Flavoured Acoustic Model and Combined Spelling to Sound for Asymmetrical Bilingual Environment

R. Lejeune, J. Baude, C. Tchong, H. Crepy, C. Waast-Richard

IBM Pervasive Computing, European Voice Technology Department
{remi.lejeune, jerome.baude, ctchong, crepy, waast}@fr.ibm.com

Abstract
The most common target of multilingual ASR aims at covering various speakers from various languages. The problem we address in this article is more specifically an asymmetrical bilingual scenario, where the same speaker may insert in his speech some foreign words using foreign pronunciations. This is a frequent situation for French as spoken in Canada, where English proper names are often spoken using English pronunciations. We explore in this article a new way of using multilingual models by enhancing a monolingual system in a measured manner (Flavoured Acoustic Models). We also present an innovative bilingual spelling to sound system based on separate decision trees, providing balanced alternatives in both languages. Our ASR results over the telephony channel show that both technologies associated with one another outperform by up to 75% monolingual systems on English pronunciations without degrading word error rate on French pronunciations.

1. Introduction
The recent works on multilingual ASR have shown several benefits of this approach: possibility to bootstrap new languages with little specific data [1], increased robustness on foreign words [2]. However, the accuracy may degrade when modeling the same phones using data from different languages together with a multilingual lexicon which increases the acoustic confusability [3].

The problem we address is ASR in an asymmetric bilingual context: in such a cultural environment, speakers of one language (hereafter called main language) often use words of another language (influencing language) and pronounce them with foreign pronunciation.

A common example is the French language as spoken in Canada (Fr_CA). Many names used in Canada are English, and speakers often say them using an English pronunciation. These names are typically very important for ASR applications like directory dialers or call routers.

A trivial approach is to train a monolingual acoustic model, using “good enough” approximate phonetic transcriptions for the influencing language words. This approach thus fully relies on the flexibility of the statistical learning process in order to capture the native pronunciations of the influencing language words. This involves the risk of poorly recognizing those words because of the confused modeling of the phones that make them up.

Another approach is to train a fully bilingual acoustic model. But this can be at cost of significant accuracy degradation for the main language [3]. Therefore we explore the way of an asymmetrical acoustic model that we later name Flavoured Acoustic Model (FAM), which aims at enabling English pronunciations for French speakers native from Canada, without degrading the recognition on French pronunciations. We describe this process in section 2.

The issue of words unknown to the basic lexicon also needs to be addressed. Indeed, ASR systems for telephony often rely on dynamic vocabularies which phonetic transcriptions cannot be easily updated by experts. They require a spelling to sound system (phonetizer). In replacement of monolingual decision tree based spelling to sound (STS), we explore bilingual alternatives that would provide satisfactory phonetic transcriptions both in English and French.

We present our results for both techniques in section 4.

2. Flavoured Acoustic Model

2.1. Overview
The process of building FAM is somewhat similar to the bilingual one, except that it aims ultimately at training the system with main language speakers only (Fr_CA speakers in our case).

Our FAM building process runs as follows: we first define the phone alphabet covering English and French. Using separate lexicons for a US English corpus and our Fr_CA corpus, we bootstrap the English phones. We then enhance the lexicon of the Fr_CA corpus with English phonetic transcriptions for proper names. We further iterate the acoustic building process, keeping a minimum amount of US English data. We perform what we will define as “native boosting” in order to put an emphasis on learning Fr_CA speakers uttering English pronunciations.

2.2. Design of the phone set
As a full bilingual system, the FAM approach requires the use of a phone alphabet that describes the main language as well as the influencing language. We have used the common alphabet described in [3], which benefits have already been demonstrated for multilingual systems [4]. This alphabet includes overlaps, as described in Figure 1.

Let us define:
Class M: phones exclusively used in the Main language
Class C: Common phones (used in both languages)
Class I: phones exclusively used in the Influencing language

From a monolingual model for Fr_CA, we need to create a new model (it0) enhanced with class I phones. This new model

![Figure 1: common phone set for Fr_CA and English](image_url)
should preserve the monolingual representation of the common phones. We achieve this goal by including in our training data some US English data, aligned with a US English acoustic model, using a separate English lexicon. In order to avoid corrupting the class C phones when they will be used for decoding French pronunciations, the amount of foreign training data introduced does not exceed 33% of the overall training data. During the acoustic model training, the contextual phonetic decision tree plays an important role in preserving and discriminating the main language representation of the common phones. Special care is taken to ensure that the question set proposed when growing the decision tree does not lack of “language discriminative” questions (for instance questions covering exactly class M or class I phones).

2.4. Training lexicons

At the bootstrap stage (it0), the lexicons are clearly separated for US English corpus and Fr_CA corpus, therefore class M and class I phones are trained on separate data. In further iterations (named it1, it2…), the lexicons remain separated. However, phonetic transcriptions covering English pronunciations are being introduced in the training lexicon when aligning Fr_CA speakers, but remain limited to legitimate candidate words like proper names.

2.5. Boosting native speakers aligned with foreign phones

The specificity of FAM approach is that the proportion of the foreign data is meant to decrease progressively, as the influencing language phones get more and more supported by main language data. In order to progressively focus the parameters of the new phones on native speakers, we boost this legitimate representation of these phones during the Forward-Backward re-estimation. During the alignment step, the best matching phonetic transcriptions of the words in the transcriptions are chosen. At this stage, we can identify the utterances where the influencing language phonetic transcriptions are proposed by the alignment process. This is quite equivalent to spotting the class I phones (for instance, more than 95% of the English proper names phonetic transcriptions contain at least one class I phone). The boosting consists of applying a bigger weight on the collected counts of mean/variance/priors for the utterances including influencing language phonetic transcriptions. This is done by duplicating these utterances several times. Therefore, the re-estimated mixtures of Gaussians for the class I & C phones are getting “closer” to the native speakers rather than the foreign speakers initially supporting them. We name this technique “native boosting”.

2.6. Iterative building process

The models are being developed iteratively. We compute the new alignments based on the previous iteration model. We name our models iterations it1, it2, it3… We name the “native boosted” models derived from the first ones it2boost, it3boost, etc.

Table 1 shows the amount of data used in the training of some of the models. The amount of Fr_CA data has been increased along the process, while we have decreased in proportion the amount of US data. Wherever native boosting was applied, we indicate between brackets the number of utterances taking into account the duplication process. it4boost2 is a model which parameters are fully re-estimated without using foreign data.

As a result of this “native boosting”, when re-aligning with the new acoustic model, and iterating this process several times, the influencing language phonetic transcriptions seem to be progressively adopted. Figure 3 shows, for each acoustic build, the percentage of individual words uttered by Fr_CA speakers in our training data that are being aligned with phonetic transcriptions containing at least one class I phone.

3. Combined bilingual spelling to sound

Given the bilingual environment, unknown words need to be phonetized using an adequate spelling-to-sound (STS) subsystem. The additional issue could be the unknown language-origin of these words. In our Fr_CA example, unknown words such as relatively rare proper names can be pronounced the French way or the English way, or both. Sometimes the spelling gives a clue as to which language applies (“Trudeau” versus “Woolsworth”), whereas the language identification is sometimes ambiguous (“Pinton”). A usual monolingual approach to STS is to train a decision tree on a corpus of aligned spellings and phonetic transcriptions. At runtime, for each letter of an unknown spelling, the tree asks questions of the surrounding letter contexts and the leaf reached determines the probability distribution of the phonemes potentially output by that letter.
The progression through the spelling expands a ranked list of candidate transcriptions, which can be pruned according to score criteria [5], [6].

For an asymmetrical bilingual system, multiple approaches have been considered, including:

- using an STS system trained on only the main language (monolingual STS)
- using an STS system trained on the union of lexicons for both languages (bilingual single tree STS)
- using two separate STS systems, each trained on a particular language (combined two trees STS)

A STS system trained on French phonetic transcriptions will fail to produce appropriate transcriptions for English words. The "lump it all together" STS system will perform better but is far from optimal, because it is indiscriminate, having been trained on a set twice as fuzzy in its letter-to-phone associations and potentially phonetizing the end of a word according to language Y's rules even though the beginning is being phonetized as language X. Separate STS systems for each language are much crisper.

We could, in a bilingual ASR context, phonetize all unknown words using both STS systems, and use all candidate transcriptions. This is sub-optimal and over-generative: in an English-French combination, for example, "French-looking" words should favor the French STS subsystem, so the outputs of the English subsystem should be discounted for these.

To address this issue, we recombine and prune the candidates coming from the separate STS subsystems using their scores. We incorporate that recombination inside the letter-expansion loop, since it allows the application of a common pruning mechanism: after the letter has been expanded into ranked and scored proposed phonetic transcriptions in both languages, we re-sort these transcriptions into a common list and prune the less likely hypotheses. A rescaling factor can be applied to each subsystem's score, either to artificially bias the global system in favor of one language, or to compensate for imbalanced dynamic ranges of scores of the individual subsystems. For instance, English being less regular in STS than French tends to produce lower scores overall.

Since we are allowed to use multiple candidate transcriptions for vocabulary words we can over generate. And not all phonetic transcription errors have the same degree of severity (due to acoustical phonetic confusion, for example). As usual for STS systems used in ASR, the ultimate measure of success is in the decoding error rate. Our results are presented in section 4.2.

### 4. Decoding results

All acoustic models that we test use LDA+MLLT front-end, 1300 context dependent phones and 100k gaussians. They are trained on the same Fr_CA dataset except FAM it4boost which is also trained with US data as described in table 1.

The flavoured decoding lexicon introduces influencing language phonetic transcriptions only for relevant words (for Fr_CA we introduce English phonetic transcriptions for proper names).

Our test set matches one of our target applications, a telephony name dialer. It includes English proper names, which are uttered using an English pronunciation. It also includes French proper names in order to check whether degradation is observed for French pronunciations. It consists in 2129 utterances, 1423 covering company names and 1116 covering city names in Canada. The vocabulary size for both tasks is over 1000 words. Each utterance is listened to and sorted into three categories: French pronunciation (labeled “French pronun.”) and Other (labeled “Other”) covering mixed or unclear pronunciations. Speakers are French-speaking inhabitants from Canada.

#### 4.1. FAM versus monolingual approaches

We compare FAM versus two different monolingual models. M1 was trained with default (naive) phonetic transcriptions for all words, including those which could apply for an English pronunciation. M2 uses a lexicon with enhanced phonetic transcriptions to better approximate English pronunciations, but still restricted to class M & C phones.

<table>
<thead>
<tr>
<th>model</th>
<th>M1</th>
<th>M1</th>
<th>M2</th>
<th>it4boost</th>
<th>it4boost2</th>
</tr>
</thead>
<tbody>
<tr>
<td>lexicon</td>
<td>mono-lingual</td>
<td>enhanced mono-lingual</td>
<td>enhanced mono-lingual</td>
<td>flavoured</td>
<td>flavoured</td>
</tr>
<tr>
<td>FR pronun.</td>
<td>5.0</td>
<td>6.0</td>
<td>5.5</td>
<td>5.3</td>
<td>5.4</td>
</tr>
<tr>
<td>other</td>
<td>19.3</td>
<td>11.5</td>
<td>11.5</td>
<td>9.7</td>
<td>8.8</td>
</tr>
<tr>
<td>EN pronun.</td>
<td>42.7</td>
<td>24.4</td>
<td>18.8</td>
<td>10.9</td>
<td>11.7</td>
</tr>
<tr>
<td>overall</td>
<td>16.4</td>
<td>11.3</td>
<td>9.7</td>
<td>7.4</td>
<td>7.5</td>
</tr>
</tbody>
</table>

Table 2: Word error rates (WER) on labeled corpus

We observe that M1 encounters difficulties to cope with English pronunciations (42.7% WER). This is easily explainable by lack of phonetic coverage. Using enhanced phonetic transcriptions like for M2 training, we reduce the WER to 24.4%, with a slight negative impact on the French pronunciations. Results are again slightly improved when training lexicon matches decoding lexicon (M2). Eventually, the best results are being obtained with FAM (it4boost), bringing 75% relative WER improvement on English pronunciations compared to the initial approach.

In order to assess the remaining contribution of US English in this error rate reduction after several FAM iterations, we also tested it4boost2. Forward-Backward re-estimated from it4boost without any US English data. The results remain relatively unchanged, which seems to indicate that the class I & C phones are well trained with Fr_CA data.

We also verify that the FAM performs as well as M1 on other telephony tasks not involving English pronunciations. This 2000 utterances test set is made of a wide variety of tasks such as booleans, digits, numbers, dates or times. We show the results in table 3.

<table>
<thead>
<tr>
<th>model</th>
<th>M1</th>
<th>it4boost</th>
</tr>
</thead>
<tbody>
<tr>
<td>lexicon</td>
<td>mono-lingual</td>
<td>flavoured</td>
</tr>
<tr>
<td>French only test set.</td>
<td>5.2</td>
<td>4.9</td>
</tr>
</tbody>
</table>

Table 3: WER on general French corpus
The slight difference observed in table 3 between monolingual and flavoured can be considered insignificant as it remains within the range of our 95% confidence intervals.

### 4.2. Spelling to sound results

We compare in table 4 several versions of STS. The baseline consists of the expert flavoured lexicon, which performances are already exposed in table 2. Quite logically, the monolingual STS performs poorly on English pronunciations. The bilingual single tree STS brings large improvement on English pronunciations, but degrades French pronunciations as a side effect. This can be explained by the drawbacks of this hybrid approach described in paragraph 3. The best results on both sides are eventually obtained with the combined two trees STS, which equalizes the expert flavoured lexicon, both on French and English pronunciations.

<table>
<thead>
<tr>
<th>model</th>
<th>lexicon</th>
<th>it4boost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>expert flavoured</td>
<td>STS monolingual</td>
</tr>
<tr>
<td>FR pronun.</td>
<td>5.3</td>
<td>4.5</td>
</tr>
<tr>
<td>other</td>
<td>9.7</td>
<td>20.0</td>
</tr>
<tr>
<td>EN pronun.</td>
<td>10.9</td>
<td>43.8</td>
</tr>
<tr>
<td>overall</td>
<td>7.4</td>
<td>16.5</td>
</tr>
</tbody>
</table>

*Table 4: comparing WER for various STS*

### 5. Conclusion

In this article we introduced the concept of Flavoured Acoustic Model. Along this training process we closely monitored the use of the influencing language data, the main language data being progressively adopted to model the foreign pronunciations. We showed the relevance of this approach in the case of French as spoken in Canada influenced by English for some proper names pronunciations. Indeed, we observed a relative gain on word error rate from 40% to 75%. We also presented a combined bilingual spelling to sound approach, and showed that it can advantageously back-off missing entries in our lexicon. It achieved same decoding performances as an expert lexicon both on French and English pronunciations.

### 6. References


