Synthesis of Spoken Messages from Semantic Representations
(Semantic-Representation-to-Speech System)

Laurence DANLOS, Eric LAPORTE

Laboratoire d'Automatique Documentaire et Linguistique
Université Paris 7
2, place Jussieu
75251 PARIS CEDEX 05

Françoise EMERARD

Centre National d'Etudes des Télécommunications
22301 LANNION CEDEX

Abstract

A semantic-representation-to-speech system communicates orally the information given in a semantic representation. Such a system must integrate a text generation module, a phonetic conversion module, a prosodic module and a speech synthesizer. We will see how the syntactic information elaborated by the text generation module is used for both phonetic conversion and prosody, so as to produce the data that must be supplied to the speech synthesizer, namely a phonetic chain including prosodic information.

Introduction

A spoken message can be produced either to utter a written text (text-to-speech system), or to communicate orally the information given in a semantic representation (semantic-representation-to-speech system). In both cases, the speech synthesizer must be provided with a phonetic chain including prosodic information in order to reconstruct the acoustic signal. As we will recall in 1., syntactic knowledge is necessary to compute the phonetic transcription of a written text and to include prosodic information in it. Hence a text-to-speech system must include a parsing module to get this syntactic knowledge. On the other hand, a semantic-representation-to-speech system can take advantage of the syntactic information elaborated when expressing the semantic representation in natural language. Therefore, we design a semantic-representation-to-speech system that generates directly from the semantic representation a phonetic string with prosodic markers, without a written stage. Our system has been designed for French but it could be extended to other languages.

1. Knowledge needed in a text-to-speech system

1.1. Spelling-to-sound conversion

The first problem encountered in synthesizing speech from written text is that of spelling-to-sound conversion. Certain languages are much easier than others in this respect. For example, about 50 rules are sufficient for the conversion of written Spanish into phonetic symbols, with a virtually zero error rate (Santos & Nombela 1982). For other languages, such as French or English, the problem is much greater. A phoneme does not generally correspond to only one grapheme, and the reverse is also true. For instance, the word *oiseau* is pronounced /wazo/ : none of its graphemes is pronounced as would be expected (i.e. /o/ for *o*, /i/ for *i*, /s/ for *s*, schwa for *e*, /a/ for *a* and /y/ for *u*).

Spelling-to-sound conversion is further complicated by the existence of non-homophonic homographs, i.e. words spelled the same but pronounced differently. The distinction between two homographs requires to know their grammatical categories (*record* is pronounced ['reko:d] if it is a noun and [ri'ko:d] if it is a verb), their inflexional features (*read* is pronounced [ri:d] in the infinitive form and [red] in the preterite), or their semantic features (*lead* is pronounced [led] when it is a noun or a verb related to the metal and [li:d] otherwise)[1].

In French, words in context raise the additional problem of liaison. A liaison occurs between a word ending in a mute consonant and a word beginning with a vowel. For example, the *n* in *mon* is pronounced in *mon arrivée* but mute in *mon départ*. However, a liaison is made only if this phonological condition is accompanied with syntactic conditions. For example, a liaison is made between a determiner and a noun as in *mon arrivée* (my arrival), but not between a subject and a verb as in *Le limon arrive* (The silt is coming).

[1] In French, semantic features are needed to distinguish only a few non-homophonic homographs, mostly technical words.

To sum up, the phonetic conversion of French texts relies on syntactic knowledge to deal with homographs and liaisons.

1.2. Prosody

A text-to-speech system supposes the storage of minimum acoustic units that allow the reconstruction of the acoustic signal for any sentence. One solution consists in the choice of diphones as acoustic units. A diphone is defined as a segment (about 1,200 for French) that goes from the steady state of a phonetic segment to the steady state of the following segment and that contains in its heart all the transitional part between two consecutive sounds.

Furthermore, the issue of increasing the naturalness of synthetic speech requires to take into account prosodic factors, namely, stress, timing (structuring of the utterance by pauses) and intonation. Intonation is characterized by the interaction of three parameters: evolution of intensity and laryngeal frequency as functions of duration.

The prosodic behavior of one speaker was therefore subjected to a systematic study. An acceptable model was extracted from this behaviour. The prosodic processing (Emerard 1977) is based on the allocation of prosodic markers (e.g. [=], [#]) at different points in a sentence. Fifteen prosodic markers were considered to be sufficient for determining suitable prosodic contours for the synthesis of French. Each marker assigns a melody and a rhythm to each syllable of the preceding word. More precisely, each marker may
- cause an interruption in the diphone concatenation,
- introduce a pause,
- affect to varying degrees the amplitude of laryngeal frequency (F_0) on the last vowel of the word,
- determine rising or falling F_0 movements.

The choice of a marker after a constituent is determined both by the syntactic category of the constituent (verbal syntagm, subordinate clause) and by its location inside the sentence, especially by the existence of a more or less complex right context. In the simple enunciative sentence *Jean part* (John is leaving), the prosodic processing has to give the following results: *Jean* [#] *part* [.] . Nevertheless, it is not possible to conclude with the following prosodic rules :
[#] is the marker assigned to [end of subject noun phrase]
[.] is the marker assigned to [end of verbal syntagm]
because in the enunciative sentence *Jean part et Marie vient* (John is going away and Mary is coming), the prosodic processing has to propose: *Jean* [=] *part* [,] *et Marie* [#] *vient* [.] . A comparison of these two sentences clearly shows that it is not possible to assign a specific marker after a constituent only on the basis of its syntactic category. It is necessary to take its right context into account. Moreover, placing prosodic markers must be carried out in a hierarchical manner. For example, the marker between the preverbal phrase and the verbal syntagm depends on the marker assigned at the end of the clause containing them; this last marker depends in turn on the marker assigned at the end of the sentence containing the clause.

To sum up, the issue of prosody is handled by placing appropriate markers in appropriate locations. This can only be done when precise syntactic information is available.

2. Production of a phonetic chain with prosodic markers

The system which translates a semantic representation into a phonetic chain with prosodic markers has been built from a written text generation system (Danlos 1986) that has been modified and completed. Let us start with a brief description of this generator.

2.1. The generator

The generator is modularized into a strategic component and a syntactic component. From a semantic representation such as

```
(1) EVENT : ACT =: GIVE-PRESENT
              ACTOR = HUM1 =: HUMAN
                          NAME = Jean
              OBJECT = TOK1 =: FLOWER
                          TYPE = anémone
              DATIVE = HUM2  =: HUMAN
                          NAME  = Marie
          GOAL = : HAPPY
              OBJECT  =  HUM2
```

the strategic component makes conceptual decisions (e.g. the decision about the order of the informations) and linguistic decisions (e.g. the decision about the number of sentences) (Danlos 1984 a and b). The output of this component is a "text template" (TT) that indicates
1) the splitting up of the text into sentences:
TT = (Sentence1. Sentence 2.)

2) for each sentence, its structure in terms of main clause and subordinate clauses:

```
Sentence1 = (Clause1 (SUB (CONJ pour que)
                           Sentence3))
Sentence3 = Clause2
```

3) for each clause, its main verb with its complementation:

```
Clause1 = ((SUBJECT HUM1) (VERB offrir)
          (OBJECT TOK1) (A-OBJECT HUM2))
Clause2 = ((SUBJECT HUM1) (VERB rendre)
          (OBJECT HUM2) (ATTRIBUTE heureux))
```

A text template is turned into a text by the syntactic component. This component applies grammar rules (e.g. reduction of a subordinate clause to an infinitive form), synthesizes the tokens and performs the morphological routines. For these operations to be carried out, a text template includes, for each sentence, syntactic information that is represented in a tree whose nodes are syntactic categories such as S (sentence), CL (clause), SUBJECT or VERB. A text template may be made up of several sentences, however we will give an example with a single sentence because the operations of phonetic conversion and entering prosodic markers are performed within a sentence, independently of the other sentences. From the semantic representation **(1)**, the text template may be:

```
(2) ((S (CL (SUBJECT HUM1) (VERB offrir)
           (OBJECT TOK1) (A-OBJECT HUM2))
       (SUB (CONJ pour que)
            (S (CL (SUBJECT HUM1) (VERB rendre)
                   (OBJECT HUM2)
                   (ATTRIBUTE heureux))))) .)
```

The syntactic component turns it into a tree whose leaves are words :

```
((S (CL (SUBJECT (NP (N Jean)))
        (VERB  a offert)
        (OBJECT (NP (DET des) (N anémones)))
        (A-OBJECT  (NP (PREP à) (N Marie))))
    (SUB  (S (CL (CONJ pour) (PPV la)
                 (VERB rendre)
                 (ATTRIBUTE heureuse))))) .)
```

The erasing of the auxiliary vocabulary leads to:

Jean a offert des anémones à Marie pour la rendre heureuse.
(John offered anemones to Mary to make her happy.)

The syntactic component contains a morphological module (Courtois 1984) that works out an inflected form (e.g. *heureuse,* the feminine singular of *heureux)* given a basic form (e.g. *heureux)* and inflexional features (e.g. feminine, singular). This module is based on a dictionary that indicates an inflexion mode for each basic form. Each inflexion mode is associated with a rule that computes inflected forms.

The only modification made to the text generation system was to replace the morphological module with a morpho-phonetic module that proceeds to both inflexion and spelling-to-sound conversion. With this modification, the syntactic component produces a tree whose leaves are phonetic words.

2.2. Inflexion and phonetic conversion

A French morpho-phonetic system has been built to compute an inflected phonetic form given an orthographic basic word and inflexional features (Laporte 1986). This system uses an intermediate phonological representation devised to optimize not only word inflexion and phonetic conversion but also liaison processing. The system works in the following way: given a basic orthographic form (e.g. *heureux),* its syntactic category and inflexional features (e.g. adjective, feminine, singular), a phonological dictionary works out its phonological representation (e.g. *ørøz).* The word is then inflected (e.g. *ørøz')* by means of a set of rules. These rules for phonological inflexion are much simpler than those that would be required for inflecting orthographic or phonetic words. By way of illustration, the feminine of the following adjectives: *bon, grand, gros, léger, petit, pris, sot, vu* can be obtained from their phonological representation with only 1 rule, whereas 3 would be required when starting from their orthographic representation and 8 from their phonetic representation (Laporte 1984). The shift from phonological words to phonetic words entails knowing where liaisons should take place. Recall that a liaison takes place when both syntactic and phonological conditions are satisfied. In the semantic-representation-to-speech system, the syntactic tree of the sentence allows us to place liaison markers at the points where a liaison is syntactically allowed. The conversion of phonological words into phonetic words is then performed by a set of straightforward rules that check the phonological conditions of liaisons at the points where a liaison marker is present. Laporte's system is represented in Fig. 1.

From the text template (2), the syntactic component with the morpho-phonetic module outputs the following tree:

	Orthographical representation		Phonological representation		Phonetic representation
Basic words	50,000	⟶ dictionary	50,000 ↓ rules		50,000
Inflected words	350,000		150,000	⟶ rules	150,000

Fig. 1.

```
(3) ((S (CL (SUBJECT (NP (N žă)))
          (VERB a ofɛʀ)
          (OBJECT (NP (DET de) (N zanemɔn)))
          (A-OBJECT (NP (PREP a) (N maʀi)))
      (SUB (S (CL (CONJ puʀ) (PPV la) (VERB ʀădʀ)
                (ATTRIBUTE øʀøz))))) .)
```

All the segmental phenomena have been taken into account and the next operation consists in entering prosodic markers in such a tree.

2.3. The prosodic component[2]

Our prosodic system is based on syntax. However, there is not an isomorphic relation between the syntax and the prosody of a sentence. For example, the syntactic structures of *Jean est parti à Paris* (John went to Paris) and *Il est parti à Paris* (He went to Paris) are nearly identical, whereas there is a prosodic marker after the noun *Jean* and none after the pronoun *Il*. Conversely, the syntactic representations of *Jean a parlé de ce problème à Marie* (John spoke about this problem to Mary) et *Jean a parlé de ce problème à Paris* (John spoke about this problem in Paris) are different although their prosodic markers are identical. As a consequence, we had to build a complete syntactico-prosodic grammar for French[3]. This grammar enables us to obtain a structure of a sentence that is isomorphic to its prosodic structure and computable from its syntactic structure. A syntactico-prosodic category corresponds

- either to a syntactic category (e.g. the syntactico-prosodic category S is equivalent to the syntactic category S),
- or to a sequence of syntactic categories (e.g. the prosodic category POV [post-verbal phrase] groups together all the complements which appear after the verbal syntagm [VS], and the prosodic category PRV [pre-verbal phrase] groups all the complements which appear before the VS),
- or to several syntactic categories (e.g. the prosodic category VC [verbal complement] corresponds to the following syntactic categories: SUBJECT, OBJECT, A-OBJECT and ATTRIBUTE).

The first operation performed in the prosodic component thus consists in transforming the syntactic tree produced by the syntactic component into a syntactico-prosodic tree. From **(3)**, this operation produces the following tree, in which the leaves are written in spelling representation for readability:

```
(4) ((S (CL (PRV (VC (NP (N Jean))))
           (VS a offert)
           (POV (CV (NP (DET des) (N anémones)))
                (CV (NP (PREP à) (N Marie)))))
        (SUB (S (CL (CONJ pour) (VS la rendre)
                    (VC heureuse))))) .)
```

Besides the syntactico-prosodic grammar, a function SEG-C has been designed for each syntactico-prosodic category C. Such a function takes two arguments: a constituent [X] of the category C and the prosodic marker x that is to appear to the right of [X]. It computes the prosodic markers that have to be entered in [X]. More precisely, if the syntactico-prosodic analysis of [X] is:

$$[X] = ([X_1]\ [X_2]\ \ldots\ [X_n])$$

then:

$$(\text{SEG-C}\ [X]\ x) = ([X_1]\ x_1\ [X_2]\ x_2 \ldots [X_{n-1}]\ x_{n-1}\ [X_n]\ x)$$

where x_1, x_2, ... x_{n-1} are the appropriate markers. As an illustration, the grammar lays down that

$$[\text{CL}] = (\text{CL}\ [\text{CONJ}]\ \text{I}\ \ [\text{PRV}]\ \text{I}\ \ [\text{VS}]\ \ [\text{POV}]\ \text{I}\ )$$

where the sign "I" following an element means that the element is either absent or present once. The function

[2] This work was supported by CNET under contract no. 857B068 with LADL.

[3] This solution was also considered by Martin (1979).

(SEG-CL [CL] x) indicates that

- when [PRV] is present, a marker f(x) must be entered after it;
- when [POV] is present, a marker g(x) must be entered after [VS];
- in any case, x is after the last constituent, i.e. [POV] when present, [VS] otherwise.

The algorithm for entering the markers works in a recursive manner by means of a function SEG. Given a constituent [X] and the marker x that is to appear to the right of [X], this function figures out the category C of [X] and calls (SEG-C [X] x). Next, the functions

$$(\text{SEG-C}_1\ [X_1]\ x_1),\quad (\text{SEG-C}_2\ [X_2]\ x_2), \ldots\quad (\text{SEG-C}_n\ [X_n]\ x)$$

are called. For example, after (SEG-CL [CL] x) has been called, the entering of the markers into [PRV] when present is executed by

$$(\text{SEG}\ [\text{PRV}]\ f(x)) = (\text{SEG-PRV}\ [\text{PRV}]\ f(x)).$$

When [POV] is present, the functions (SEG [VS] g(x)) and (SEG [POV] x) are called, otherwise the function (SEG [VS] x) is called. The function SEG is first applied to the root of the arborescent syntactico-prosodic structure of the sentence involved and to its final punctuation mark ("." "," "?" ";" ":") which corresponds to a prosodic marker. When the recursion is over, the auxiliary vocabulary is erased, leaving a phonetic chain with prosodic markers. As an example, the function SEG applied to **(4)** leads to the following result:

(5) žã [=] a ofɛʀ [$] de zanemɔn [=] a maʀi [,] puʀ la
ʀãdʀ [$] øʀøz [.]

(Jean [=] a offert [$] des anémones [=] à Marie [,] pour la rendre [$] heureuse [.])

3. Algorithm and results

The phonetic chain with prosodic markers produced by the system are forwarded to the speech synthesizer developed at CNET (Courbon & Emerard 1982). The chart in Fig. 2 depicts the whole algorithm for generating spoken messages from semantic representations.

An implementation of the system has been developed in COMMON-LISP in the domain of terrorism crime newspaper reports. It produces phonetic chains with prosodic markers such as the ones shown below. Again, orthographic words replace phonetic symbols for readability. The syntactic conditioning of liaisons is

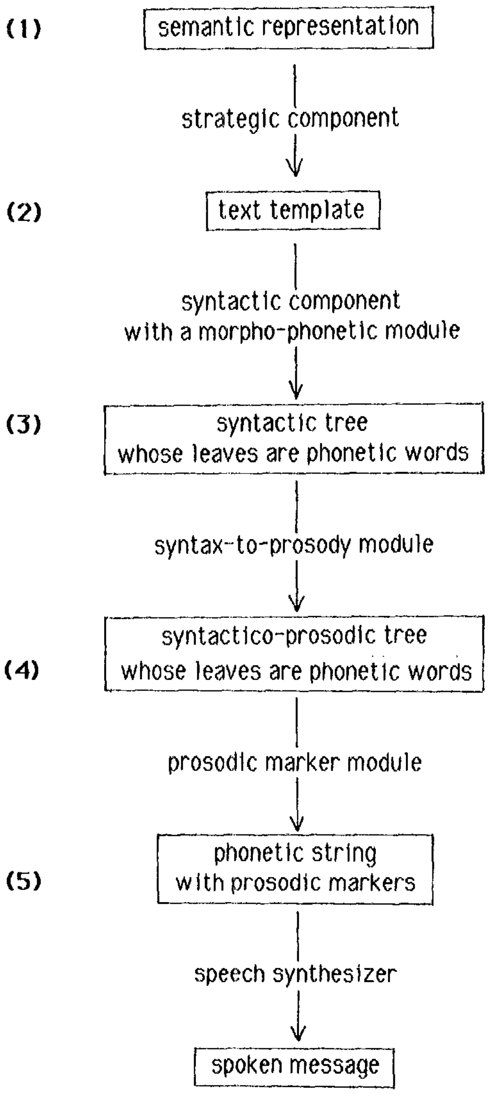

Fig. 2

marked with the sign [-]. We present three syntactically different versions of the same terrorism crime to emphasize the robustness of the syntactic component and the entering of appropriate prosodic markers according to syntax.

Version 1. *Indira Gandhi* [#] *a été assassinée* [$] *mercredi à New-Dehli* [.] *Des* [-] *extrémistes sikhs* [=] *ont tiré* [@] *sur* [-] *le premier ministre indien* [,] *alors que* [-] *elle* [-] *partait* [$] *de* [-] *son domicile* [=] *à pied* [*] *pour se rendre* [$] *à* [-] *son bureau* [.]
(Indira Gandhi was assassinated Wednesday in New-Dehli. Sikh extremists shot the Indian Prime Minister as she was leaving her home on foot to go to her office.)

Version 2. *Des* [-] *extrémistes sikhs* [#] *ont assassiné* [@] *Indira Gandhi* [*] *mercredi à New-Dehli* [.] *Ils* [-] *ont tiré* [@] *sur* [-] *le premier ministre indien* [,] *alors que* [-] *elle* [-] *partait* [$] *de* [-] *son domicile* [=] *à pied* [*] *pour se rendre* [$] *à* [-] *son bureau* [.]
(Sikh extremists assassinated Indira Gandhi Wednesday in New-Dehli. They shot the Indian Prime Minister as she was leaving her home on foot to go to her office.)

Version 3. *Mercredi à New-Dehli* [,] *des* [-] *extrémistes sikhs* [=] *ont assassiné* [@] *Indira Gandhi* [,] *en tirant* [@] *sur* [-] *le premier ministre indien* [*] *alors que* [-] *elle* [-] *partait* [$] *de* [-] *son domicile* [=] *à pied* [*] *pour se rendre* [$] *à* [-] *son bureau* [.]
(Wednesday in New-Dehli, Sikh extremists assassinated Indira Gandhi by shooting the Indian Prime Minister as she was leaving her home on foot to go to her office.)

Conclusion

The semantic-representation-to-speech system developed in COMMON-LISP produces a spoken message of about 35 words in less than 1 second.

In our system, only the strategic component is domain dependent. The lexicon and discourse structures used to build the text templates are domain dependent linguistic data. The rest of the system is domain independent. Let us recapitulate the data and rules integrated in it:

- a syntactic component which can apply the French grammar rules whatever the structure of the texts and the syntax of the sentences;

- a complete phonological dictionary of the 50,000 basic forms of French and a set of rules for obtaining a phonetic text from a phonological text;

- a complete syntactico-prosodic grammar of French and a set of rules that enable us to enter prosodic markers in a sentence whatever the syntax of the sentence;

- a speech synthesizer and a synthesis software.

Of course, these data and rules are only valid for French but it must be clear that the same kind of data is required for other languages and that the algorithm should be similar.

Bibliography

COURBON, J. L., & EMERARD, F., 1982, "SPARTE: A Text-to-Speech Machine Using Synthesis by Diphones", *IEEE Int. Conf. ASSP*, pp. 1597-1600, Paris.

COURTOIS, B., 1984, "DELAS : Dictionnaire Electronique du LADL, mots Simples", Rapport technique du LADL, nº 12.

DANLOS, L., 1984 a, "Conceptual and Linguistic Decisions in Generation", in *Proceedings of COLING 84*, Stanford University, California.

DANLOS, L., 1984 b, "An Algorithm for Automatic Generation", in *Proceedings of ECAI 84*, T. O'Shea éd., Elsevier Science Publishers BV, Amsterdam.

DANLOS L., 1986, *The Linguistic Bases of Text Generation*, Cambridge University Press, Cambridge.

EMERARD, F., 1977, *Synthèse par diphones et traitement de la prosodie*, Thèse de troisième cycle, Université de Grenoble III.

LAPORTE, E., 1984, "Transductions et phonologie", DEA, Université de Paris 7.

LAPORTE, E., 1986, "Application de la morpho-phonologie à la production de textes phonétiques", *Actes du séminaire "Lexiques et traitement automatique des langages"*, Toulouse.

MARTIN, Ph., 1979, "Un analyseur syntaxique pour la synthèse du texte", *Actes des 10º Journées d'études sur la parole*, pp. 227-236, Grenoble.

SANTOS, J. M., & NOMBELA, J. R., 1982, "Text-to-Speech Conversion in Spanish: A Complete Rule-Based Synthesis System", *IEEE Int. Conf. ASSP*, pp. 1593-1596, Paris.