ABOUT IMPROVING RECOGNITION OF SPONTANEOUSLY UTTERED FRENCH CITY-NAMES

D. Jouvet, K. Bartkova, L. Delphin-Poulat, A. Ferrieux, X. Lamming, J. Monné, C. Raix

France Télécom R&D – DIH/IPS
2 avenue Pierre Marzin, 22307 Lannion, France

ABSTRACT

This paper deals with the recognition of French city-names over the telephone. This recognition task, critical in many applications, involves a 40,000 city-name vocabulary, ranging from short monosyllabic words to long official compound-names. Data collected from a field experiment are analyzed, and several ways of improving speech recognition performance are investigated. This includes a careful checking of the pronunciation lexicon, acceptation of shorter forms (common names), adaptation of the acoustic models and introduction of specific noise models as well as a few frequent words and expressions to facilitate out-of-vocabulary data rejection. Experiments show that all these techniques help improving the overall recognition performances and nicely combine together.

1. INTRODUCTION

Recognizing city-names is mandatory in many applications such as directory assistance, tourism information, etc. However this task is quite difficult in France as it implies a large vocabulary (40,000 city-names). Furthermore, some names are short monosyllabic words, while other ones, such as long official compound-names, are frequently abbreviated in shorter common names. Spontaneous speech data collected from a field experiment was used to investigate this difficult speech recognition task.

After a brief overview of the speech recognition system used, the speech data is analyzed in section 2. As many city-names are similar, some a priori information can be efficiently used to re-order the N-best candidates in order to favor large and frequently requested cities. Section 3 investigates the pronunciation lexicon. Two aspects were considered: first, the pronunciations of city-names were checked and some pronunciations variants were introduced in the lexicon; then, short common city-names were added to the official city-name lexicon. Section 4 recalls on the improvement resulting from the adaptation of the acoustic model parameters. Finally section 5 deals with spontaneous speech artifacts and shows that adding noise models and frequently used words and expressions into the recognition model improves recognition performance.

This work was partially supported by the SMADA European project. The SMADA project is partially funded by the European Commission, under the Action Line Human Language Technology in the 5th Framework IST Programme.

2. EXPERIMENTAL SET-UP

For investigating the city-name recognition task, and evaluating speech recognition performance, we used spontaneous speech data collected from a field experiment.

The speech recognizer is HMM based. The acoustic analysis computes energy and 9 Mel-frequency cepstral coefficients every 16 ms. First and second order temporal derivatives, estimated on a 5-frame window, are introduced in the modeling. Context-dependent phoneme models [1] are used, and the acoustic modeling relies on mixtures of Gaussian densities that were estimated on a large task-independent speech corpus.

2.1. Database

The database was collected in a field experiment where people from different regions of France were calling a green number providing access to a directory assistance demonstrator. Nation-wide information on residential subscribers was available, that is about 23 million subscriber entries. The city-name database used in this paper corresponds to answers to the question asking for the city-name. The evaluations reported here are conducted on 4,000 thousands utterances (test set) that were hand transcribed.

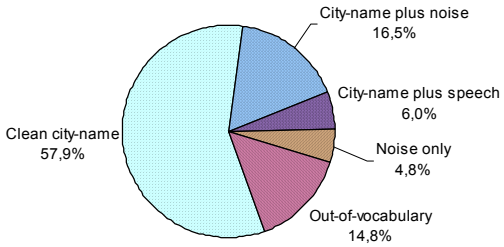

Figure 1: Analysis of field data.

Figure 1 represents the distribution of the data. 58% of the utterances are "clean" utterances, i.e. correspond to the pronunciation of a city-name and do no contain extra speech or significant noise. 17% contain a city-name plus noise loud enough to be annotated during the hand transcription. 6% contain extra speech signal, besides the city-name (before and/or after). 5% of the tokens contain only noise signals loud enough to trigger the energy-based endpoint detector. Finally 15% of the tokens correspond to out-of-vocabulary data. The speech recognizer has to reject these noise and out-of-vocabulary tokens in order to make the dialog more fluent.

2.2. A priori knowledge: city size

As mentioned before, the vocabulary size is large: about 40,000 city-names. As some city-names are rather similar, the raw error rate is quite large. However, several ways of improving the service quality can be envisaged. One possibility is to go through the list of N-best recognized hypotheses, up to an accepted answer or to a pre-defined limit. Another possibility is to ask for extra information, such as the department. Of course these dialog mechanisms may be dependent on a confidence score.

Because of such possible processing, the N-best candidates are often considered in this paper for analyzing the speech recognition performance.

In order to increase the probability of occurrence of the correct answer at the beginning of the N-best list, some a priori knowledge can be used. Here, the city-size (number of inhabitants) is considered. Such information is handled as a unigram language model, the a priori probability associated to each city-name being proportional to its population. This is very efficient for increasing the frequency of the correct answer in the 2 or 3 best candidates (this can be observed in Figures 3, 4 and 8, by comparing curves "with prior" to curves "no prior"). The figures also show a very large improvement for the first best hypothesis.

3. PRONUNCIATION VARIANTS

City-name pronunciations cannot be always predicted from the spelling. Hence the pronunciation of city-names have been checked and corrected if necessary. Moreover, as people do not always use the official name, especially when it is a long one, it is important to add common names into the lexicon.

3.1. Analysis of city-names

Among the 40,830 city-names present in out reference city-name vocabulary, only 58% are made of a single word. The remaining ones are compound names containing from 2 to 8 words. The frequency of the city-names according to their length (number of words) is represented in Figure 2.

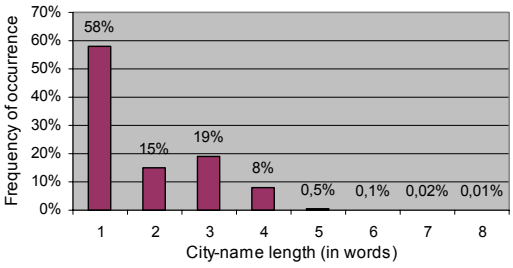

Figure 2: City-names frequency according to their length.

People tend to simplify compound names in particular when they live in the city, or when they are familiar with it, or simply because they find the name too long (more then 4 words). In such cases, common shorter names are often used instead. It also happens that some people are not even aware of the full non-abbreviated official city-name. Therefore, when a simplification is possible, it is crucial to add the abbreviated common forms of the city-names into the recognition vocabulary.

3.2. Introducing common abbreviated city-names

In order to create "abbreviated" names an analysis of the city-names was carried out. It was found, for instance that, when city-names are composed of 2 words (15% of the names), in 53% of the cases the first word turned out to be inseparable from the second word as it is either a definite article ("*le*", "*la*", "*les*") or an adjective like "*saint*", "*grand*" (large), "*petit*" (small), etc. In 3-word city-names (19% of the names), for 60 % of them the second word is a preposition, mainly "*sur*" (on). In these cases, and when the first word is not a very common name (such as "*château*" (castle), "*forêt*" (forest), etc.) a cut before the preposition is considered as possible. In 4- and 5-word city-names (8 % of the names), in about 70 % of the cases, the first word is the adjective "*saint*". The second word is the name of the saint and the remaining words give more precision about the location. In these cases a cut after the name of the saint is highly probable.

Table I: Examples of common abbreviated names.

Complete city-name	Short common forms
• *Montrieux sur le Loir*	• *Montrieux*
• *Les Monts Verts*	• <no short version>
• *Port Saint Louis du Rhône*	• *Port Saint Louis*
• *Saint Sulpice de Cognac*	• *Saint Sulpice*
• *Saint Etienne du Gué de l'Isle*	• *Saint Etienne* • *Saint Etienne du Gué*

A set of rules is applied for creating the short common names. For example, no cut is allowed after an adjective when preceding the noun, after a number, an article or a very common word in first position. This way, for the city-name *"Montrieux sur le Loir"*, two names are accepted after processing: a long one, which is its full name and a short one (see Table I). In this case the cut is carried out before the preposition *"sur"* (on). On the other hand, the city-name *"Les Monts Verts"* though 3-word long, has only one (long) form, because the first word is an article and the second word *"monts"* (hill) is a very common noun, which cannot unambiguously specify a city. No cut is allowed after the first word when it is a very common noun such as *"port"* (port) in *"Port Saint Louis du Rhône"* and the cut for the short version is located only after the name of the saint. A city-name containing the name of a saint with a simple complement such as *"Saint Sulpice de Cognac"* has two forms, while with a more complex complement, such as *"Saint Etienne du Gué de l'Isle",* 3 forms are possible.

The reference city-name database used contains 40,830 city-names, of which only 37,848 are actually different. Cities with same names can be found in different areas of France. When "shorter" city-name forms are introduced the number of different city-names is increased by about 4,000 names. However, short forms close to the place of caller can be considered as more probable than further remote ones. Nevertheless such a priori knowledge is not used in the experiments described in this paper.

3.3. Pronunciation variants

Sometimes city-names can have more than one pronunciation. When a city is situated in a part of the country having a specific language such as Breton, Basque or Alsatian languages, city-

names can have a different phonetic form when uttered by natives of the region and when uttered by people from other parts of the country. Differences can exist even for the same city-name designating cities in different parts of the country. In order to recognize a correct (local) native pronunciation and also a less native and more French rule-based one, pronunciation variants are accepted for approximately 10 % of the French city-names. For example, the city *"Trébeurden"* (in Brittany) can be uttered with a native pronunciation [trebœrdɛ̃] or with a more French one [trebœrdɑ̃]. For the Alsatian city *"Artolsheim"* the correct pronunciation is [artɔlsaɪm] but [artɔlsɛm] is also accepted. Many other pronunciation variants, which are not counted in the 10% mentioned above, are also taken into account, such as the variants associated to the mute schwa (ə, which can be pronounced or not) and the optional silence which can occur between words in compound names.

3.4. Recognition results

Improvement due to the introduction of common abbreviated city-names, as well as the one due to the checking and correction of pronunciation variants of the city-names, was evaluated on field data. 72 utterances (that is 2.2% of the data) can be handled correctly only when abbreviated names are introduced in the lexicon. Figure 3 displays the cumulative percentage of correct answers (city-names) in the N-best hypotheses. Results are reported with speech recognition only (no prior) and after re-ordering the N-best lists using city-size a priori knowledge (with prior). To make the figure more readable, these last curves are limited to the 5-best hypotheses.

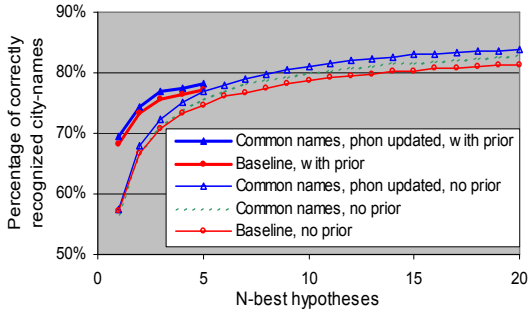

Figure 3: Pronunciation variants & common short names.

Figure 3 shows the improvement due to both the introduction of common abbreviated names and lexicon pronunciation checking (curves "Common names, phon updated" vs "Baseline"). Without prior, a 2.3 % absolute increase of correct answers is observed in the 5- to 20- best hypotheses. The middle dotted-line curve shows that more than half of the improvement is due to the introduction of common names. Taking into account city-size knowledge (curves "with prior") leads to a 1.3% absolute increase of correct answers in the 3 best candidates.

4. ACOUSTIC MODELING ADAPTATION

It is well known that adapting acoustic parameters to the task context improves recognition performances. Several adaptation techniques (MLLR, Bayesian and incremental adaptation) were previously evaluated on a medium size vocabulary voice portal task, and it was shown that all techniques lead to significant recognition performance improvement [2]. Here, incremental adaptation was applied on the city-name model, in a supervised mode. Only clean utterances from the training set were used to perform adaptation. The model was first evaluated on a purely acoustic basis (i.e. no a priori language model was used). It can be seen in table II that for the rejection threshold A, substitution, false rejection and false alarm rates were decreased. Of course, the operating point can be modified by adjusting the rejection threshold (see for example Figure 7). For a similar false rejection rate as the initial model (rejection threshold B) the substitution and false alarm rates are reduced more.

Table II: Recognition performances

	Substitution	False reject	False alarm
No adapt.	34.08%	8.45%	67.42%
Adapted, Thresh. A	33.07%	5.86%	63.77%
Adapted ,Thresh. B	31.31%	8.51%	52.58%

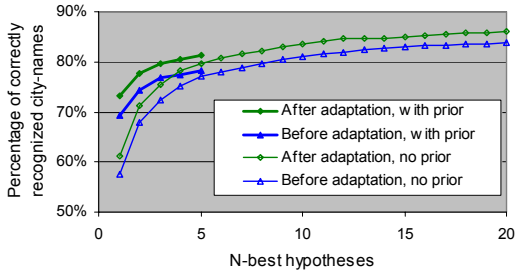

Figure 4: Acoustic models task adaptation.

Figure 4 reports the improvement due to the task adaptation of the acoustic models. The curves marked "After adaptation" correspond to the N-best hypotheses obtained with the model "Adapted, Thresh. A" in Table II. These curves show a large improvement in the N-best solutions, even after re-ordering using the city-size information.

5. SPONTANEOUS SPEECH

In real applications, the recognizer has to deal with many artifacts. The speech signal may be surrounded by noisy parts (which may even span over the speech part), or the city-name can be preceded by hesitations and unpredicted words. Hence the recognizer must include a rejection mechanism in order to deal with out-of-vocabulary data. One classical way for rejecting such data is to include a filler model in parallel of the vocabulary words. The rejection / keep decision thus amounts to comparing a log-likelihood ratio to a pre-defined threshold, which is linked to the weight of the filler model. Here we investigate some ways of making the filler model more detailed and thus more efficient.

5.1. Adding specific noise models and frequent words

Besides the generic rough noise model, which is part of the basic rejection model, specific noise models were investigated. A few models were first trained from another annotated speech corpus. These specific noise models, corresponding to expiration, inspiration, handset, cough, bips, etc, were added into the recognition grammar. They could occur before and after

the vocabulary city-names (to capture surrounding noises), as well as in parallel to the city-names (to facilitate rejection of noise tokens). This is indicated by "& Noise" in the following figures.

Frequent out-of-vocabulary words were analyzed on another set of field data. Hesitations ("*euh*", "*ben*") and a few words ("*à*" (in), *"alors"* (then), *"c'est"* (it is), etc) were observed as being frequently used before the city-names. Allowing these words before the vocabulary is indicated by "& Spot" in the following figures (they help spotting the city-name in the user answer).

Finally frequently used words occurring in the answers, instead of the expected city-name, were also added in parallel to the city-names. This includes, for example, words or expressions such as *"oui"* (yes), *"non"* (no), *"pardon"* (sorry), *"je ne sais pas"* (I don't know), etc. They are indicated by "& OOV", as they should facilitate rejection of out-of-vocabulary data.

5.2. Recognition results

The following figures report error rates on various data subsets. As there is always a trade-off between false rejection and false alarm (and substitution) the operating curve of the adapted model is represented in each graph. This curve is obtained by varying the rejection threshold (filler model probability).

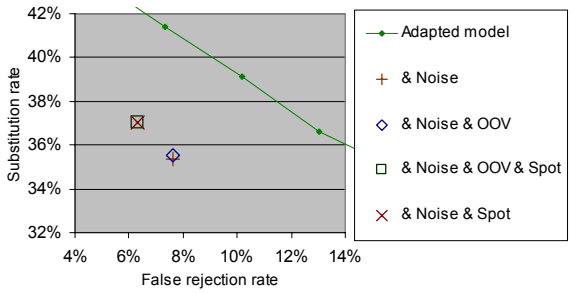

Figure 5: Error rates on "city-names plus noise" data.

On data surrounded by noise, a significant improvement is observed in all cases (Figure 5). The detailed noise models facilitate a better alignment of the relevant speech signal with the vocabulary models.

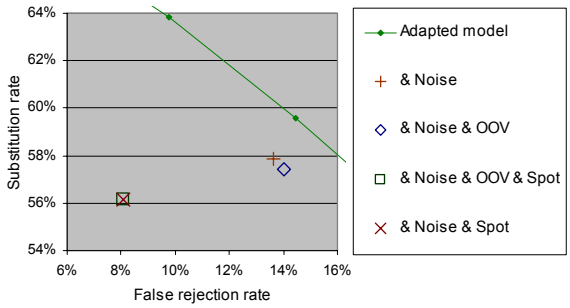

Figure 6: Error rates on "City-names plus speech" data.

On city-names surrounded by extra speech, it clearly appears (Figure 6) that introducing and allowing frequently used words before the vocabulary words improves performances.

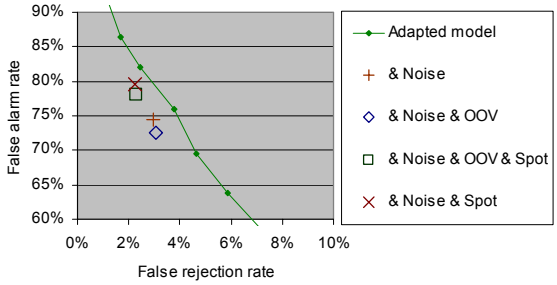

Figure 7: Error rates (False alarm vs False rejection).

Figure 7 shows that all configurations improve performances by providing lower false alarm and false rejection rates. However, it is the addition of the noise models alone, or along with out-of-vocabulary words, that provide the best configuration on this field data. The proportion of city-names surrounded by extra speech is not large enough to emphasize a global improvement by adding the "& Spotting" configuration.

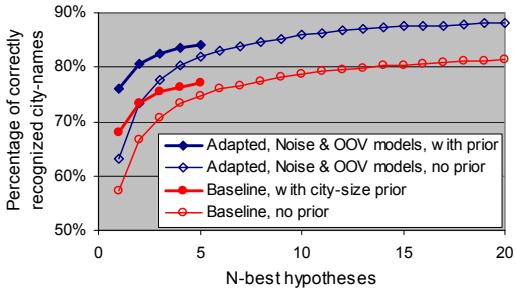

Figure 8: Overall improvement of correct recognition rate.

6. CONCLUSIONS

This paper shows that for a difficult task like the recognition of the French city-names, many ways for improving the overall recognition performance can be investigated. For example, city-size provides useful a priori information, which can be efficiently handled in an application. Extra information such as the department name may also be requested. The department name or number is sometimes spontaneously provided by the user, along with the city-name, however such information is not yet handled by the system. This is part of further planned improvements.

Having a correct pronunciation lexicon and allowing common abbreviated forms is of paramount importance. Here again, task adaptation proves to be efficient. A detailed modeling of noise artifacts, as well as the introduction of frequently used words and expressions is a promising way of improvements, which have to be investigated deeper, together with the adaptation of noise models to field data. As shown in Figure 8, all the techniques nicely combine together.

7. REFERENCES

[1] Jouvet D, Bartkova K. & Stouff A., "Structure of allophonic models and reliable estimation of the contextual parameters", *Proc. ICSLP'94*, Yokohama, Japan, 18-22 Sept. 1994.

[2] Delphin-Poulat L., "Comparison of Techniques for Environment / Application Adaptation", *ITRW Workshop on Adaptation Methods for Automatic Speech Recognition*, Sophia-Antipolis, France, 29-30 Aug. 2001, pp.139-142.