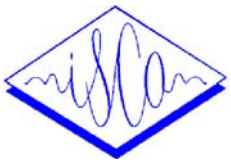
NAVIGATING GERMAN CITIES BY SPONTANEOUS FRENCH QUERIES



Harouna Kabré and Alexander Waibel

Interactive Systems Laboratories, ILKD
Department of Computer Science
University of Karlsruhe
76128 Karlsruhe, Germany
{kabre,waibel}@ira.uka.de

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ABSTRACT

This paper reports our efforts on the adaptation of a baseline system trained on clean speech to a task for which French native speakers uttered some Spontaneous French queries while driving a car. When the system is retrained on the new task acoustic data the Word Error Rate (WER) is decreased by 60% compared to our baseline system initial performance on the new task. We show that on spontaneous queries, 1/4 of this improvement could be achieved without prior system retraining by a more accurate Language Modelling which takes into account the noises and spontaneous speech effects and by a carefull grapheme/phoneme transcription of foreign words. We also describe the integration of this French system in our Multilingual Navigation System.

1. INTRODUCTION

Speech technology is more and more spreading to new applications over the last years. The navigation of foreign countries while driving a car is one of the new application that will be of high interest in the future due the growing demand in tourisms, in business's trips, etc.. Whereas applications in this domain so far mostly have been restricted to hand free operation of the telephone, the demand for other functionality such as requesting information for a street location in a foreign country or controlling radio and cassete in a car with navigation systems is steadily growing. Unfortunately recognizing navigation queries is a challenging task for at least three reasons [4]:

- the noisy car environment which leads to performance decreases compared to a quite environment.
- the large number of confusable street and city names that have to be recognized by the system, where a high confusability will also lead to performance degradations.
- the lack of knowledge on the foreigner competence for pronouncing foreign words (E.g. French pronouncing german street and city names).

It results a mismatch between the training and testing conditions for a speech recognition system trained in a lab environment which in turn leads to a dramatic decrease of its performance. The question is wether or not an adapation is possible without a prior acoustic recording of speeech data in the target acoustic environment.

This paper reports our efforts on the porting of a baseline system trained on clean read speech to the recognition of in-car spontaneous speech data collected from French native speakers pronouncing german city names. Starting with a French recognizer designed in the Janus III environment (see section 5), we show that on spontaneous queries, 1/4 of this improvement could be achieved without prior acoustic data recording by an accurate Language Modelling (LM) which takes into account the noises and spontaneous effects and by a carefull grapheme/phoneme transcription of foreign words. Moreover the integration of our French system into the Multilingual VODIS project is described.

An overview of our methodology is firstly introduced (section 2) followed by a description of our test database and Language Modelling (section 3 and 4). We discuss our experiments (section 5 and 6) before concluding the paper in section 7.

2. METHODOLOGY

2.1. Dealing with the Problem of Noises

In our task of in-car speech recognition, given the noisy acoustical environment inherent to the driving scenario and the noise-sensitiveness of phonetically-based recognizers, different noises are distinguished. There are static background noises and non-human noise events due the car environment and human noises due to spontaneous speech. In a first experiment we ignored all the noises in the transcriptions since our baseline system has been trained on clean speech. In the second experiment we included a transcription of some human noises and included their phonetic strings in the system dictionnary.

2.2. Deriving German City and Street Names Pronunciation by French Speakers

The problem of native speakers pronouncing foreign words is a complex problem due to the lack of knownledge about this phenomena. It is obvious that it depends on the competence of each native on the foreign language. Unfortunately this effect resulting from cross-language approximation is unknown and the solution used in our task is empirical.

For our task, we analysed a subset of French speakers pronouncing some German street and city names. Out of 10 speakers we distinguished a group of novices and a group of those having some knowledge of german language. From the two groups we oberved that the speakers (in particular the novices) tend to use some kind of visual rules to utter foreign words. For example as can be seen in table 1, novices tend to pronounce Acherstrase negleting the difference between /a$\tilde{}$/ and /a/ whereas the other group has almost reach the correct pronunciation which is Acherstracheu. Even in this case the french accent is noticable.

The general approach we adopted is to start with the orthographic input of german street and city names. Then a simple model which deletes and inserts some phonemes is successively applied to it. It results in 2-3 versions of each word for which we derived a phonetic transcription for the recognizer dictionnary.

3. DATABASES

For training the baseline system we used the BREF_1 database which averages 12 h of read speech of native French speakers. There are 80 speakers who read a part of the French newspaper "lemonde". Since the vocabulary is large enough (20000 words), adding the missing street and city in the system dictionnary allows it to handle to a certain performance some spontaneous queries without retraining.

For testing we used the French VODIS database [1], which has been collected at INRIA in France. It consists of 200 speakers (102 males/98 females) of French car navigation speech. The data is sampled at 11.025 kHz with a resolution of 16 bits in three different car environments using four different microphones. For our experimentations only the close-talk microphone is used which averages 28000 sentences. From these sentences 3200 are kept for testing. Among the 3200 sentences 1513 are real spontaneous French queries and the remaining are some spontaneous responses to prompt messages displayed by the navigation system.

4. LANGUAGE MODELLING

The test of the new task text data against the baseline system vocabulary before the different interpolation of LM showed a 41.5% OOV (Out Of Vocabulary) rate which indicates that our new task is different. In an effort to decrease this OOV rate, we firstly performed an interpolation between the training transcriptions and our task text data. Then we secondly included the human noises as discussed in previous subsection. Note that since some human noises sounds are similar to some phonetic sounds (E.g. /eh/ is transcribed to /e e h/), the addition of noises reduces the OOV rate. This is useful in our case

	Input	Pronunciation
novice	Acherstra se	a S E R s t R a z
expert	Acherstra se	a S E R s t R a S 2
novice	Heidelberg	2 d 2 l b E R
expert	Heidelberg	2 d 2 l b E R g

Table 1: Some transformations realised by some noviced and experts of German Language on some city and street names.

because the test database contain such events which are considered as words.

5. EXPERIMENTAL SETUP

5.1. The Baseline System

The speech-to-speech translation system JANUS-III is a joint effort of the Interactive Systems Labs at Carnegie Mellon University, Pittsburgh, and the University of Karlsruhe, Germany. For a detailed description, refer to [5].

The janus core engine is language independent and it has already been shown that the underlying recognition methods and techniques can be applied to several languages [Multi]. The French baseline system has been designed and trained on the 76 speakers of the BREF_1 database [6] and tested on the 6 remaining speakers. The system is a standard Hidden Markov Model (HMM) based system with 44 phonemes and one silence modelled by a three states left to right HMM, except for the silence which has only one state. The audio data is sampled at 16 khz with a resolution of 16 bits. The MFCC coeffcients are extracted every 10 ms over a window of 20 ms. The first and second derivatives of MFCC are computed and we formed a final acoustic vector which includes zero crossing and the energy values. The 43-size vector is reduced with an LDA (Linear Discriminant Analysis) to 24-sized vector from which a standard JANUS training sequence is run. The final system achieved a word error rate of 12.4% [7].

5.2. Results

We carried out two experiments, both on language modelling and system retraining. The baseline system achieved 35.4% WER (see Table 2) when testing on the new task acoustic data. The use of our best LM interpolation improved in order of 12%. When the noises are transcribed before the LM estimation the error reduction is up to 15% WER relative. If we compared to the solution of retraining the system (i.e. 60% WER) we observe that both grapheme/phoneme transcription and LM brings 1/4 of the improvement.

6. NAVIGATING GERMAN CITIES

The final system obtained is integrated in a Multilingual Navigation System [1]. As can seen in Figure 1 the navigation system consists of five modules. The system that is able to handle French navigation queries. For example when the user utters the following request: " Ou se trouve la rue Kaiserstrase ?", the hypothesized output of the recognizer is then fed into a semantic case-frame parser. The output of the parser is piped into a general manager. Within this general manager a dialogue subsystem decides if the parsed output is specific enough to be given as input to the navigation. If the parsed sentence does not include a specific destination, the general manager initiates a clarification dialogue with the user to further narrow the actual destination. Some examples of clarification dialogues are discussed in details elsewhere [4]. As soon as the destination of the query is fully specified, the general manager retrieves the necessary map coordinates from the map database and passes this information to the navigation system. So the route is calculated and directions are synthetized to the user.

Systems	WER	Relative Error Reduction
		(compared to baseline system)
S1: BREF LM	35.4%	-
S2: interpolated LM	31.3%	12%
S3: S2 + human noises	30.1%	15%
S4: S3 + acoustic retraining	14%	60%

Table 2: Performance of the French Recognizer tested with different Language Models.

7. SUMMARY

We have shown that compared to a solution for which the baseline system is retrained on car speech, a carefull Language Modelling of the task by taking into account some noises inherent to the spontaneous speech and by using some empirical rules of native French pronouncing german street and city names 1/4 of improvement is achieved which allows the setup of a car navigation system by spontaneous French queries.

Both the integration and the evaluation of the system are on the way. In the future this part of the task will be persued while investigating more deeply the problem of foreign word pronunciation which seems to be very sensitive for our system performance.

8. ACKNOWLEDGEMENT

We acknowledge the contribution of Martin West phal to this paper for his great help.

9. REFERENCES

[1] VODIS. *http://www@werner.ira.uka.de/ ISL.speech.vodis.html*, 1999.

[2] Lvcsr. M. Finke, T. Zeppenfeld, M. Maier, L. Mayfield, K. Ries, P. Zhan, J. Lafferty, A. Waibel: Switchboard April 1996 Evaluation Report, DARPA, 1996.

[3] Multi *Language Independent and Language Adaptive Large Vocabulary Speech Recognition*, T. Schultz and A. Waibel, ICSLP 98, Vol. 5, pp. 1819-1822, Sidney 1998.

[4] P. Geutner, M. Denecke, U. Meier, M. Westphal and A. Waibel. *Conversationnal Speech Systems for On-board Car Navigation and Assistance*, ICSLP 98, vol. 4, pp. 14447-1450, sidney 1998.

[5] Janus. M. Woszczyna, M.Finke, D.Gates, M.Gavalda, T.Kemp, A.Lavie, A.McNair, L.Mayfield, M.Maier,I.Rogina, K.Shima, T.Sloboda, A.Waibel, P.Zhan, T.Zeppenfeld: *Janus II advances in spontaneous speech translation*, in Pro. ICASSP-96, pp 409 ff, Atlanta, ISBN 0-7803-3192-3, 1996.

[6] Elra. *http://www.icp.inpg.fr/ELRA/fr/ inforequest.html*, 1999.

[7] Kabré, *Large Vocabulary French Recognition in the JANUS III environment*, ISL report, Dec. 1998.

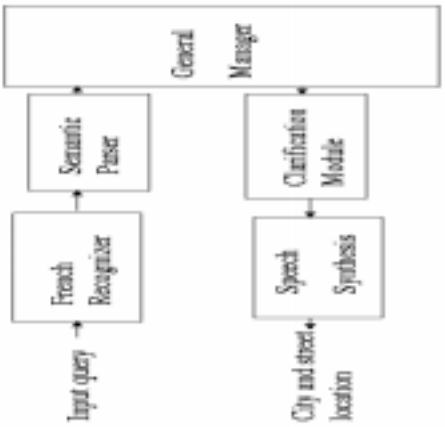


Figure 1: Overview of the General Manager Architecture