# SIMULATION OF A FRENCH READING MACHINE FOR THE BLIND

D. O'Shaughnessy, M. Lennig, and P. Mermelstein
Université du Québec

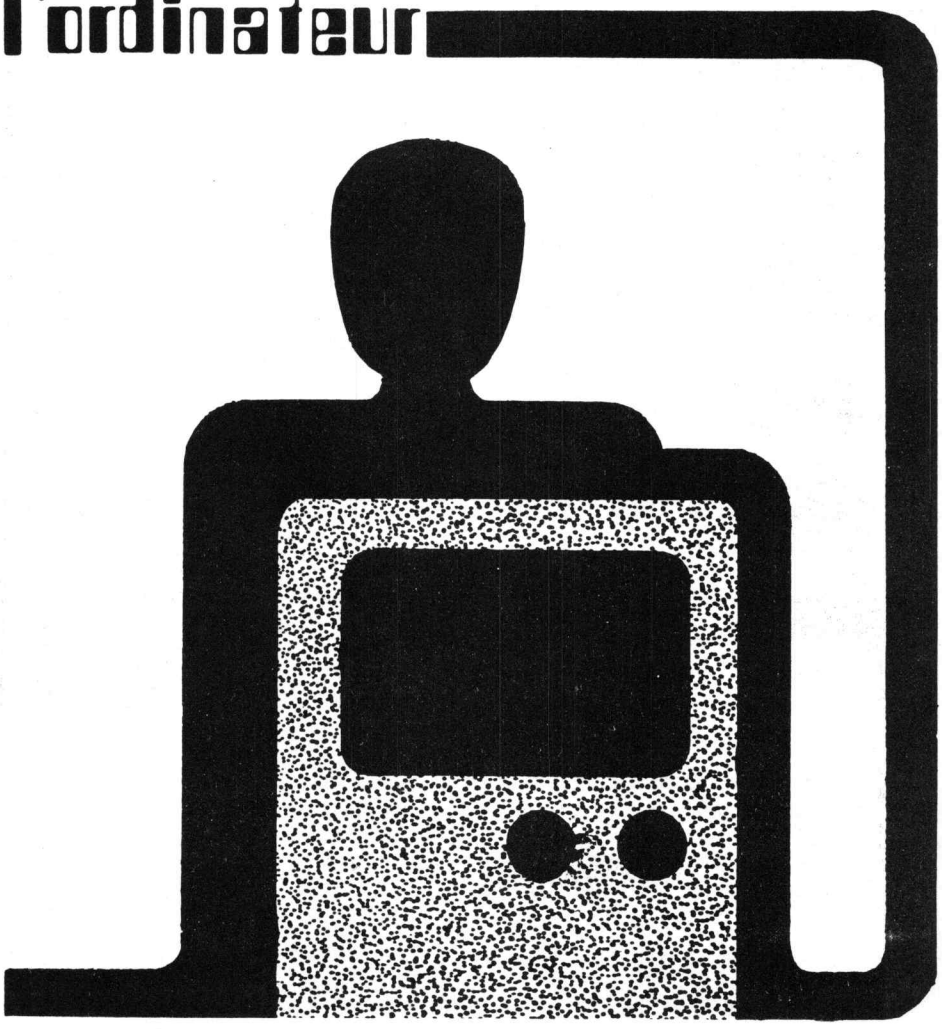
M. Divay
Institut Universitaire de Technologie de Lannion, France

Reprinted from the Proceedings of the

Tiré-à-part des comptes-rendus de

**Canadian Man-Computer Communications Society**
**L'Association canadienne des communications entre l'homme et l'ordinateur**

7th / 7ième
Conference
June 10–12 Juin, 1981
Waterloo, Ontario



Copies of this and previous CMCCS Conferences are available from:

Canadian Information Processing Society
5th Floor, 243 College Street
Toronto, Ontario, Canada
M5T 2Y1

Price:

| | | |
|---|---|---|
| 7th Conference | 1981 | $23.00 (non member) |
| | | $20.00 (member) |
| 6th Conference | 1979 | $ 7.00 |
| 5th Conference | 1977 | $ 7.00 |
| 4th Conference | 1975 | $ 5.00 |

Proceedings of the 1st, 2nd, and 3rd Conference are out of print.



On peut se procurer des copies de compte rendu de la présente conférence et des conférences antérieures de L'ACCHO à l'adresse suivante:

L'Association canadienne de l'informatique
243, rue College, 5e étage
Toronto (Ontario) Canada
M5T 2Y1

Prix:

| | |
|---|---|
| 7e conférence 1981 | $23 pour un non-membre |
| | $20 pour un membre |
| 6e conférence 1979 | $ 7 |
| 5e conférence 1977 | $ 7 |
| 4e conférence 1975 | $ 5 |

Les comptes rendus des première, deuxième et troisième conférences sont épuisés.

# SIMULATION OF A FRENCH READING MACHINE FOR THE BLIND

D. O'Shaughnessy*, M. Lennig*, P. Mermelstein* and M. Divay**

** INRS-Telecommunications, Université du Québec*
*** Institut universitaire de Technologie de Lannion, France*

## ABSTRACT

A reading machine for the blind generates speech output from typed text input. First, an optical character recognizer produces a binary image from a printed page, and this is interpreted as a string of symbols, using template matching techniques. Second, a linguistic processor translates letters into phonemes, and provides intonation appropriate to the syntax of the message. In French, this requires deciding when to pronounce usually- silent word-final letters, and using pronunciation rules different from English. Syllables have more regular durations in French, and pitch tends to rise more at the end of a French word than an English word. A final component interprets the phonemes as variations in vocal tract resonances and amplitudes. Sound output comes from a speech synthesizer modified for French by the addition of nasalized and front rounded vowels. Without making such acoustic rule modifications, the speech would have a heavy English accent.

## RÉSUMÉ

Un dispositif de lecture destiné aux aveugles produit une version parlée d'un texte dactylographié. D'abord, un dispositif reconnaissant les caractères optiques produit une image binaire d'une page imprimée, image qui est interprêtée comme une chaîne de symboles, grâce à des techniques d'assortiment par gabarit. Ensuite, un processeur linguistique traduit les lettres en phonèmes et donne l'intonation appropriée à la syntaxe. En français, ce processeur décide quand il faut prononcer les finales ordinairement muettes, basé sur des règles de prononciation différentes de celles de l'anglais. Les syllabes sont plus longues qu'en anglais et le ton a tendance à monter davantage à la fin d'un mot. Un dernier élément interprète les phonèmes comme des variations des résonances et amplitudes du conduit vocal. La production des sons provient d'un synthétiseur de la parole modifié pour le français par l'addition des voyelles antérieures arrondies et des nasalisées. Sans ces modifications aux règles de l'acoustique, la parole aurait un fort accent anglais.

## 1) INTRODUCTION

A machine which is capable of accepting printed text as input (in the form of loose pages, newspapers, books, or even images on television screens) and producing speech as output (the spoken version of the text) represents a socially-desirable and technically-challenging venture. Such a machine would be of great benefit to persons with vision disabilities, since so much communication these days involves the printed page. Few blind people can read braille, but virtually all understand speech and thus a reading machine would provide access to much information which must now be recorded on tape by sighted humans reading the text aloud.

One such machine recently was developed for the English language (KURZWEIL, 1976), and a second similar device is nearing production (CALDWELL, 1978). However, no French reading machines currently exist. Our effort has been to simulate such a machine, to demonstrate the quality of synthetic speech with French text as input.

## 2) COMPONENTS

A reading machine consists of 3 components: optical character recognition (OCR), linguistic processing, and speech synthesis.

### 2.1 Optical Character Recognition (OCR)

The OCR component does image-to-grapheme translation, i.e., it converts an image of printed text into the sequence of characters (letters, numbers, and symbols) which the image contains. In our system, the first state of the OCR device employs an automatic roll-feed page scanner (manufactured by Stewart-Warner Corp.), which converts a page image to a digital representation, by interpreting the image as thousands of black or white points. On a white page, of course, contiguous sets of black points would appear to the human eye as letters.

Our scanner resolves the image to an accuracy of 208 points per inch; thus using a standard type spacing of 10 or 12 characters per inch, each symbol would be sampled 21 or 17 times across its width. Vertical sampling ranges from 16 points per symbol for short letters such as 'a' to 28 points for tall ones like 'j'. Each sample is quantized to 8 bits on a black-to-white scale. A threshold decides which of the gray points should be considered black (parts of a symbol) or white (part of the background). The 208 points per inch is adequate to resolve fine variations in most characters, although small text such as in telephone books and classified newpaper advertisements would require better resolution.

To interface to this OCR hardware device, software has been developed to translate the digital image representation into a string of graphemes. The programs accept a standard IBM type font, and have an error rate of about 0.2%, using template matching as the pattern recognition technique.

The programs segment the image into regions corresponding roughly to letters, using constraints on the size of the letters and heuristics such as the fact that typed pages usually consist of horizontal lines uniformly spaced in the vertical and horizontal dimensions. In French text, an accent mark may be classified into a region separate from the rest of the letter, but these regions are jointed back together using contextual information. Each input region is compared to a series of stored region templates, to find which letter the input symbol corresponds to. Most of the letters have one stored template, but certain symbols, whose bit pattern representations can differ significantly due to the resolution of the scanning device, have two stored templates. Such a simple template set is possible only because the input type is restricted to the one font used to train the system.

The templates are ordered in sequence from most likely to least likely, according to rough statistics of the French language. Thus more frequently occuring vowels such as 'e' and 'a' are tested early, and rare consonants such as 'w' and 'k' come later.

Since the comparison search terminates when a suitable match is found, the testing of more likely candidates first reduces the average computation time. In the scanning process, some versions of the same symbol may appear bigger than others. Thus the comparison routine centers all regions according to their maximum limits, and then shifts them horizontally and vertically for better matches if the input and stored templates are sufficiently similar.

## 2.2 Linguistic Processing

The linguistic processor component does grapheme-to-phoneme ("letter-to-sound") translation, i.e., it converts the sequence of letters and punctuation into a string of French phonemes. It also produces a set of intonation markers suitably detailed to allow specification of the pitch and duration for each phoneme. The first task is performed by rules which convert graphemes into phonemes according to the inherent properties of the French language. Each language can be described by its own set of rules, and some languages can be more compactly described than others. For instance, Spanish has a relatively small, comprehensive set of letter-to-sound rules, while English is considerably more complex (HUNNICUT, 1976; ELOWITZ et al, 1976). French lies somewhere between the two.

### 2.2.1 Letter-to-phoneme translation:

A combination lexicon and letter-to-phoneme approach is likely optimal. A compact set of letter-to-sound rules can be used to cover the majority of words in a language, while the lexicon would consist primarily of words not easily described by letter-to-sound rules (which usually would include very common words whose pronunciation deviates from the normal rules, and some foreign words). To increase the power of the lexicon, common prefixes and suffixes could be stored there, and input words would first be decomposed into affixes and roots before look-up in the dictionary. Thus it is not necessary to store all derivations and conjugations of a word in the lexicon, but only its basic components. This is especially helpful in French, with its many verb terminations.

The letter-to-phoneme component developed at Université de Rennes and CNET-Lannion (DIVAY ET GUYOMARD, 1977) consists of a set of linguistic rules and an interpreter to decide which rules apply in a given phonetic situation. The rules are ordered so that liaison (e.g., "nous avons") and enchainement (e.g., "il écoute") are examined first, then word final /s/ (which could be silent (e.g., "chats", "chiens") or not ("hélas")), then common words, regular words, and finally elision of certain word final letters. The special handling of common words enables the program to function more quickly, since a large percentage of words will not have to go through the large set of regular letter-to-phoneme rules. The rules are context-sensitive in the form:

$$a \longrightarrow b \;/\; c + d$$

(the symbol sequence 'a' is transformed into the sequence 'b' if 'a' is found in the context 'c - d', that is, 'c' on the left and 'd' on the right). For example, a vowel followed by 'm' or 'n' would be considered a nasalized vowel if the ensuing letter were a non-nasal consonant, but not if a vowel ensued (e.g., "lent" versus "ennemi").

### 2.2.2) Intonation:

The main parameters of intonation, pitch and duration, vary with several parameters of the input text. Phonemic variation, such as low vowels having lower pitch and longer durations, can be treated directly from the list of input phonemes. But since the pitch and durations of the components of a word also depend upon syntactic structure and semantics, it is necessary to do further linguistic text analysis. The lack of an adequate intonation model in many speech synthesizers is a major cause of the "foreign" accent of speech synthesized by rule.

An algorithm has been developed to accept as input the punctuation and a simple parts-of-speech analysis as well as the phonemes of a sentence, and yield an intonation contour as output. A list of function words (articles, pronouns, prepositions) in the lexicon helps distinguish the syntactically less important words from the content words (nouns, verbs, adjectives), which have longer durations and greater pitch changes.

We modelled the fundamental frequency and durations of phonemes after the findings of an analysis of natural French speech (O'SHAUGHNESSY, 1980). It was found that vowel duration varied widely as a function of nasality and the features of ensuing consonants (nasalized vowels and those before voiced consonants being long vowels), and

that consonants could either shorten or lengthen in consonant clusters, depending on the difficulty of articulation and the proximity of points of articulation (e.g., the durations of /s/ and /t/ in "reste" would be shorter than those in "messe" or "mette", due to the similarity of articulation). Fundamental frequency tends to rise on important words and remain steady on less important words, but the rapidity of variation seems to be less than in English.

The durational algorithm assigns a duration to each phoneme in a given sentence, and then modifies it according to its phonetic context. Thus, for example, vowels in general and nasalized vowels in particular are assigned long durations, while unvoiced consonants are assigned short durations. Then these base durations are adjusted by the immediate context: (a) if a consonant occurs in a cluster, its duration is shortened if the consonants have similar phonetic features or lengthened if the consonant cluster is difficult to pronounce, (b) vowels are shortened if followed by unvoiced consonants and lengthened if followed by voiced consonants (especially voiced fricatives). At the next higher-level context, that of the word, durations are shortened for words of several syllables and for words presumed to be less important and therefore less stressed.

Punctuation affects duration in that durations are lenghened in the word just prior to a punctuation mark (this mark is assumed to accompany a short pause in the speech). Since some sentences can be fairly long without internal punctuation marks, a simple syntactic analysis chooses likely candidates for phonological boundaries, and lengthens the words just prior to them.

The fundamental frequency (Fo) algorithm assigns a general trend, known as the declination line, to major clauses in the sentence, in which the Fo starts with a positive offset and gradually declines with time. Major modifications to this general pattern are due to punctuation: a period marking the end of the sentence causes a major fall in Fo, while a question mark causes the Fo to remain high throughout with a major rise at the end of the sentence. (If the question sentence contains a word such as "qui", "comment", "pourquoi", etc., the intonation resembles that of a statement instead). Other punctuation marks, such as commas, result in Fo rises of lesser magnitude just before the pause which marks the punctuation location. Within phrases (phonological units between presumed pauses), Fo varies according to the presumed importance of the word and the number of syllables in the word. The importance of a word relates to its sentence stress level: either stressed or unstressed. Unstressed words have a falling Fo (but not to so low a value as to go below the declination baseline, as that would indicate the end of the sentence). Stressed words usually have a Fo rise on their final syllable, and occasionally on the first syllable as well (in these polysyllabic cases the middle syllables would have slight Fo falls).

2.3) Speech Synthesis

The speech synthesis component (FLANAGAN and RABINER, 1973; FLANAGAN ET AL, 1970) translates a string of phonemes and intonation markers into the actual speech output, and consists of two parts: a phoneme-to-parameter converter, and a vocal tract model. The vocal tract model is currently simulated in Fortran software, but will soon be replaced by a hardware LSI chip capable of real-time synthesis.

The phoneme-to-parameter converter involves a fundamental choice of approach. One could record components of natural speech, and process them by linear predictive (LPC) analysis for reduced storage space. Then as various sounds were needed in sequence for a given sentence, the stored components could be transformed and concatenated to yield the output speech. Alternatively, synthesis can be done phonemically by direct simulation of the established parameters of speech (i.e., manipulation of the formant resonances, bandwidths, and amplitudes).

The LPC concatenation approach has less of a machine accent, but is limited to simulating the voice of the original speaker, and is subject to the degradation of LPC speech, as well as requiring more storage space. The phonemic approach potentially can produce perfect speech, but in practice usually gives a machine accent.

We have chosen the second approach and have modified an existing software synthesizer developed by Klatt at M.I.T. (and now in use the Telesensory Systems reading machine) to handle French phonemes (Klatt, 1980). For example, we added the nasalized and front rounded vowels which exist in French but not in English, and deleted the English phonomenon of diphthongization. Another major

change occurred in the so-called liquids /l/ and /r/. The French /l/ is articulated more forward in the oral cavity, with less tongue tip contact with the hard palate. This tends to raise the second formant resonance compared with the English /l/. The English /r/ is a retroflex sonorant with a markedly low third formant; the French /r/, on the other hand, is pronounced with-out retroflex of the tongue tip, and usually with a velar point of constriction. It is often accompanied by frication noise, and indeed has no voicing (vibration of the vocal cords) at all in contexts where the /r/ is adjacent to an unvoiced consonant in a syllable (e.g., "parc", "trac" - unvoiced; "large", "vrai" - voiced).

3) SUMMARY:

Our first year of development of the French reading machine has involved research and development oriented toward a non-real-time demonstration. We are now able to exhibit a reading machine simulation in which:

1) the OCR device produces a symbol string from a typed page,

2) this string is then processed by successive software programs to do pattern recognition, linguistic processing, and speech parameter production.

3) the parameter string then drives a software speech synthesizer, which puts the speech into a digital waveform.

The synthetic speech, produced in non-real-time, has been judged by informal listening tests to be intelligible and to have an acceptable quality. More objective perceptual tests will be performed in the near future.

In the upcoming year, hardware to implement the reading machine in real-time will be designed to replace some of the functions of the software programs. Further computation time savings will result from an implementation of the system using a microprocessor. Upon interfacing of all components, a real-time demonstration, producing speech output as the text is being scanned, should be feasible. This second demonstration would show how well the system functions under expected operating conditions, i.e., with a human operating the OCR device and listening to the output speech simultaneously.

REFERENCES

Beauchemin, N. (1972) "Corrélation des durées sous l'accent en français", 7th Intern. Cong. on Phonetic Sciences (1971), A. Rigault & R. Charbonneau (Mouton: the Hague), 860-865.

Benguerel, A. (1971) "Duration of French vowels in unemphatic stress", Language & Speech 14, #4, 383-391.

Caldwell, J. (1978) "Flexible, high performance speech synthesizer", J. Acoustical Society of America 64, S1, S72.

Choppy, C. et Lienard, J. S. (1975) "Un algorithme de prosodie automatique sans analyse syntaxique", Journées d'Etudes, Toulouse.

Divay, M. et Guyomard, R. (1977) "Conception et réalisation sur ordinateur d'un programme de transcription grapémo-phonétique", thèse de 3ème cycle, Université de Rennes.

Elowitz, H., Johnston, R., McHugh, A., and Shore, J. (1976) "Letter-to-Sound Rules for Automatic Translation of English Text to Phonetics", IEEE Trans. on ASSP, Vol. ASSP-24, no. 6, 446-458.

Flanagan, J., Coker, C., Rabiner, L., Schafer, R., and Umeda, N. (1970) "Synthetic voices for computers", IEEE Spectrum 7, 22-45.

Flanagan, J. and Rabiner, L. (1973) SPEECH SYNTHESIS (Dowden, Hutchinson, and Ross, Stroudsburg, Pennsylvania).

Hunnicutt, S. (1976) "Phonological Rules for a Text-to-Speech System", Journal Assoc. Computational Linguistics, microfiche 57.

Klatt, D. (1980) "Software for a cascade/parallel formant synthesizer", J. Acoustical Society of America 67, 3, 971-995.

Kurzweil, R. (1976) "The Kurzweil reading machine: a technical overview", Science, Technology, and the Handicapped, AAAS Meeting, 3-7.

Mettas, O. (1969) "Etude sur la durée des consonnes dans l'un des parlers parisiens", Studia Linguistica 22, #2, 91-103.

O'Shaughnessy, D. (1981) "A study of French vowel and consonant durations", J. of Phonetics, to appear.

Quino, J. et Teil, D. (1970) "La synthèse de la parole à partir des digrammes phonétiques", Revue Acoustique 3, numero 9.

Smith, A. (1977) "The timing of French, with reflections on Syllable timing", Works in Progress, U. of Edinburgh 10, 97-108.

Vassiere, J. (1971) "Contribution à la synthèse par règles du français", thèse de 3ème cycle, Université de Grenoble.

Wajskop, M. (1979) "Segmental durations of French intervocalic plosives", Frontiers of Speech Communication Research, B. Linblom & S. Ohman (Academic Press: N. Y.), 109-123.