a syllable?" In T. Otake and A. Cutler (Eds.), *Phonological Structure and Language Processing: Cross-Linguistic Studies*, pp. 95-123. Mouton de Gruyter, Berlin/New York, 1996.

[3] J. Boelte, "The role of mismatching information in spoken word recognition", PhD Thesis, Freie Universitat Berlin, Berlin, 1996.

[4] G. Booij, "The Phonology of Dutch", Clarendon Press, Oxford, 1995.

[5] C.P. Browman and L. Goldstein, "Tiers in articulatory phonology, with some implications for casual speech" In J. Kingston and M.E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pp. 341-376, Cambridge University Press, Cambridge, 1990.

[6] CM. Connine, D.G. Blasko and D. Titone, "Do the beginnings of spoken words have a special status in auditory word recognition?", *Journal of Memory and Language*, Vol. 32, pp. 193-210,1993.

[7] CM. Connine, D.G. Blasko and J. Wang, "Vertical similarity in spoken word recognition: Multiple lexical activation, individual differences, and the role of sentence context", *Perception and Psychophysics*, Vol. 56, pp. 624-636, 1994.

[8] CM. Connine, D. Titone, T. Deelman, and D. Blasko, "Similarity mapping in spoken word recognition", *Journal of Memory and Language*, Vol. 37, pp. 463-480, 1997.

[9] A. Cutler, "Syllable omission errors and isochrony", In H.W. Dechert and M. Raupach (Eds.), *Temporal Variables in Speech*, pp. 183-190, Mouton, The Hague, 1980.

[10] A. Cutler, "Spoken word recognition and production", In J.L. Miller and P.D. Eimas (Eds.), *Speech, Language and Communication*, Vol. 11 of E.C Carterette and M.P. Friedman (Eds.) *Handbook of Perception and Cognition*, pp. 97-136, Academic Press, NY, 1995.

[11] A. Cutler and S. Butterfield, "Rhythmic cues to speech segmentation: Evidence from juncture misperception", *Journal of Memory and Language*, Vol. 31, pp. 218-236, 1992.

[12] A. Cutler and T. Otake, "Assimilation of place in Japanese and Dutch", Proc. of the Fifth International Conference on Spoken Language Processing, Sydney, December (in press).

[13] Davis, W.D. Marslen-Wilson and M.G. Gaskell, "Ambiguity and competition in lexical segmentation", 1997 (manuscript).

[14] B. Dejean de la Batie and D. C. Bradley, "Resolving word boundaries in spoken French: Native and non-native strategies", *Applied Psycholinguistics*, Vol. 16, pp. 59-81, 1995.

[15] W. van Donselaar, C. Kuijpers and A. Cutler, "How do Dutch listeners process words with epenthetic schwa?", Proc. of the Fourth International Conference on Spoken Language Processing, pp. 149-152, Philadelphia, 1996.

[16] W. van Donselaar, C Kuijpers and A. Cutler, "Facilitatory effects of vowel epenthesis on word processing in Dutch", *Journal of Memory and Language*, to appear.

[17] E. Dupoux, K. Kakehi, K. Hirose, C. Pallier and J. Mehler "Epenthetic vowels in Japanese: A perceptual illusion", submitted.

[18] L. Ellis, A.J. Derbyshire and M.E. Joseph, "Perception of electronically gated speech", *Language and Speech*, **Vol.** 14, pp. 229-240, 1971.

[19]J. Fokes and Z. S. Bond, "The elusive/illusive syllable", *Journal of Phonetics*, Vol. 50, pp. 102-123, 1993.

[20] U.H. Frauenfelder and C Floccia, "The recognition of spoken words", In A. Friederici (Ed.), *Language Comprehension: A Biological Perspective*, pp. 1-40, Springer, Heidelberg, 1998.

[21] M.G. Gaskell and W.D. Marslen-Wilson, "Phonological variation and inference in lexical access", *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 22, pp. 144-158, 1996.

[22] M.G. Gaskell and W.D. Marslen-Wilson, "Integrating form and meaning: A distributed model of speech perception", *Language and Cognitive Processes*, 1997.

[23] M.G. Gaskell and W.D. Marslen-Wilson, "Mechanisms of phonological inference in speech perception", *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 24, pp. 380-396,1998.

[24] K. Kakehi, K. Kato and M. Kashino, "Phoneme/syllable perception and the temporal structure of speech", In T. Otake and A. Cutler (Eds.), *Phonological Structure and Language Processing: Cross-Linguistic Studies*, Mouton de Gruyter; Berlin/NewYork, pp. 125-143, 1996.

[25] M. Kashino, A. van Wieringen and L. Pols, "Cross-language differences in the identification of

intervocalic stop consonants by Japanese and Dutch listeners", Proc. of the Second International Conference on Spoken Language Processing, pp. 1079-1082 Banff, Canada, 1992.
[26] C.J. Koster, "Word Recognition in Foreign and Native Language", Foris, Dordrecht, 1987.
[27] C. Kuijpers and W. van Donselaar, "The influence of rhythmic context on schwa epenthesis and schwa deletion in Dutch", *Language and Speech,* Vol. 41, pp. 87-108, 1998.
[28] C. Kuijpers and W. van Donselaar, "Phonological variation and phoneme identification in Dutch", forthcoming.
[29] C. Kuijpers and W. van Donselaar and A. Cutler, "Phonological variation: Epenthesis and deletion of schwa in Dutch", Proc. of the Fourth International Conference on Spoken Language Processing, pp. 94-97, Philadelphia.
[30] A. Lahiri and W.D. Marslen-Wilson, "The mental representation of lexical form: A phonological approach to the recognition lexicon", *Cognition,* Vol. 38, pp. 245-294.
[31] B. Lindblom, "Phonetic invariance and the adaptive nature of speech", In B.A.G. Elsendoorn and H. Bouma (Eds.), *Working Models of Human Perception,* pp. 139-173, Academic Press, London, 1988.
[32] P.A. Luce, D.B. Pisoni and S.D. Goldinger, "Similarity neighborhoods of spoken words", In G.T.M. Altmann (Ed.), *Cognitive Models of Speech Processing,* pp. 122-147, MIT Press, Cambridge, MA, 1990.
[33] W.D. Marslen-Wilson, A. Nix and M.G. Gaskell, "Phonological variation in lexical access: Abstractness, inference and English place assimilation," *Language and Cognitive Processes,* Vol. 10, pp. 285-308, 1998.
[34] W.D. Marslen-Wilson and P. Warren, "Levels of perceptual representation and process in lexical access: Words, phonemes, and features", *Psychological Review,* Vol. 101, pp. 653-675, 1994.
[35] J.G. Martin and H.T. Bunnell, "Perception of anticipatory coarticulation effects", *Journal of the Acoustical Society of America,* Vol. 69, pp. 559-567,1981.
[36] J.G. Martin and H.T. Bunnell, "Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences", *Journal of Experimental Psychology: Human Perception and Performance,* Vol. 8, pp. 473-488, 1982.
[37] J.F. Matter, "A la recherche des frontieres perdues", PhD Thesis, University of Utrecht, 1986.
[38] J. McAllister, "The processing of lexically stressed syllables in read and spontaneous speech", *Language and Speech,* Vol. 34, pp. 1-26, 1991.
[39] J.L. McClelland and J.L. Elman, "The TRACE model of speech perception", *Cognitive Psychology,* Vol. 18, pp. 1-86, 1986.
[40] J.M. McQueen, D.G. Norris and A. Cutler, "Lexical influence in phonetic decision-making: Evidence from subcategorical mismatches", *Journal of Experimental Psychology: Human Perception and Performance,* to appear.
[41] G. Mehta and A. Cutler, "Detection of target phonemes in spontaneous and read speech", *Language and Speech,* Vol. 31, pp. 135-156, 1988.
[42] W. Milberg, S. Blumstein and B. Dworetzky, "Phonological factors in lexical access: Evidence from an auditory lexical decision task", *Bulletin of the Psychonomic Society,* Vol. 26, pp. 305-308, 1988.
[43] C.B. Mills, "Effects of the match between listener expectancies and coarticulatory cues on the perception of speech", *Journal of Experimental Psychology: Human Perception and Performance,* Vol. 6, pp. 528-535.
[44] D. G. Norris, "Shortlist: A connectionist model of continuous speech recognition", *Cognition,* Vol. 52, pp. 189-234, 1994.
[45] B. van Ooijen, "Vowel mutability and lexical selection in English: Evidence from a word reconstruction task", *Memory and Cognition,* Vol. 24, pp. 573-583,1996.
[46] T. Otake, K. Yoneyama, A. Cutler and A. van der Lugt, "The representation of Japanese moraic nasals", *Journal of theAcoustical Society ofAmerica,* Vol. 100, pp. 3831-3842,1996.
[47] I. Peretz, I. Lussier and R. Beland, "The roles of phonological and orthographic code in word stem completion", In T. Otake and A. Cutler (Eds.), *Phonological Structure and Language Processing: Cross-Linguistic Studies,* pp. 217-226, Mouton de Gruyter, Berlin, 1996.
[48] D.B. Pisoni and P.A. Luce, "Acoustic-phonetic representations in word recognition", *Cognition,* Vol. 25, pp. 21-52,1987.
[49] I. Racine and F. Grosjean, "La reconnaissance des mots en parole continue: Effacement du schwa et frontiere lexicale", Actes des Journees d'Etudes Linguistiques, Nantes, 1997.
[50] J. Segui, "The syllable: A basic perceptual unit in speech processing?", In H. Bouma and D.G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes,* pp. 165-181, Erlbaum, Hillsdale, NJ, 1984.
[51] J. Vroomen and B. de Gelder, "Lexical access of resyllabified words: Evidence from phoneme monitoring", *Memory and Cognition,* in press.
[52] P. Warren and W.D. Marslen-Wilson, "Cues to lexical choice: Discriminating place and voice", *Perception and Psychophysics,* Vol. 43, pp. 21-30, 1988.
[53] D.H. Whalen, "Subcategorical mismatches slow phonetic judgments", *Perception and Psychophysics,* Vol. 35, pp. 49-64, 1984.
[54] D.H. Whalen, "Subcategorical phonetic mismatches and lexical access", *Perception and Psychophysics,* Vol. 35, pp. 49-64, 1991.
[55] D.H. Whalen, C.T. Best and J.R. Irwin, "Lexical effects in the perception and production of American English /p/ allophones", *Journal of Phonetics,* Vol. 25, pp. 501-528, 1997.
[56] F. Wijnen, E. Krikhaar, and E. den Os, "The (non)realization of unstressed elements in children's utterances: evidence for a rhythmic constraint", *Journal of Child Language,* Vol. 21, pp. 59-83,1994.