Canadian French Text-To-Speech Synthesis: Modeling an Optimal Set of Realizations for Dialect Markers

Evelyne Tsoukermann, Lucie Ménard, and Marise Ouellet

Bell Laboratories - Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974-0636, USA
[evelyne,lucie,marise]@research.bell-labs.com

ABSTRACT
We report on the development of a text-to-speech system for Canadian French which builds on both our theoretical and practical experience from the European French system [8, 16]. In order to extend to an additional dialect, we took an approach which would permit us to rapidly build the new TTS system. Specifically, we (1) identified speech markers for the new dialect, (2) determined optimal implementation points for these markers, at the phone inventory, dipphone recording, text analysis, and duration levels. Potential units were filtered based on generalizable spectral properties and on frequency of occurrence.

1. TOWARDS LINGUISTIC OPTIMALITY
The construction of a Canadian French TTS system based upon the existing European French one raised two main issues: on the one hand, we needed to model as many phonetic realizations as possible which are specific to French Canadian in order to reach the highest level of quality; on the other hand, we needed to determine the implementation issues so that the development of the new system could be as fast as possible. It was essential to optimize the choice of phonetic elements to be considered.

In fact, we took into account a minimum number of elements, among all the distinctions existing between the two dialects of French, which are strongly related to the perception of the French Canadian dialect. The chosen elements are called geographical markers which are defined as cues indicating dialectal group membership [3]. These markers are of top importance from a perceptual point of view, that is, they are related to the listener’s perception of the “accent” of Canadian French speakers [4]. It is also shown that these markers vary among speaking styles. Indeed, considering the differences between European and Canadian French, the more formal the speaking style is, the less such dialectal markers will be realized. Benefiting from earlier studies [6] which have focused on phonetic elements specific to the Canadian variety of French, the following markers have been included in our system and examples are given in Table 1:

1. Nasal vowel realization. Nasal vowels in French Canadian show a difference of quality, when compared to the same vowels in French. For instance, the mid-vowels [ɛ] and [ɔ] are sometimes perceived as being higher in Canadian French. The nasal back low vowel [ɔ] in European French, is often fronted in Canadian French. Moreover, the central vowel [ɛ], even though still present in reference books, is nevertheless not produced anymore in most areas of France, whereas it is still active in French Canadian.

2. The realization of mid-oral vowels. The phonological opposition between mid-vowels (high and low) is rendered by a different distribution of the mid-high [ɛ] or [ɔ] versus the mid-low [ɔ] or [ɛ]. In fact, [ɛ] or [ɔ] sometimes appear in Canadian French instead of [ɛ] and [ɔ] in European French.

3. The presence of an oral back and low vowel. It is shown that the lacking of the front low vowel [ɔ], producing [ɔ], is not realized in European French, particularly in word final position.

4. Intrinsic and contextual lengthening. In the Canadian variety of French, some vowels such as [ɔ, ɑ, a, e, ë, o], [ɪ, ʊ, ʌ] have inherited an intrinsic lengthening from an earlier stage of French, whereas this duration does not occur in the European dialect. Note that the mid-vowel [ɛ], theoretically the lengthened counterpart of [ɛ], is realized with a lower aperture, which thus brings a particular timbre to the vowel. This phenomenon has to be taken into account, whether by adding new units or by adding rules to the duration module.

5. The tense/lax opposition for high vowels. We know that the three tense high vowels [i y u] are in contextual alternation with the lax counterparts [ɪ Y U]. Whereas the lax ones appear in a word-final syllable (optional in non word-final) closed by any consonant except for the lengthening phones [ɪ, v, z, ʒ], the tense ones are realized in any other cases. The lax ones are characterized by shortened duration and more open aperture, thus creating some specific spectral properties.

6. The affrication of oral stop dental consonants before high front vowels and glides. Dental consonants [n] and [d] are pronounced with frication when followed by a front high vowel [i], [ɪ], [y] and [Y] or a front glide [j] and [ŋ], thus yielding to the sounds [ɲ] and [ŋj]. The phenomenon sometimes spreads over the word boundaries.

2. IMPLEMENTATION
The six markers listed above needed to be generated by the system in modifying the existing modules. Therefore, once the optimal selection of linguistic elements was thoroughly reviewed, we looked for the best way to implement these elements, within the different parts of the system.

2.1 Overview of the System
The Bell Laboratories European French TTS system was designed following the architecture of multilingual TTS
Table 1: Examples of marker realizations in European (EF) and Canadian French (CF).

<table>
<thead>
<tr>
<th>Marker</th>
<th>Word</th>
<th>IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Nasal vowels</td>
<td>lundi</td>
<td>/l̪diː/</td>
</tr>
<tr>
<td>2 Mid-oral vowels</td>
<td>alphabet</td>
<td>/aˈlɑfət/</td>
</tr>
<tr>
<td>3 Back low vowel</td>
<td>Canada</td>
<td>/kə′nædə/</td>
</tr>
<tr>
<td>4 Lenghtening</td>
<td>aise</td>
<td>/ez/</td>
</tr>
<tr>
<td>5 High vowels</td>
<td>prētrīz</td>
<td>/prɛtʁi/</td>
</tr>
<tr>
<td>6 Affrication</td>
<td>fortuit</td>
<td>/fɔrti/</td>
</tr>
</tbody>
</table>

systems [7]. Four main modules are part of each machine: the diphone or larger unit inventory or stored acoustic units, the text analysis module, the duration model, and the intonation model. We report on the changes made in each of these modules. The selected dialect markers were handled by the system either (1) by the addition of phones, or (2) by the addition of rules in the grapheme-to-phoneme module, or (3) by the construction of a completely new duration model. We describe, in the following sections, the steps involved in each process.

2.2 Diphone Inventory
This section presents the methodology used in order to select the units for the phone inventory. It will also be shown that each step involves an optimal selection of criteria. First, whereas the European diphone inventory is made of 36 units (including 14 vowels, 18 consonants, and 3 glides), the French Canadian inventory consisted of a total of 43 phones - 20 vowels, 20 consonants, and 3 glides. For vowels, we count 4 nasal vowels and 16 oral ones. At this stage, lax high vowels [I Y U], corresponding to marker 5, as well as the affricates [tʃ] and [dʒ] (marker 6), have been added to the inventory of units to be recorded. The mid-long vowel [ɛ], for words such as “prétre” [prɛtʁ], having a specific quality in French Canadian, is also part of the new subset of units. The low back vowel [a] and the nasal [ɛ], which are not realized by the European system, have also been generated and recorded for the Canadian version.

2.3 Recordings and Corpus
We recorded a completely new database described earlier with a professional native Canadian French speaker, in an anechoic room. The corpus consisted mainly of logatomes, embedded in the same carrier sentence: “C’est X que je dis” (“It is X that I say”) where X is the logatome. In order to account for inter-word as well as intra-word phenomena, we took all the combinations of 43 phones (43^{43}=1849) and removed most of the non valid pairs, such as V + glides (where V is a vowel), stop + stop, or fricative + fricative. Most diphones were placed in a phonotactically balanced logatomes made of the vowels /a/ and the consonant /t/. That is, for the diphone /tv/, the logatome is /atVtɑ/, for CV (where C is a consonant), the logatome is /tCVtɑ/, for CC /tCCtɑ/, and so on. For VC diphones, we used a closed syllabic structure with the consonant in coda position, since the position has strong timing influence on the consonant. For instance, for the diphone “tv”, the logatome is “atovta” and not “atovata”.

With regard to context-sensitive units, the selection was optimal as well. Due to their spectral unsteadiness, French phones can be roughly divided in 2 groups, the stable and the unstable phones. Unstable phones (U) are typically the glides and the liquids whereas stable phones (S) are the vowels and the other consonants. Additionally, in the context of voiceless steps, liquids tend to lose their voiced quality. In order to generate diphone units which would present the spectral properties of their surrounding segments, diphones containing an unstable phoneme were recorded in triphones, were the unstable unit is followed or preceded by a stable phone, i.e., SUS. This stable phone representing the context consisted of 4 different – in their level of opening – oral vowels [i e a], 1 nasal vowel [ɛ] and 2 occlusive consonants – anterior and posterior [p k]. The total number is close to 1000 units.

Finally, we wanted to capture polyphones in frequently occurring function words, in order to reach a high quality synthesis. Polyphones were extracted from top frequency function words in French, and consisted of 3-phone to 5-phone units in inter- and intra-word position. Following Bimboz’s model [1], we decided that a polyphone candidate was any number of unstable phones surrounded by a stable phone. Polyphones were selected according to acoustic and textual criteria. The acoustic criteria relies on the acoustic stability of the phones and their position in the polyphone. The textual criteria deals with the occurrences and the frequencies of the polyphones. In order to collect these polyphones from textual data, the following steps were involved:

1. Identify function words and take the top 200 most frequent ones. We used our large collection (35 million words) of French corpora (such as Le Monde, the HANSAR corpus, and other French newspapers);
2. Extract collocation of function words from 2 to 4 words. For 2 word collocations, we got 7,000 occurrences, for 3 words, 14,000 occurrences, and 315,000 occurrences for 4 words. For instance, a 2-word collocation consisted in the sequences “pour du”, “bien trop”, a 3-word would be “et bien a”, “il y a”, and 4-word collocations are of the following type: “de dessous par des”, “car comme pour la”.
3. Convert orthographic words to phonetic words, using the new module of text analysis updated with the Canadian rules described in the following section. This reduced the number of occurrences from 10 percent for the 2 word collocations to 50 percent for the 4 word collocations;
4. Extract all the polyphone candidates, such as SUS, where U is of length 1 to 3. A total of 1,000 SUS triphones, 460 SUUS 4-phones, and 37 SUUS 5-phones were retained;
5. Select the top frequency polyphones in each category; 780 polyphones were retained.

To sum up, a few thousands diphones and polyphones
were recorded, segmented on an SGI machine using Waves software analysis, and clustered automatically.

2.4 Text Analysis
This module is the front end of the system; it takes the input orthographic strings and converts them into phonetic symbols. Additionally, syllabification and stress marks are also assigned to the phonetic strings. The text analysis model designed for the European French system consists of several hundred rules [9] for handling pronunciation of French words. More specifically, it is divided into four components, (1) the morphological analysis of words, (2) the language models used to disambiguate homographs and other contextual phenomena such as liaison, (3) the grapheme-to-phoneme conversion rules, and (4) the syllabification and stress assignment.

All the changes necessary for the Canadian version occur in the third component, i.e., at the level of the grapheme-to-phoneme rules, where phonemes specific to Canadian French are implemented.

As mentioned above, even though most of these differences used to exist in European French, the nowadays spoken language tends not to maintain anymore some distinctions such as the open /ɛ/ vs. closed /e/ and tends to assimilate the front rounded nasal [ɔ] to its front un-rounded one [ɔ].

The modification of the rules occurred, for instance, in the following:

1. variation between mid vowels [ɛ] and [ɔ] in European French and their counterpart [ɛ] and [ɔ] in Canadian French, for instance in word final position, where the orthographic string “ée” in “respect” (respect) is realized [ɛ] in European French and [ɔ] in Canadian French.

2. realization of the nasal [ɔ] instead of [ɛ] in the European system, for words such as “un” (one) [ʊ], or “lundi” (Monday) [lud̩].

3. laxed high vowels [ɪ], [ʏ], and [ʊ] were also generated by rules, converting graphical representations such as (“ou”, “ui”, “uy”, etc.) in word final syllables closed by a consonant not subject to lengthening effect. This is the case of the vowel “ii” as in the word “vite”, lax in Canadian ([vɪt]), but tense in European ([vɨt]). However, in “vite”, the vowel “i” remains tense in both varieties ([vɪt]), since the syllable is closed by a consonant (in this case [i]) having a lengthening effect on the preceding vowel.

In order to keep versions of the text analysis updated in the two systems, we used a modular architecture, consisting of ordered rule files reflecting usage of both languages. These files are called by a main procedure keeping track of the updates. A total of about 50 rules were added or modified to account for Canadian French.

2.5 Duration
One of the main distinctions between European and Canadian French lies in the duration. It has been shown that the temporal organization of segments as well as longer units, such as words and phrases, are different from one variety to the other. Intrinsic duration, associated to many words, plays an important part in the perception of French Canadian accent. In order to take into account these facts, we had to design a new duration model, based on representative data. The duration model relies on evidence extracted from data, where a phone is represented by a vector of features [11]. Each feature corresponds to the factors selected in the studied language. A multiplicative model computes a value for each of the phone.

2.5.1 Corpus selection The duration model is based on duration measurements performed on a set of 219 sentences that have been selected by a greedy algorithm from a 500,000 word corpus. The set of texts consisted of excerpts of “Le Monde” Newspaper [2]. The utterances were converted in phonetic strings by the text analysis module. Only the declarative sentences of a length varying from 6 to 25 words were retained. Prior to the application of the greedy algorithm, the factors having an important effect in duration in French Canadian have to be carefully selected. Indeed, a factor defines a characteristic feature of the language and can sometimes be divided into subcategories or levels. The levels refer to the value of a factor. For instance, the factor (or feature) “voicing”, can be given the value voiced or unvoiced. The choice of factors was inspired by our knowledge of temporal properties of French Canadian [5].

First, the existence of some intrinsically long phones in Canadian French, shows the need to include some segmental criteria in the factor selection (for instance, [a], by nature longer than [ɛ]). Moreover, the influence of adjacent units belonging to the same syllable, yields to consider contextual factors, mainly the preceding and following phones. When the factors deal with phone position, classes of phones have been defined, based on their common behaviour. Therefore, we grouped the 43 original phones in the following 13 classes:

- Vowels: [i, y, u] [l, Y, U] [ɛ, ɨ, ɛ, o, ɔ, a, ɔ] [a, ɔ, ø];
- Consonants: [p, t, k, tʃ] [b, d, g, dʒ] [ʃ, s] [v, z, ʒ] [n, m, p, θ];
- Glides: [j, w, y].

However, word- penultimate vowel lengthening, characteristic of Canadian French, (such as “beauté”, [bœtɛ] “beauty”), where the first vowel [o] is lengthened, has to be taken into account in the corpus selection process. For instance, we grouped, under the category of “position from the right”, a set of elements with codes a maximum of positional information: position (1) of phoneme in syllable, (2) of syllable in word, (3) of word in phrase, (4) of phrase in utterance. Each of the 4 subfactors is associated with the values “final position”, “penultimate position” and “internal position”. Similarly, the factor “position from the left”, consists of the same 4 subfactors. Each subfactor can have one of the following levels: “initial position”, “second position”, or “internal position”.

Finally, the temporal behavior of segments in Canadian French varies as a function of its position, i.e. onset, nu-
class, or coda. The factor referring to the type of syllabic component, that is, onset, nucleus or coda, was thus taken into account. Since different temporal properties can be observed based on the effect of clitics, we also classified each word as being a content word or a function word. When it is a function word, the element was coded when surrounded by 0, 1, or 2 function words.

As a result, 13 factors were retained. Each phone of the phonetically transcribed corpus can be represented by a 13-feature vector, where each feature receives a value. We run a greedy algorithm [12] using as input the large news corpus; the algorithm covers all the existing events with the smallest number of sentences. The sentences were recorded by the same native Canadian French speaker who recorded the unit inventory. They were downsampled and segmented on an SGI machine, using the Waves speech analysis software. Then, the duration measurements were performed.

2.5.2 The duration model. The 219 sentences yielded 13,310 phones on which the duration model was built. The model resulted in a tree representing subsets of phones grouped by the similarity of their behavior. Among the set of selected factors stated above, the following were retained as having an influence on duration:

- **Vowels:**
  1. Accentuation: this is a positional factor, since the accent in French is generally on the word-final syllable;
  2. Position of syllable into the word: the values “final”, “penultimate” or “internal” were assigned since right has more influence than left;
  3. Syllabic structure: open or closed syllable;
  4. Type of segment: the phone identity consisting of the terminal nodes of the tree was given its duration values.

- **Consonants:**
  1. Consonants were split between glides, [l], [r], nasals, stops, and fricatives;
  2. Voicing vs non-voicing;
  3. For occlusive consonants, closure and burst;
  4. The liquid [l] was separated based on its onset or coda position;
  5. The liquid [r] was split depending on its following and preceding phones.

The statistical model provided very realistic durations for Canadian phones. Indeed, we have reached correlations up to 79%, with an average of 69%. Even though this value seems quite low, note that it is applicable on sets of a few hundred units. Besides, we have reduced the total number of levels (values) for each factors, by mapping those which showed a similar behavior. This certainly adds to the generalizable value of the model.

2.6 Intonation

There are differences in the intonational contour of European and Canadian varieties of French, but the current model would need to be refined in order to handle the specificities of Canadian French. In fact, we can observe that the range of fundamental frequency between peaks and valleys, within an intonational phrase, as well as the global effect of declination (decrease of fundamental frequency on the overall sentence) are realized in distinct ways from one dialect to the other.

3. CONCLUSION

The first full-fledged Canadian French TTS system was recently built based on our knowledge and experience with European French. We quickly determined the main differences between the two dialects, identified the specific markers, and implemented the new modules. Within five months, a new comprehensive system was built that handles unrestricted text. The system was widely tested on a variety of texts, and evaluated by two listeners. We are currently considering more extensive system evaluation with users who would give their input for improvement.

REFERENCES