French Speech Recognition in an Automatic Dictation System for Translators: the TransTalk Project

Julie Brousseau[‡], Caroline Drouin[‡], George Foster[†], Pierre Isabelle[†],
Roland Kuhn[‡], Yves Normandin[‡], and Pierre Plamondon[†]

[†] Centre d'Innovation en Technologies de l'Information (CITI)
1575 Boul. Chomedey, Laval, Québec, Canada, H7V 2X2

[‡] Centre de recherche informatique de Montréal (CRIM)
1801 McGill College, Montréal, Québec, Canada, H3A 2NA

Abstract

This paper describes a system designed for use by professional translators that enables them to dictate their translation. Because the speech recognizer has access to the source text as well as the spoken translation, a statistical translation model can guide recognition. This can be done in many different ways—which is best? We discuss the experiments that led to integration of the translation model in a way that improves both speed and performance.

1 Introduction

The TransTalk project attempts to integrate speech recognition and machine translation in a way that makes maximal use of their complementary strengths. Professional translators often dictate their translations first and have them typed afterwards. If they dictate to a speech recognition system instead, and if that system has access to the source language text, it can use probabilistic translation models to aid recognition. For instance, if the speech recognition system is deciding between the acoustically similar French words *cheveux* (hair) and *chevaux* (horses), the presence of the word *horses* in the English source text will guide it to the correct choice.

We have implemented a prototype of TransTalk that takes as input an English text and a spoken French translation, and yields French text. An earlier paper [1] focused mainly on the machine translation aspects of TransTalk. Since that paper was written, we have made improvements in both components of the prototype—for instance, the French speech recognition component now handles continuous speech. The current paper will focus on the unique problems of large vocabulary speech recognition in the French language [2], and features of our system designed to deal with these problems, as well as presenting experimental results.

2 The Transtalk System

We have tried out three quite different versions of our TransTalk prototype. Version 1, described in [1], was capable only of isolated-word speech recognition. The two more recent versions both carry out continuous mode dictation over a vocabulary of 20,000 French word forms and expressions. In version 2, each spoken French sentence generates a list of the 200 most probable word sequence hypotheses, which is sent to the translation module. The translation module uses its knowledge of the corresponding English sentence to choose one of the 200 French word sequence hypotheses as the final output of the complete system. In version 3, the translation module is applied much earlier—in fact, before recognition begins. From each English source sentence the translation module generates a *dynamic vocabulary* of French words likely to occur in the translation. When recognition of the spoken French sentence takes place, the recognizer is allowed to consider only words occurring in the current dynamic vocabulary. These different ways of integrating the translation module into the speech recognizer are described in more detail in section 5 (below).

In going from version 2 to version 3, we also made changes in the recognizer unrelated to the way in which the translation module is integrated into it. In the version 2 recognizer (described in detail in [4]) acoustic modeling is carried out by triphone HMMs with shared codebook output distributions (mixture Gaussian distributions with one codebook of Gaussian densities per phone). Recognition is performed in several passes. The first two passes (forward-backward) use a bigram language model and intra-word but not inter-word triphones to produce a looped word-pair graph. An N-best search algorithm then generates the 200 most probable utterances, which are rescored and reordered with inter-word triphones before being passed on to the translation module. All this computation takes about 92.6 times real-time—even before translation model rescoring takes place.

The version 3 recognizer no longer employs triphones as the basic acoustic unit. Instead, the basic unit is now the *tree state*, i.e., each state of the HMM for a given phone is represented by a context-dependent decision tree [5]. So far, our experience with tree states both in English and in French speech recognition has been extremely positive. As compared with triphones, tree states use considerably less memory and speed up recognition, while yielding better performance. The other major change made to the recognizer was to use discrete output distributions, rather than the shared codebook approach employed in version 2. The motivation here was speed alone (in general, discrete distributions yield a faster system with poorer performance).

The net effect of these two changes—both increasing speed, one tending to improve performance, the other tending to diminish it—was a baseline (i.e.

without translation module) version 3 system that took 15.8 times real-time and had slightly worse speech recognition performance than version 2. We will see that integration of the translation module yielded a version 3 system that was considerably faster and performed better than the baseline.

3 French Speech Recognition

In the system's 20,000-word French lexicon each entry has a phonetic representation based on 44 phones including 20 vowels and 24 consonants. To enhance the effect of the translation module, certain frequent expressions such as de la or tout de suite are entered as units. The base pronunciations were obtained automatically, using a set of grapheme-to-phoneme rules which take into account phonetic idiosyncrasies of the French spoken in Quebec (such as assibilation and vowel laxing), and were then manually verified.

Other idiosyncrasies of the French language have to be taken into account when building a speech recognition system. Dictionary explosion due to elision and homophones is one of them. Elision causes the last vowel of some function words (de, la, ne, que) to be omitted when the following word begins with any vowel. Our system handles these contracted units as separate words.

Homophones are more frequent in French than in English, and often cause word errors in French speech recognition. In our 20,000 word vocabulary there are 6,020 words such that for each of them, at least one homophone exists. Thus, 30% of our lexicon is made up of homophones, compared to 5% for the 19,977 WSJ word lexicon [4]. Where only one member of a set of homophones is predicted by the English source sentence, the translation module will be particularly useful (e.g. *children* predicts *enfants* rather than *enfant*).

A major challenge that must be overcome in French speech recognition is liaison. Liaison appears in continuous speech and occurs when a consonant at the end of a word, orthographically present but not pronounced in the isolated word, is pronounced in the presence of a vowel at the beginning of the next word. Whether or not this consonant is pronounced is difficult to predict—it depends on a complex interaction between orthography, syntax, semantics, and other factors.

We have tried several methods for handling liaison during acoustic training and testing. Currently, possible liaisons are derived automatically in the lexicon by a set of simple rules. For instance, a word ending in "s" (such as les) will generate the liaison phone /z/. The liaison phone is represented as a separate field in the lexicon. During training and recognition this "liaison field" will be active if the following word begins with a vowel. However, even when the field is active, the acoustic realization of liaison is optional; if it occurs, an insertion penalty is imposed. Seven consonants can participate in liaison: /z,t,n,r,p,v,k/.

Note that our rules for generating liaison are not perfect. For example, they will wrongly permit liaison after a proper noun (as in *Gervais /z/ avait deux chiens*) or even after the conjunction *et* (inserting an erroneous /t/). More work is needed here. The same is true of language models for French—long-distance agreement between words is an important feature of French, and conventional N-gram models handle it badly.

4 Description of Corpora

Most of the sentences in the acoustic training corpus came from *La Presse*, a French-language Montréal daily newspaper: 88 different people read an average of 256 La Presse sentences each, yielding 22,528 spoken sentences. 19 speakers each read 25 French sentences from Hansard, the Canadian parliamentary corpus, yielding 475 spoken sentences. Finally, 29 speakers read 43 phonetically balanced French sentences [3], yielding 1,247 spoken sentences. These 24,250 (total) spoken sentences were read in continuous mode, with verbalized punctuation.

The bigram French language model was trained on a French corpus of 2.2 million words drawn from 40 Hansard files. The translation model was trained on a bilingual corpus of 45.6 million English words and 47.0 million French words from 943 Hansard files (only the existence of a large bilingual corpus such as Hansard makes TransTalk possible in the first place!)

The TransTalk system was tested on 300 Hansard sentences (6,639 words) not used for acoustic, language model, or translation model training. These sentences meet two criteria: they contain 40 or fewer tokens (including punctuation), and they are completely covered by the recognizer's 20,000-word lexicon. They were dictated in continuous mode, with verbalized punctuation, on a Sennheiser HMD 410 by 10 speakers (5F-5M) not used for training.

5 Applying Translation Models

Given a source sentence in English and a spoken French translation of that sentence, our system must try to generate the transcription of the French sentence. A variety of translation models, all requiring bilingual training data, will yield probability estimates for French words, given the English source sentence [1, 2]. The models differ mainly in the extent to which they take into account the position of given words in the English and in the French sentence. For the purposes of this paper, one need only understand that some of these models are crude and computationally cheap, while others are sophisticated and expensive. What models should be chosen, and how can the probabilities they generate be applied to constrain speech recognition?

In version 1 of TransTalk (described in [1]) a simple combined translation-language model (TLM) is applied to the output of an isolated-word recognizer. For each acoustic token, the recognizer generates n word hypotheses; the TLM is applied during a Viterbi search to yield a French sentence with n words. The TLM has a component which predicts the next French grammatical category given the preceding two categories, as in a standard triPOS language model. However, the probability of a word, given its category and the corresponding

Model	Word % Correct	Sent. % Correct	Rank of best sent.
SR	80.7%	4.0%	54/200
trigram	84.5%	8.7%	19/200
tri + TM	86.0%	12.7%	11/200
best	90.0%	22.0%	1/200

Table 1: Rescoring N-best Hypotheses (N=200)

English sentence, is provided by a simple translation model.

Version 2 employed the N-best technique: the recognizer generated the N most probable hypotheses (based on acoustics and a bigram language model), and the translation model was used to rescore these hypotheses. The advantage of this approach is that the most sophisticated, computationally expensive translation models can be applied to score entire French sentence hypotheses in a small amount of time (for any reasonable value of N). The disadvantage is that these translation models are brought to bear only after the recognizer has made irreversible decisions about word choice.

In version 3, the order is reversed: the same translation model as in version 2 is applied before recognition begins on the French sentence, generating a *dynamic vocabulary* from the English sentence that is smaller than the total lexicon of the recognizer. Recognition proceeds normally (using only the bigram French language model) except that the recognizer is only allowed to consider words in the dynamic vocabulary. Before implementing version 3, we knew that this approach would at least have the advantage of speeding up recognition, by reducing the search space. We could not know in advance whether