Reduction and assimilatory processes in conversational French speech: Implications for speech synthesis

Danielle Duez

Laboratoire Parole et Langage, CNRS ESA 6057, Aix en Provence, France
email: duez@lpl.univ-aix.fr; fax: (33) 04 42 59 50 96
Introduction
Speakers adaptively tune phonetic gestures to the various needs of speaking situations (Lindblom, 1990). For example, in informal speech styles such as conversations, speakers speak fast and hypoarticulate, decreasing the duration and amplitude of phonetic gestures and increasing their temporal overlap. At the acoustic level, hypoarticulation is reflected by a higher reduction and context-dependence of speech segments: Segments are often reduced, altered, omitted, or combined with other segments compared to the same read words.

Hypoarticulation does not affect speech segments in a uniform way: It is ruled by a certain number of linguistic factors such as the phonetic properties of speech segments, their immediate context, their position within syllables and words, and by lexical properties such as word stress or word novelty. Fundamentally, it is governed by the necessity for the speaker to produce an auditory signal which possesses sufficient discriminatory power for successful word recognition and communication (Lindblom, 1990).

Therefore the investigation of reduction and contextual assimilation processes in conversational speech should allow us to gain a better understanding of the basic principles that govern them. In particular, it should allow us to find answers to the questions as why certain modifications occur and others do not, and why they take particular directions. The implications would be of great interest for the improvement of speech synthesis. It is admitted that current speech-synthesis systems are principally able to generate highly intelligible output. However, there are still difficulties with naturalness of synthetic speech, which is strongly dependent on contextual assimilation and reduction modelling (Hess, 1995). In particular, it is crucial for synthesis quality and naturalness to manipulate speech segments in the right manner and at the right place.

This paper is organised as follows. Section 1 summarises perceptual and spectrographic data obtained for aspects of assimilation and reduction in oral vowels (Duez, 1991), voiced stops (Duez, 1995) and consonant sequences (Duez, 1998) in conversational speech. Reduction means here a process in which a consonant or a vowel is modified in the direction of lesser constriction or weaker articulation, such as a stop becoming an affricate or fricative, or a fricative becoming a sonorant, or a close vowel becoming more open. Assimilation refers to a process that increases the similarity between two adjacent (or next-to-adjacent) segments. Section 2 deals with the interaction of reduction and assimilatory processes with factors such as the phonetic properties of speech sounds, immediate adjacent context (vocalic and consonantial), word class (grammatical or lexical), position in syllables and words (initial, medial or final), position in phrases (final or non-final). Section 3 summarises some reduction-and-assimilation tendencies. Section 4 raises some problems of how to integrate reduction and contextual assimilation in order to improve naturalness of speech-synthesis and proposes a certain number of rules derived from results on reduction and assimilation.

Reduction and contextual assimilation

Vowels

Measurements of the second formant measured in CV syllables occurring in conversational speech and read speech showed that the difference in formant frequency between the CV boundary (locus) and the vowel nucleus (measured in the middle of the vowel) was smaller in conversational speech. The frequency change was also found to be greater for the nucleus than for the locus. Moreover, loci and nuclei did not change in the same direction. The results were interpreted as reflecting differences in coarticulation, both an anticipatory effect of the
subsequent vowel on the preceding consonant, and/or formant undershoot (as defined by Lindblom, 1963).

**Voiced stops.**

Perceptual and acoustic data on voiced stops extracted from the conversational speech produced by two speakers revealed two consistent tendencies: 1) There was a partial or complete nasalisation of /b/’s and /d/’s in a nasal vowel context, that is, a preceding and/or a succeeding nasal vowel: at the articulatory level, there was the velum-lowering gesture partially or totally overlapped with the closing gesture (for an illustration of complete nasalisation, see the following example)

<table>
<thead>
<tr>
<th>Phonicological</th>
<th>Identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>/πA</td>
<td>δA/</td>
</tr>
</tbody>
</table>

2) There was a weakening of /b/ into the corresponding approximant fricative /B/, semivowel /w/ and approximant (labial) and the weakening of /d/ into the corresponding fricative /z/, sonorant /l/, approximant /dental/, or its complete deletion. These changes were assumed to be the result of a reduction in the magnitude of the closure gesture. The deletion of the consonant was viewed as reflecting the complete deletion of the closure gesture. Interestingly, assimilated or reduced consonants tended to keep their place of articulation, suggesting that place of articulation is one of the consonantal invariants.

**Consonant sequences**

A high number of heterosyllabic [C1#C2] and homosyllabic [C1C2] consonant sequences were different from their phonological counterparts. In most cases, C1’s were changed into another consonant or omitted. Voiced or unvoiced fricatives and occlusives were devoiced or voiced, reflecting the anticipatory effect of an unvoiced or voiced C2. Voiced or unvoiced occlusives were nasalized when preceded by a nasal vowel, suggesting a total overlapping of the velum-lowering gesture of the nasal vowel with the closure gesture. Similar patterns were observed for a few C2’s. There were also some C1’s and C2’s with only one or two features identified: Voicing, devoicing and nasalisation were incomplete, reflecting partial contextual assimilation. Other consonants, especially sonorants, were omitted, which may be the result an extreme reduction process. An illustration of C1 omission can be seen in the following example.

<table>
<thead>
<tr>
<th>Phonicological</th>
<th>Identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>/λμΕτάλειο/</td>
<td>/μΕτάλειο/</td>
</tr>
</tbody>
</table>

In some cases, there was a reciprocal assimilation of C1 to C2. It was particularly obvious in C1C2’s, where the manner and place features of C1 coalesced the voicing feature of C2 to give a single consonant (/sd/=>/zl/, /fz/=>/zl/, /sv/=>/zl/, /fz/=>/zl/, /tv/=>/dl/). An illustration can be found in the following example.
Thus, two main trends in assimilation characterised consonant sequences: 1) Assimilation of $C_1$ and $C_2$ to nasal vowel context and 2) voicing assimilation of $C_1$ to $C_2$, and/or $C_2$ to $C_1$. In all cases, $C_1$ and $C_2$ tended each to keep their place of articulation.

**Factors limiting reduction-and-assimilation effects**

**Segment properties and adjacent segments.**

Vowels as well as consonants underwent different degrees of reduction and assimilation. The loci and the nuclei of the front vowels were lowered, while those of the back vowels were raised, and there was little change for vowels with mid-frequency $F_2$. Nucleus-frequency differences exhibited greater changes for back vowels than for front vowels, for labials as well as for dentals. Data obtained for voiced stops revealed a higher identification rate for dentals than for labials, suggesting that the former resist reduction and assimilatory effects more than the latter. This finding may be due to the fact that the degree of freedom is greater for the lips than the tongue which is submitted to a wide range of constraints. Consonant sequences also revealed a different behaviour for the different consonant types. Omitted consonants were mostly sonorants. Moreover, differences were observed within a same category. The omitted sonorants were /l, ® or m/, those reported as different were /n/ changed into /m/ before /p/.

The above findings suggest a lesser degree of resistance to reduction and assimilatory effects for sonorants than for occlusives and fricatives. Sonorants are consonants with a formantic structure: They are easily changed into vowels or completely deleted. Similarly, voiced occlusives are less resistant than unvoiced occlusives which have more articulatory force (Delattre, 1966). The resistance of speech-segments to the influence of reduction and contextual assimilation should be investigated in various languages: The segments which resist more are probably those which in turn exert a stronger influence on their neighbours.

**Syllable structure and position in a syllable.**

Mean identification scores were higher for homosyllabic $C_1C_2$’s than for heterosyllabic ones. The highest identification scores were for sequences consisting of a fricative plus a sonorant, the lowest scores for sequences composed of two occlusives. In heterosyllabic sequences, the $C_1$’s not equal to their phonological counterparts were mostly in coda position. Moreover, in $C_1C_2$-onset sequences there was a slight tendency for $C_2$’s to be identified as a different consonant. The data suggest a stronger resistance of onset-speech segments, which is in total conformity with the results found for articulatory strength (Straka, 1964). Moreover, onset segments have a higher signalling value for a listener in word recognition.

**Word class**

Word class had no significant effect on the identification of voiced plosives, but a significant effect on the identification of $C_1$’s in consonant sequences. Grammatical words did not react in the same way to the influence of reduction and assimilatory processes. For example, the elided article or preposition (de =/>/d∅/) was often omitted in $C_1#C_2$’s as $C_1$ as well as $C_2$. It was also
often changed into an /h/ when it was an intervocalic voiced stop preceded by a nasal vowel. On the opposite, in phrases consisting of je /Ζ/ (personal pronoun) + verb (lexical word), the /Ζ/ was maintained while the first consonant of the verb was mostly reported as omitted, or at least changed into another consonant.


\[ \text{et je vais te dire} \text{ (and I am going to tell you)} \]

Phonological \[ /ΕΖΕτ} ← δι®/ \]

Identified \[ /ΕΖΕτ} ← δι®/ \]

Final prominence

In French, the rhythmic pattern of utterances mainly relies on the prominence given to final syllables at the edge of a breath group (Vaissière, 1991). As final prominence is largely signalled by lengthening, final-phrase syllables tend to be long, compared to non-final phrase syllables. Phrase-final segments resist the influence of reduction and assimilatory processes which are partly dependent on duration (Lindblom, 1963). Prominent syllables showed a larger formant excursion from the locus to the nucleus than non-prominent ones. Voiced plosives and consonant sequences perceived as phonological were located within prominent syllables.

Tendencies in reduction and assimilation

Natural speech production is a balance between an articulatory-effort economy on the part of the speaker and the ability to perceive and understand on the part of the listener. These two principles operate, at different degrees, in all languages, in any discourse and everywhere in the discourse, within syllables, words, phrases and utterances. Thus, the acoustic structure of the speech signal is characterised by a continuous succession of (more or less) overlapping and reduced segments, the degree and the extent of overlapping and reduction being dependent on speech style and information. Reduction and assimilatory processes are universal since they reflect basic articulatory mechanisms, but they are also language-dependent to the extent that they are ruled by phonological and prosodic structures of languages. Interestingly, the regularities observed here suggest some tendencies in reduction and contextual assimilation specific to French.

Nasalisation

There is a universal tendency for nasality to spread from one segment to another, although the details vary greatly from one language to another and nasalisation is a complex process that operates in different stages. For example, the normal path of emergence of distinctive nasal vowels begins with the non-distinctive nasalisation of vowels next to consonants. This stage is followed by the loss of the nasal consonants and the persistence of vowel nasalisation, which therefore becomes distinctive (Ferguson, 1963; Greenberg, 1966). Interestingly, investigations of patterns of nasalisation in modern French revealed different nasalisation-timing patterns (Duez, 1995; Ohala and Ohala, 1991) and nasalisation degrees depending on consonant permeability (Ohala and Ohala, 1991). The fact that nasal vowels may partially or completely nasalise adjacent occlusives has implications for speech synthesis since sequences containing voiced or unvoiced occlusives preceded by a nasal vowel are frequent in common adverbs and numbers.

C2 dominance
In languages such as French, the peak of intensity coincides with the vowel while in some other languages, it occurs earlier in the syllable and tends to remain constant. In the first case, the following consonant tends to be weak and may drop while in the other case, it tends to be reinforced. This characteristic partly explains the evolution of French (for example, the loss of the nasal consonant in the process of nasalisation) and the predominance of CV syllables (Delattre, 1969). It also gives an explanation to the strong tendency for occlusive or fricative C1’s to be voiced or devoiced under the anticipatory effect of a subsequent unvoiced or voiced occlusive or fricative, and for sonorants to be vocalised or omitted.

Resistance of prominent final-phrase syllables

In French, prominent syllables are components of a hierarchical prosodic structure, and boundary markers. They are information points which predominantly attract the listener's attention (Hutzen, 1959), important landmarks which impose a cadence on the listener for integrating information (Vaissière, 1991). They are crucial for word recognition (Gee and Grosjean, 1987) and the segmentation of the speech stream into hierarchical syntactic and discourse units. Thus, the crucial role of the prominence pattern in speech perception and production may account for its effect on the reduction and contextual assimilation of speech segments.

Implications for speech synthesis

The fact that speech production is at a same time governed by an effort-economy principle and perceptual needs has crucial implications for speech-synthesis. Perceived naturalness has proven to strongly depend on the fit to natural speech, listeners being responsive to an incredible number of acoustic details and performing best when the synthesis contains all known regularities (Klatt, 1987). As a consequence, the improvement of synthetic naturalness at the segmental level requires detailed acoustic information, which implies in turn a fine-grained knowledge of linguistic processes operating at different levels in the speech hierarchy, and in particular a good understanding of reduction and assimilation processes in languages.

Concatenation-based synthesisers

There are two types of synthesisers: formant and spectral-domain synthesisers, and concatenation-based synthesisers. Concatenation-based synthesisers are based on the concatenation of natural speech units of various sizes (diphones, demi-syllables, syllables and words) recorded from a human speaker. They present a certain number of advantages and disadvantages mainly related to the size of units. For example, small units such as diphones and demisyllables do not need much memory but do not contain all the necessary information on assimilation and reduction phenomena. Diphones which are units extending from the central point of the steady part of a phone to the central point of the following phone contain information on consonant/vowel and vowel-consonant transitions but do not cover coarticulation effects in consonant sequences. In contrast, demisyllables which result from the division of a syllable into an initial and a final demisyllable (Fujimura, 1976) cover most coarticulation effects in onset and coda consonant sequences actually present in words but not in sequences resulting from the elision of an optional /œ/. Systems based on the concatenation of larger units such as syllables and words (Lewis, and Tatham, 1999; Stöber, Portele, Wagner and Hess, 1999) solve some of the above problems since they contain many coarticulatory and
reduction effects. However, they also need to be context-knowledge based. For example, Lewis and Tatham (1999) described how syllables have to be modified for concatenation in contexts other than those from which they were excised. Stöber, Portele, Wagner and Hess (1999) proposed a system using words possessing the inherent prosodic features and the right pronunciation. In concatenation-based systems, the quality and naturalness of synthesis require the selection of appropriate concatenation units or word instances in the right contexts, which implies the knowledge of regularities in reduction and assimilatory processes. In French consonant sequences, the assimilation of an occlusive to a preceding nasal vowel was shown to depend on its location within syllables (final or initial) and its membership in either homosyllabic or heterosyllabic sequences. Coarticulatory effects were also found to be limited by final prominence. Thus, different timing patterns of nasalisation can be obtained for occlusives by integrating in the corpus different instances of the same syllables or words produced in both phrase-final and phrase-internal positions. Similarly, the problem of grammatical words which tend to sound “too loud and too long” (Hess, 1995) can be solved by recording different instances of these words in different contexts. This procedure should be particularly useful for grammatical words whose prominence depends on their location within phrases. The personal pronoun (il=>ιλ) may be clearly articulated in phrase-final position, on the opposite, the /l/ is deleted when /ιλ/ is followed by a verb, that is, in phrase-internal position. In the latter case, it constitutes with the verb a single prosodic word. Some verbal phrases consisting of the personal pronoun (je /Ζ↔+ verb) were also shown to present considerable and complex reduction. In some cases there was elision of /ζ↔/ and assimilation of voicing of /ζ/ to the following consonant. In other cases, there was deletion of /ζ↔/ and partial or complete reduction of the verb-initial consonant. As verbal phrases are frequently used, different instances as a function of context and styles might be added in the corpus.

Rule-based synthesis: Rules of reduction and assimilation

In formant and spectral-domain synthesisers where the generation of the acoustic signal is derived from a set of segmental rules which model the steady state properties of phoneme realisation and control the fusion of strings of phonemes into connected speech (Klatt, 1987), output can be improved (at least partly) by the incorporation of reduction and contextual-assimilation rules in the text-to-speech system. For example, the present results suggest that we should include the following rules for consonants located in non-prominent syllables: (1) rules of nasalisation for voiced intervocalic occlusives followed and/or preceded by a nasal vowel, and for unvoiced and voiced syllable-final plosives preceded by a nasal vowel and followed by another consonant; (2) rules of devoicing or voicing for voiced or unvoiced syllable-final obstruents before an unvoiced or voiced syllable-initial obstruent; (3) rules of vocalisation for syllable-final sonorants in heterosyllabic sequences and (4) rules of deletion of /l/ into personal pronouns.

An illustration can be found in the following tentative rules. The formalism of these rules follows that of Kohler (1990) and has the following characteristics. Rules are of the form X => Y / W ___ Z where X is rewritten Y after left-hand context W and before right-hand context Z, respectively. In the absence of Y, there is a deletion rule. Each symbol is composed of a phonetic segment, V and C for vowels and consonants, respectively, and # for syllable boundary. Vowels and consonants are defined as a function of binary features. +/- FUNC means function/non-function word marker. As assimilated, reduced and omitted consonants were mostly located into non-prominent syllables, the feature /-PROM/ is not represented.
- Change of intervocalic voiced plosives into their nasal counterparts after a nasal vowel

\[
\begin{bmatrix}
  \text{C} \\
  \text{− nas} \\
  \text{+ voice} \\
  \text{+ occl}
\end{bmatrix} \Rightarrow \begin{bmatrix}
  + \text{nas} \\
  V
\end{bmatrix}
\]

-Nasalisation of voiced-or-unvoiced-stops before any syllable-initial consonant

\[
\begin{bmatrix}
  \text{C} \\
  \text{− nas} \\
  \text{+ occl}
\end{bmatrix} \Rightarrow \begin{bmatrix}
  + \text{nas} \\
  V
\end{bmatrix}
\]

-Voicing of unvoiced obstruents before syllable-initial voiced obstruents

\[
\begin{bmatrix}
  \text{C} \\
  \text{− voice} \\
  \text{+ obst}
\end{bmatrix} \Rightarrow \begin{bmatrix}
  + \text{voice} \\
  V
\end{bmatrix}
\]

-Devoicing of voiced obstruents before unvoiced syllable-initial obstruents

\[
\begin{bmatrix}
  \text{C} \\
  \text{+ voice} \\
  \text{+ obst}
\end{bmatrix} \Rightarrow \begin{bmatrix}
  \text{− voice} \\
  V
\end{bmatrix}
\]

-Vocalisation of sonorants before any syllable-initial consonant

\[
\begin{bmatrix}
  \text{C} \\
  \text{+ son}
\end{bmatrix} \Rightarrow \begin{bmatrix}
  V
\end{bmatrix}
\]

-Deletion of /l/ in the function word /il/ before any syllable-initial consonant

\[
\begin{bmatrix}
  \text{L} \\
  \text{+ func}
\end{bmatrix} \Rightarrow \begin{bmatrix}
  \varnothing
\end{bmatrix}
\]

References


