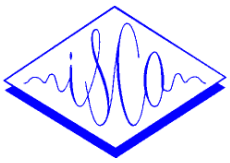

Experiments in Prosody for the Generation of Oral French

Craig Thomas, Michael Levison*, Greg Lessard***

*School of Computing / **French Studies, Queen's University, Kingston, Ontario, Canada
craigt@webhandcentral.com, levison@cs.queensu.ca, lessardg@post.queensu.ca

Abstract

This paper describes a series of experiments designed to allow VINCI, a natural language generation environment, to produce not just orthographic, but also phonetic output in French with the help of the MBROLA speech generator. In order for this to be possible, it is necessary for the syntactic representation of utterances generated by VINCI to contain phonetic symbols and prosodic information, including pitch and length. We show that the resulting system is capable of generating oral output found to be largely acceptable according to a set of human evaluators and describe a number of pratical applications of the system.

1. Introduction

VINCI is a natural language generation system which accepts files specifying the lexicon, syntax and morphology of a language, and creates utterances in a form specified by the user. Utterances may be generated randomly or with a measure of semantic control. The original purpose of the system was to produce test exercises for language learning. It has, however, become a general-purpose tool used to study a variety of linguistic phenomena, including the generation of short texts (fairy tales), linguistic humour (puns and riddles) and limericks. It has facilities for adaptive control and for detailed comparison of a learner answer with an expected answer, which can be applied to the investigation of learner errors. In principle, the system can be used for any natural language, though most of our (non-humour) research relates to French.

Until recently, VINCI's output has taken a written form displayed on the computer screen. We have now experimented with the production of oral output.

At the simplest level, this is achieved by including phonetic strings in the lexicon and providing phonetic morphological rules analogous to orthographic ones. VINCI can then produce output as a sequence of SAMPA-like symbols instead of, or as well as, its normal orthographic form. With some intermediate adjustments to cater for differences in representation of these symbols, this phonetic output can then be passed to the MBROLA program ([1]), which in turn produces the requested speech. This Concept-to-Speech approach may be distinguished from the Text-to-Speech model in which oral output is obtained from a previously generated orthographic representation (see [2] for discussion).

If this was all we did, the result would be the well-known monotone typical of early speech synthesizers. For the result to be useful to a language learner, we must incorporate prosody to make the result as close as possible to the utterances produced by human speakers. This is the topic of the current paper.

MBROLA is a generalized speech synthesizer based on diphones, which accepts a database corresponding to the language to be spoken and a phonological representation of an utterance, and produces the requested speech. In this research, we have used a database for a female voice speaking Parisian French, though other French voices might easily be substituted.

Input to MBROLA consists of a sequence of phonemes and their durations in milliseconds. Each phoneme may also be accompanied by a "pitch-pair" which alters the phoneme's fundamental pitch. Figure 1, for example, displays such a sequence for the words "l'adulte ne ..." in which the pitch of the phoneme /a/ changes 50% of the way through from its base level of 80 Hz to 103 Hz. (The synthesizer interpolates the change linearly.)

```
l 80 (50, 80)
a 80 (50, 103)
d 60 (50, 80)
y 80
l 80
t 80
n 80
@ 80 (50, 126)
```

Figure 1. An example of MBROLA input for the fragment "l'adulte ne ..."

Such pitch-pairs, along with changes to the phoneme length, are used to implement the components of the prosody. In order to produce them, VINCI must be made to generate prosodic information as well as SAMPA output. In addition, since prosody and syntax are tightly bound together in French, prosodic descriptions must be built into the syntactic representation of each utterance.

2. French Prosody

We are concerned here with three aspects of prosody: phoneme length, pitch and stress.

The lengths of French phonemes, some of which are context-dependent, have been studied extensively by Simon [3] and Brichler-Labaeye [4]. We have already mentioned an intermediate program which converts the IPA output produced by VINCI to the form required by MBROLA. Information about lengths is incorporated into this program, which creates the second component for each MBROLA phoneme. (See Figure 1.)

Information about pitch changes was not available in the form we needed. It would have been possible to analyze a corpus of oral French. This, however, would have required us to transcribe the individual sentences into standard orthography, and to parse each into its constituent clauses and phrases in order to determine the relation between phrases and changes of pitch. Accordingly we have conducted a limited series of experiments to obtain this data. (See section 3 below.)

Stress in French is a complex phenomenon whose physical instantiations may include syllable length, intensity, and pitch. French is a fixed stress language in that the last syllable of the syntactic unit (typically the clause) receives stress, while all preceding syllables are unstressed. For example, "la FILLE", "la petite FILLE", "la petite fille maLADE", where the stressed syllable is shown in upper case. MBROLA represents stress as length.

3. First Experiment

VINCI was used with a simple syntax to generate a series of sentences, both the orthographic and phonetic forms being produced. The sentences included a subject, a verb and optionally a direct object and an adverb. We allowed sentences which were interrogative, declarative or imperative, positive or negative, with subjects being nouns, pronouns or proper names, using transitive or intransitive verbs.

For each of the 80 reasonable combinations, two instances were created, the more meaningful being selected so as to reduce the effects of semantic anomaly. These were spoken by native French speakers, digitally recorded as WAV files, and analyzed by software called Praat [5]. Praat produces a graph of fundamental frequency (pitch) vs time, on which the words of the sentence can be overlaid, as shown in Figure 2 below.

Qualitative analysis of these "Praat pictures" indicates that the curves can be adequately approximated by about eight pitch levels, and strongly suggests that pitch is a cumulative combination of three factors: word effects, phrase effects and sentence effects. Each (long) word appears to have its own rise-fall curve, which is superimposed upon the phrase curve. Different types of phrases each have their own curve. For example, a noun phrase, which often consists of a determiner and a noun, has a rising pitch on the determiner and a falling pitch on the noun. Sentence effects are similar. Declarative sentences tend to start at the speaker's resting frequency, show a rising pitch and then a fall to a frequency below the resting frequency. Interrogative sentences tend to show an overall rising pitch. These results are generally in line with previously observed tendencies (see [6] for examples).

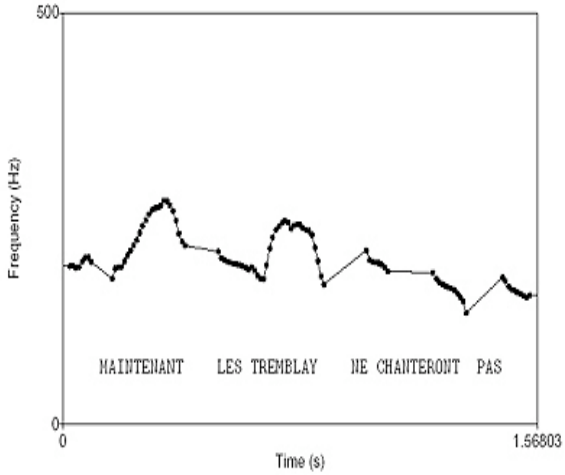

Figure 2. A Praat picture.

Some evidence for paragraph-level intonation effects was obtained by asking two of the francophones to read generated fairy tales. The sample, however, was too small to draw conclusions with confidence.

To confirm these initial conclusions, symbols indicating pitch changes were hand-encoded into the phonetic strings which VINCI created for the test sentences, and these were submitted to MBROLA, which "spoke" the results. The linguist among us (Lessard) listened to these and suggested changes, mostly to the phoneme lengths, which exhibited more contextual variations than the earlier works had suggested. In the course of this analysis, some minor deficiencies in the MBROLA database became apparent. For example, occlusive consonants which form the final phoneme of an utterance normally remain unreleased; but the database provides only explosive versions for these consonants, causing resulting sentences to sound slightly odd to a native speaker.

We also wished to determine whether all three of the observed factors: word, phrase and sentence effects, were essential. Different combinations were therefore encoded into the same sentences and spoken by MBROLA. Native speakers were then asked to judge which verison of the sentence sounded most natural. In every case, the version containing all three effects was judged best.

The next step was to have the prosodic markings inserted into the phonetic strings by computer rather

than hand-encoding. For this purpose, 28 symbols were added to the phonetic alphabet, representing absolute or relative changes of pitch, syllable breaks and stress markers. Intra-word pitch markers were simply included within the phonological spelling in the lexicon. Pitch variations related to phrases or sentences were inserted as new lexical entries, and these new "words" were introduced into the syntax rules. Stress marks were added, in accordance with the rule mentioned above, by the intermediate program between VINCI output and MBROLA input. With these additions, prosodically marked phonetic output is generated automatically.

4. Second Experiment

Using the facilities described above, a new set of sentences was created, similar in form and variety to the ones produced originally. These were generated by VINCI, synthesized by MBROLA, and recorded. Five judges, all native French speakers, were asked to listen and evaluate the quality of the results. Each judge was given a random selection of twelve sentences and asked to rank their quality on a five-point scale, as shown in Figure 3 below.

For each sentence you hear, please indicate one of the five replies:

strongly disagree (SD)
disagree (D)
neutral to the statement (N)
agree (A)
strongly agree (SA).

Q1 The voice clearly pronounced words and was understandable.
Q2 The rise and fall of the voice sounded natural.
Q3 The speed at which the sentence was read was reasonable.
Q4 Overall, this sentence sounded very similar to the way a human might have spoken it.

Please add any comments you may wish to make on the sentence.

Figure 3. The questions put to the judges

The aggregated results are shown in Table 1 below. In one or two cases questions were missed, leading to totals below 60. Since the judges were all professors or graduate students in French Studies with language teaching experience, they usually pinpointed the deficiency where a rank was N or lower. Most were minor problems, easily fixed: a phoneme too long, missing phonetic variants in the MBROLA database, and so on. Nonetheless, 68.7% of evaluations were in the A or SA categories.

	SD	***D***	***N***	***A***	***SA***
Q1	1	9	8	21	20
Q2	1	9	12	18	18
Q3	0	5	3	25	25
Q4	3	9	13	25	8
Total	5	32	36	89	71

Table 1. The results aggregated for all judges

5. Applications

The ability to map between oral and written stimuli allows for a number of applications which we are currently exploring. Some of these hinge on other features of the system, including a complex set of mechanisms for evaluating user input and the ability to adaptively drive generation based on pedagogical requirements and the analysis of previous responses. (Further details may be found in [7] and [8].) In what follows, we describe three examples.

The first is the traditional *dictée*, in which learners are presented with utterances spoken by the system and requested to provide written transcriptions. This allows for the testing of a range of agreement phenomena which have written but not oral representations. Compare, for example, the written and oral versions of the sentence:

Les marchandises sont arrivées vers deux heures.
'The merchandise arrived around two o'clock'
[lemarSâdizsôtaRivevERd0z%R]

where the third line contains the oral output in VINCI's SAMPA-like notation. In written French, the past participle must agree in gender with *marchandises* and in number with the subject noun phrase. The final *-es* on *arrivées* reflects this. However, neither of these phenomena appear in the oral equivalent. Since the generation environment produces all utterances from its grammatical representation, it 'knows' the correct gender and number as well as the structure of the sentence. A student typing:

Les marchandises sont arrivés vers deux heures.

will see the diagnostic:

Vous avez mis la forme masculine: arrivés
mais “Les marchandises” est féminin.
Il faut mettre “arrivées”.

A second application of the program is in the area of the recognition and interpretation of intonation. Learners

often have difficulty perceiving intonation in their second language; perhaps as a result of this, many tend to impose their first language intonation rules on their second language utterances. Although there is evidence for the value of a teaching context in which learner intonation is identified and compared with a model used as a source of feedback, (see [9] for discussion), it remains that learning contexts of the sort remain relatively underused. Several variants exist. In the simplest, exercises are based on the use of a small set of utterances either pre-recorded or spoken by an instructor, and learners are asked to repeat them, with qualitative feedback being provided by an instructor. In a more technologically advanced model, intonation is represented by means of an intonation visualizer, but the set of target utterances remains closed and typically small. The combination of VINCI, MBROLA and Praat provides a extension to this approach, both in terms of the range of utterances which learners may experience (since utterances are generated on the fly and may be in principle of unlimited number), and in the richness of the feedback available.

In one framework, VINCI (by means of MBROLA) generates a set of output utterances, some declarative, some interrogative, etc. using the grammar described in Section 2 above, and learners are asked to identify which intonation is being presented. At a more advanced level, each of these utterances forms a 'target' which learners are asked to imitate. Praat is applied both to the target utterance generated by the system and to learner attempts to mimic these. Multiple attempts are possible with, it is hoped, convergence on the model. Preliminary experiments with this model suggest that learners find it interesting. It remains to be demonstrated whether it leads to measurable improvements in subsequent performance.

A third application takes the form of a tool for the teaching of phonetic transcription to linguists. This may function at one of two levels. In one, the lexicon is composed of the phonemes of a given language (French, for example) together with their phonetic characteristics (thus `[i]` carries the traits 'front, closed, oral, unrounded') and VINCI syntax rules generate phonotactically possible sequences of syllables (for example: `[le-ta-mir-stro-me-li]`. Learners transcribe these into a modified IPA representation. Errors of transcription can be diagnosed with some precision since the system already 'knows' the appropriate traits associated with each symbol. At a more advanced level, oral utterances to be transcribed will be similar to those shown for the *dictée* above, but VINCI generates not an orthographic transcription, but rather a phonetic transcription, including representation of intonation.

6. Conclusions

Our research to date has demonstrated the ability to marry a language generator and a speech synthesizer to produce oral output close to human speech. In subsequent work we will examine longer and more complex sentences and paragraphs, taking into account factors which add variation to generated utterances, including different voices, phonostylistic factors, and random variations from one utterance to the next.

Acknowledgements

The research described here was made possible by a Standard Research Grant from the Social Sciences and Humanities Research Council of Canada.

References

[1] Dutoit, Thierry (2002). The MBROLA Project. http://tcts.fpms.ac.be/synthesis/mbrola.html

[2] Dutoit, Thierry (1996). *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers. Dordrecht.

[3] Simon, Pela (1967) *Les consonnes françaises*. Paris: Editions Klincksieck.

[4] Brichler-Labeaye, Catherine (1970). *Les voyelles françaises*. Paris: Editions Klincksieck.

[5] Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

[6] Léon, Pierre and Martin, Philippe (1969). *Prolégomènes à l'étude des structures intonatives*. Paris: Didierions Klincksieck.

[7] Levison, Michael and Lessard, Greg. (2000). A Multi-Level Approach to the Detection of Second Language Learner Errors. *Literary and Linguistic Computing*. *15/3:313-322*.

[8] Levison, Michael; Lessard, Greg; Danielson, Anna Marie and Merven, Delphine. (2001). From Symptoms to Diagnosis. In *CALL – The Challenge of Change*. (K. Cameron, ed.) Exeter: Elm Bank, pp. 53-59.

[9] Lepetit, D. (1992). *Intonation française: enseignement et apprentissage*. Toronto: Canadian Scholars Press.