Crossroads
The ACM Student Magazine

Automatic Generation of French Speech

by ***Craig Thomas***

Introduction

In the past 60 years, scientists have created numerous exotic speech synthesis devices that attempt to generate sounds indistinguishable from those made by a human being. Early approaches utilized custom-built synthesizers that attempted to reproduce the same type of sounds as our vocal tracts, with varying degrees of success [5]. Over time, however, the sophistication of computer technology has advanced to the point where computer sound cards are able to produce digitized recordings. This means that special hardware is no longer required. In fact, the digital signal processing chip found on all modern sound cards produces sounds of higher quality than that of earlier hardware. Building on the speech synthesis theories developed concurrently with the technology, high-quality, low-cost concatenative speech synthesizers have been created that are capable of reproducing high quality speech [3]. Although synthesized speech sounds nearly human, there is an element of everyday speech that automated methods lack. The missing link is the **prosody**, the patterns of stress, inflection, and intonation in a language.

Consider the sentence "You are John Smith." Depending on how the sentence is spoken, there may be many different ways to interpret the actual meaning. With falling pitch at the end, this is a declaration. With rising pitch, it becomes a question. Indeed, as **Figure 1** shows, the sentence changes meaning depending on where we place stressed words.

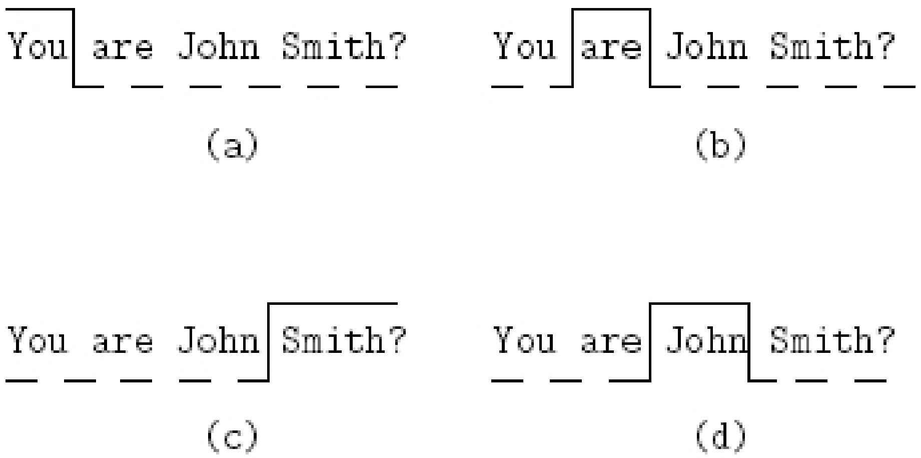


Figure 1: Different kinds of questions depending on tone. Solid lines indicate stressed words. (a) Question to clarify identity, stress is placed on "You". (b) Question expressing disbelief, stress is placed on "are". (c) Question to clarify last name, stress is placed on "Smith". (d) Question to clarify first name, stress is placed on "John".

In the first case, the sentence can be one to clarify the identity of the other person. For example, the speaker could go on to say "*You* are John Smith? I thought *he* was John Smith!" In the second instance, the speaker is expressing disbelief at the situation. For example, the speaker could go on to say "You *are* John Smith? I thought you were someone else." In the third instance, the speaker is clarifying whether or not Smith is indeed the proper last name. For example, the speaker could go on to say "You are John *Smith*? I thought you were John *Stewart*." In the final sentence, the speaker is clarifying whether or not John is indeed the proper first name. For example, the speaker could go on to say "You are *John* Smith? I thought you were *Mike* Smith." It is such differences in pitch and stress that automated speech synthesis methods have difficulty capturing and reproducing.

What is Prosody?

Prosody refers to changes in pitch, stress (energy), and syllable length that are audible in a speech signal [2]. *Pitch* refers to the fundamental frequency of the sound wave produced. Each sentence we speak has a pitch contour associated with it that can be broken down into smaller sequences of elementary contours associated with linguistic phenomena. *Stress* in some models is thought to be related to loudness or phonetic force, but in others is believed to be a function of syllable length and timing [7]. Some models disregard stress and claim that prosody is simply comprised of pitch changes. In our research, stress is treated as a function of syllable length. The final component, *syllable length*, is used to indicate boundaries and stress, and also contributes to the overall rhythm of an utterance. As mentioned before, prosody is the missing link that many current speech synthesizers lack. The lack of prosody results in sentences that sound mechanical and monotonic.

How does computer science factor into the generation of prosody and this study in particular? The computer's role is, of course, everywhere: in the pitch analysis program (Praat), in generating prosodically marked phonological sentences from a grammar, in producing the sound output (speech) from the generated output, and in the programming and modifications involved with the natural language generation system. Many different natural language generation systems exist that are capable of producing utterances in a given language. Very few, however, have attempted to incorporate a speech synthesis device. Natural language generation systems have many advantages and can be used in a variety of different ways.

So, You Want to Learn French…

VINCI is a natural language generation system that allows users to create a grammar that models a natural language by specifying the syntax, morphology, semantics and lexical relations. The original purpose of VINCI was to create test exercises for language learning, but it has grown to become a general-purpose tool to examine a variety of linguistic phenomena. Simply put, users of the system can write grammars describing the features of a language. The system can then generate utterances in the language that was specified. In its application to language learning, VINCI typically generates random utterances of some particular type, and invites the learner to carry out a transformation on them. **Figure 2** shows a simple example in French. Other examples are discussed in [6].

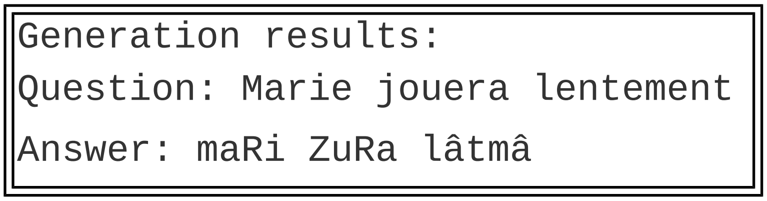
```
Question: L'homme a mangé une pomme.
Answer: L'homme l'a-t-il mangée?
```

Figure 2: An example of a generation from VINCI. In this particular example, learners are given a declarative sentence, and asked transform it to a question, making the object of the verb a pronoun.

The system also silently generates the expected results for comparison with the learner's reply. In fact, VINCI can perform a very extensive comparison, indicating the type of error the learner has made. Such exercises have also been incorporated into an adaptive system, which varies the precise nature of the next "question" based on the results of the comparison. Systems like this can be used to investigate the learner's misunderstandings which lead to the errors.

The system can go one step further and help to test a learner's understanding as well as their ability to conjugate or translate words. For example, a grammar has been created that automatically generates a complete fairy tale. The fairy tale tells a story of heroes, kings, and princesses, with the story and characters being generated randomly each time. A series of questions is generated along with the fairy tale. These test a learner's comprehension of whole segments of text rather than isolated sentences.

The next stage is to have the generation system create oral sentences and questions rather than words on a screen. The basis of such a system is simple. The current VINCI French lexicon contains over 3,000 entries that incorporate both phonological and orthographic representations of dictionary words. The phonological representations of the dictionary words consist of symbols from a variant of the International Phonetic Alphabet (IPA) where a single ASCII character corresponds to a single phoneme (a phoneme is simply a basic unit of sound). Since an IPA representation of a word is conceptually no different from an orthographic spelling, VINCI can already generate the phonological output relating to an utterance (**Figure 3**).



```
Generation results:
Question: Marie jouera lentement
Answer: maRi ZuRa lâtmâ
```

Figure 3: An example of a generation from VINCI showing both orthographic and phonological variations on a sentence. The question contains the orthographic output and the answer contains the phonological representation.

The phonological output of the system is directly available as input to a speech synthesizer. However, if a synthesizer were used to produce the spoken equivalent of this phonological output, the voice would seem flat and mechanical. In particular, a method to produce prosody in VINCI and a speech synthesizer capable of interpreting prosodic events are needed.

MBROLA, A Speech Synthesizer

MBROLA is a well-known speech synthesizer that takes as input a series of IPA characters and a prosodic description of the sentence and outputs speech [4]. The simplest input to MBROLA is in the form of an IPA character followed by a duration expressed in milliseconds (**Figure 4**).

```
l 80
a 80
d 60
y 80
l 80
t 80
n 80
@ 80
```

Figure 4: An example of MBROLA IPA input. The first column is the MBROLA IPA character to synthesize, and the second column is a duration in milliseconds to apply to the phoneme. This example is "l'adulte ne" from the full French sentence "l'adulte ne pleurerait jamais davantage" (*The adult would never cry any more*).

The above sequence, if synthesized, would sound monotone and non-human. So, in addition to just a phoneme and duration, MBROLA also accepts a construct known as a pitch pair. Pitch pairs have two numbers, the first being a percentage through the current phoneme to apply a new pitch and the second being a new pitch to apply at that moment in time (**Figure 5**).

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

Figure 5: An example of MBROLA IPA input along with corresponding pitch events.

The third column contains pitch pairs in brackets.

For example, in **Figure 5**, 50% of the way through the phoneme 'a', a new pitch of 103 Hz is applied. The synthesizer linearly calculates which pitch to apply to phonemes starting from a pitch of 80 Hz 50% of the way through phoneme 'l' to a target pitch of 103 Hz 50% of the way through phoneme 'a'. This allows for a great degree of control when implementing prosodic features in synthesized speech.

As we can see, the phonological output of VINCI is nearly in the form needed for MBROLA. In fact, we can trivially convert the output from VINCI into the desired input for MBROLA. The VINCI output, however, contains no information to describe prosodic events. This is the problem we need to address.

A Language Definition with Prosody

Our first task is to discover the kinds of prosodic changes introduced by human speakers. To perform this task, a number of native French speakers were asked to lend their voices for recording. Using VINCI, a series of sentences were generated that captured a wide variety of sentence types (interrogative, declarative and imperative, positive and negative, with subjects being nouns, pronouns, or proper names, using transitive and intransitive verbs). Each speaker was asked to read 10 of the generated sentences, and the results were recorded and stored as PCM WAVE files sampled at 22,050 Hz. A linguistic analysis tool called "Praat" was used to plot the voices in frequency versus time plots so that the relative locations of the pitch levels could be determined. In order to gain this information, the words of each sentence were overlaid with their associated Praat picture and plotted so that the phenomena could be easily observed. An example is shown in **Figure 6**.

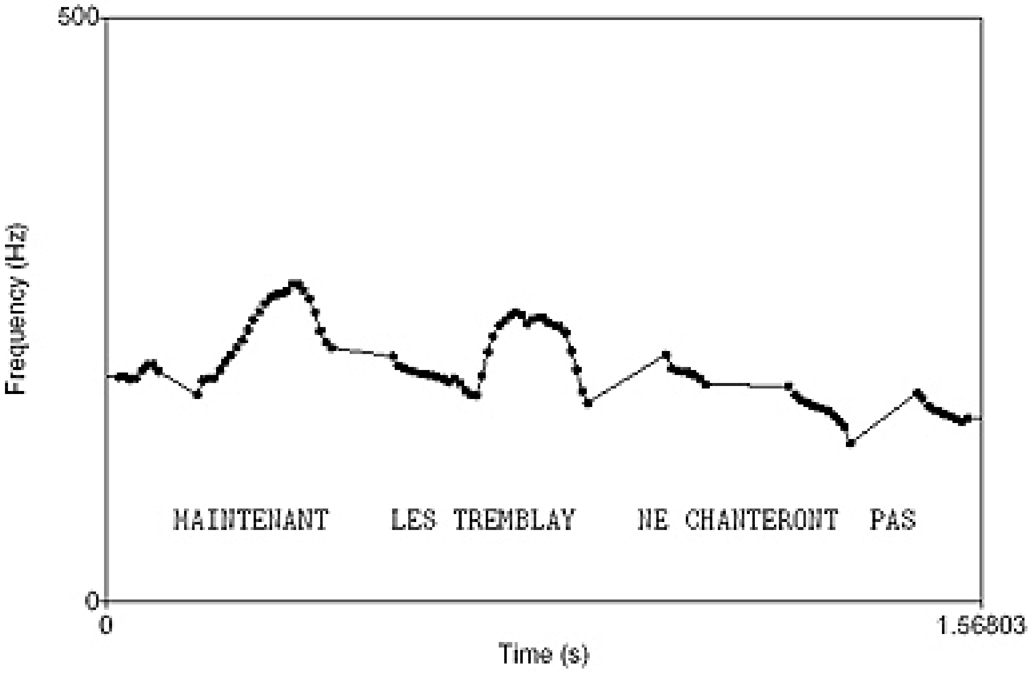


Figure 6: The pitch curve associated with the phrase "Maintenant Les Tremblay ne chanteront pas"
(*Now the Tremblays will not sing*). The words of the sentence are overlaid with the rises and falls in pitch.

This operation was carried out for a number of sentences, and general trends were observed and solidified empirically into a set of rules. Firstly, it was noted that the curves could be approximated by about eight pitch levels. It was also discovered that there are three levels of prosodic phenomena occurring. At the first level, each word has its own prosodic curve. At the second level, each type of phrase structure has a prosodic curve. So, for example, a noun phrase (NP) is normally composed of a determiner (e.g. "the") and a noun (e.g. "dog"). Every noun phrase has a particular curve associated with it that is described by a lowering of pitch after the determiner and then a rise again after the noun. The phrase level and word level events are cumulative, that is to say, a rise on the word and fall on the phrase are additive. At the third and final level, there is an overall phenomena that occurs on the entire sentence. An example of an overall curve in a declarative sentence is shown in **Figure 7**.

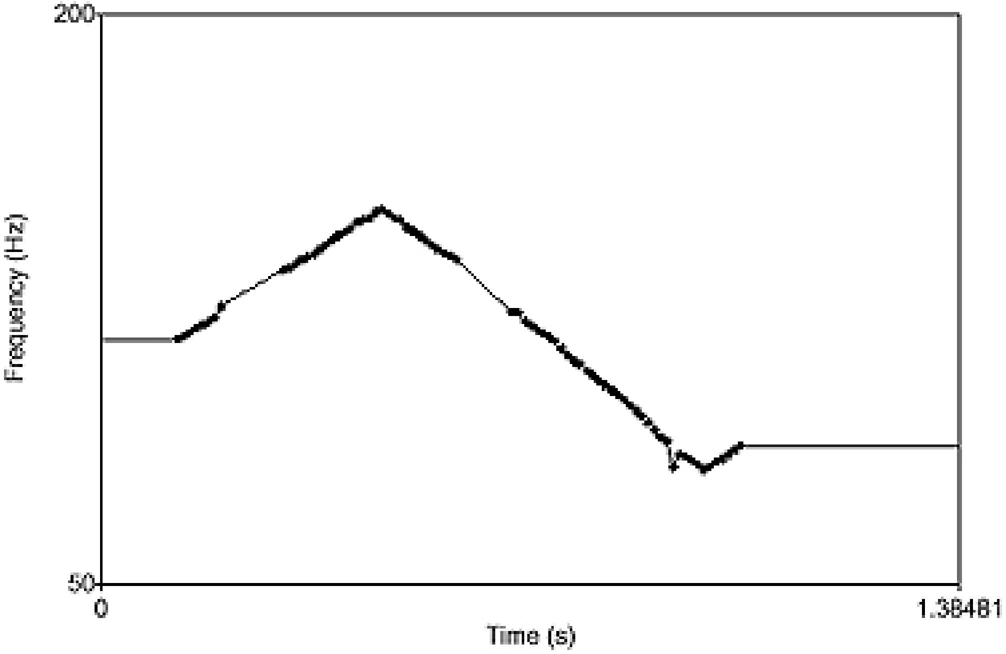


Figure 7: An example of a declarative sentence level prosodic curve. All declarative sentences are characterized by a
start at or near pitch level 1 (normal resting voice) followed by a rise near the center of the sentence
preceded by a drop to a pitch level of 0 (or sentence final pitch).

When all three levels are accumulated, we get an overall prosodic curve such as the one shown in **Figure 8**.

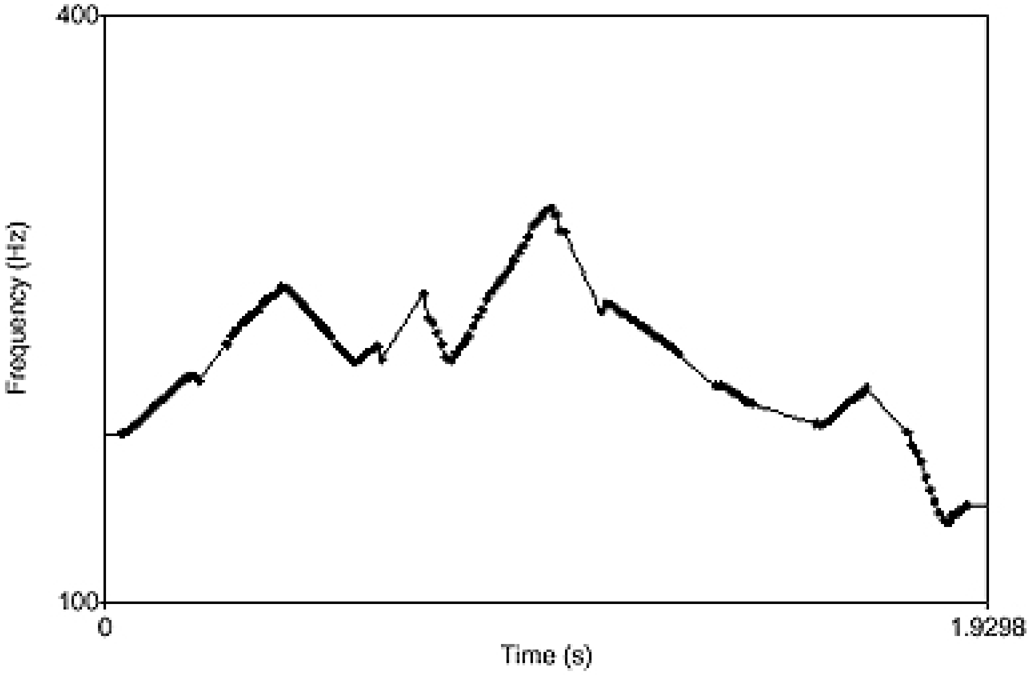


Figure 8: An example of a synthesized utterance containing word, phrase, and sentence level prosody. Notice that

the overall effects are cumulative to create localized prosodic events.

To add prosodic information to VINCI, a series of 28 "prosodic letters" were added to the French lexicon. These letters describe rises and falls in pitch as well as stress symbols. The words consisted of absolute pitch levels (expressed in Hz) and relative changes which capture steps up and down from the current pitch level. For prosodic events occurring inside of words, the letters are placed directly inside the phonological forms in the lexicon. Phrase and sentence changes are included as new "words" in the lexicon and can be incorporated into the syntax. Additional prosodic words were defined and placed into the syntax to add stress to some words that required them. For example, in French, the last word of each sentence normally has a stress on the last syllable.

New sentences were now generated in which all prosodic markings were introduced automatically. The newly-generated sentences were synthesized using MBROLA and a series of judges with linguistic backgrounds were asked to listen and evaluate the quality of the results. The types of sentences produced were similar to the ones produced originally and consisted of interrogatives, declaratives, and imperatives. Each judge was given a random selection of sentences and asked to rank on a scale the quality of the output. In spite of some trivial comments from reviewers regarding simple changes, the majority of the judges responded favorably to the synthesized sentences indicating that they believed the sentences to sound natural.

The current results show that our approach is very promising for shorter sentences, but yields no information on how longer sentences sound with prosodic events. Obviously, a future study must investigate the extension of our results to longer and more complex sentences. In addition to exploring longer sentences, research can also begin to tackle the problem of generating emotional speech. Neutrally intoned sentences are acceptable for a wide variety of applications, however, emotional speech would be beneficial for some. Imagine a fairy tale that is read with emotion conveying anger, excitement, nervousness, or joy. Such emotional events would help to make a fairy tale more compelling and could also help individuals understand the mood of the story. Another concurrent research stream deals with the area of speech recognition. A phonetic parser could be utilized to extract relevant phonetic information from spoken input and produce the phonological equivalent. Such a parser could be utilized in order to have second language learners speak a response to the system. Such a system could help diagnose errors in second language learners and could help correct pronunciation problems. The current VINCI system could also be extended to speak a sentence and have a student type in the associated orthographic stream for verification in a simple dictation exercise. As we can see, there are a number of future directions and applications stemming from this research.

References

1

Di Cristo, A. 2000. A prosodic model for text-to-speech synthesis in French. In *Botinis, A. (Ed.) Intonation: Analysis, Modelling and Technology*. Kluwer Academic Publishers.

2

Dutoit, T. An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers.

3

Dutoit, T. 1997. High quality text-to-speech synthesis: An overview. *Journal of Electrical and Electronics Engineering*, 17, 1, 25-37.

4

Dutoit, T. 2002. The MBROLA Project. <**http://tcts.fpms.ac.be/synthesis/mbrola.html**>

5

Klatt, D. 1987. Review of text-to-speech conversion for English. *Journal Acoustic Society America*, 82, 737-793.

6

Levison, M., and Lessard, G. 1996. Using a language generation system for second language learning. *Computer Assisted Language Learning*, 9, 2-3, 181-189.

7

Sluijter, A.M.C., Shattuck-Hufnagel, S., Stevens, K.N., and Van Heuven, V. 1995. Supralaryngeal Resonance and Glottal Pulse Shape as Correlates of Stress and Accent in English. In *Proceedings of the XIIIth International Congress on Phonetic Sciences*, 2, 630-633.

Biography

At the time this article was written, Craig Thomas (**craigt@webhandcentral.com**) had successfully defended his MSc and was working as Chief Technical Officer of WebHand Central Inc, a web application solutions company. Craig previously received his BScH in Computing and Information Science from Queen's College in 2001 and won the Distinguished Undergraduate Thesis Award for a topic on manuscript reconstruction from fragments. Craig's areas of interest in computer science include Internet technology, low level computer architecture, and natural language systems. In his spare time, Craig enjoys swimming, playing guitar, and travelling.