IDIAP COMMUNICATION

IDIAP

Martigny - Valais - Suisse

SWISSCOM "AVIS" PROJECT (No. 392)

Advanced Vocal Interfaces Services

Technical Report for 1997

J. M. Andersen G. Caloz H. Bourlard

IDIAP–Com 97-06

December 1997

Dalle Molle Institute for Perceptive Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] Dalle Molle Institute for Perceptive Artificial Intelligence PO Box 592 CH-1920 Martigny, Switzerland

SWISSCOM "AVIS" PROJECT (No. 392)

Advanced Vocal Interfaces Services
Technical Report for 1997

INTRODUCTION

This report covers the period of January 1-December 31, 1997 and describes the research and development activities performed at IDIAP for Swisscom in the framework of the "AVIS" F&E-Project No 392 on "Advanced Vocal Interfaces Services".

As initially planned in the contract, the IDIAP R&D activities actually focused on:

1. Task 1: Continuous speech recognition over the telephone line, in view of the development of better and automatic telecom services (e.g., in the framework of "Blue Window" or in view of partial automation of 111 service).

   As summarized in the following, this task involved:

   (a) Task 1.1: Initiating of a sub-project on linguistic modeling of spontaneous Swiss French requests.

   (b) Task 1.2: Setting up a first reference system for Swiss-French continuous speech recognition, with initial tests on the (Swiss-French) Polyphone database (see below).

2. Task 2: Management and distribution of the Swiss-Polyphone database, with extension to and preliminary testing on GSM data:

   (a) Task 2.1: Distribution and management of the Swiss-Polyphone data.

   (b) Task 2.2: Specifications and collection of GSM data, with initial recognition tests.

1 AVIS PROJECT

1.1 Task 1.1: Initiation of the Swiss-French Project

The goal of this sub-project is to set up the required basis in language and dialog modeling for interactive voice response systems by selecting partners with the appropriate expertise. These partners should complement IDIAP's expertise in speech recognition and should collaborate with them to develop complete systems.

After discussions with potential partners, it was decided to ask ISSCO (Geneva) and LIA/EPFL (AI Lab of EPFL) to submit a join project proposal. This **T0+6 deliverable** is included to this report as **Annex A** and has been submitted to Swisscom for comments and approval.

In the meantime, an active collaboration between LIE/EPFL and IDIAP was initiated. The following activity report is part of **T0+12 deliverable** and summarizes some of the actions that have already taken place in 1997:

1. A. Rozenknop, student at ENST-Paris, came for a 3 month internship at LIA/EPFL. The objective of his stay was to install a running version of the speech recognition STRUT (Speech Training and Recognition Unified Tool – see Task 1.2) environment. The environment, as well as some language models, were provided to LIA/EPFL for installation and testing. A. Rozenknop also benefited of a 2 weeks training at IDIAP (in July 1997), during which he got the necessary information to get started with STRUT.

2. R. Aragues, exchange student at EPFL, is currently doing his diploma project at LIA/EPFL. The objective of his work is to extend the STRUT installation realized by A. Rozenknop, and especially to test the environment on more realistic examples such as sentences taken out of the Polyphone base. Once again, the examples and languages models will be provided by IDIAP. The overall objective of this first collaboration is to build a small corpus of word graphs produced by the STRUT recognizer for typical sentences taken from the Polyphone base. The small corpus will be very useful for better tuning of the research on coupling between speech recognition and language processing that will be jointly undertaken by LIA/EPFL and IDIAP in various frameworks (AVIS, FNRS, ...). In particular, for LIA/EPFL, it will serve as a basis to better evaluate the complexity of the adaptation of a stochastic syntactic analyzer to an input in the form of a word graph produced by a speech recognizer. For IDIAP, it should also be the opportunity to carry out some research work dedicated to the enhancement of the algorithms dedicated to word graph production (extension of the existing Noway module).

1.2 Task 1.2: Baseline Continuous Speech Recognition System

1.2.1 Goals

The goal of this task was to produce a baseline system for continuous speech recognition in French. This system had to be tested on French Swiss Polyphone data (telephone speech) described in Task 2.1. Finally, it was expected that preliminary results would be reported on a specific subset of Polyphone including 500 simulated operator (111) calls.

1.2.2 Overview of results

To achieve the above goals, a wide range of software has been installed, adapted and tested, followed by a series of experiments.

The installed software includes STRUT (Speech Training and Recognition Unified Tool) [4], NOWAY (a large vocabulary continuous speech recognition system developed in the framework of THISL, a European project in which IDIAP is an active partner)), and BDLex (a French dictionary for speech recognition) as the major components.

The initial experiments focused on Polyphone isolated speech recognition, where a recognition rate of 94.8 % was reached with a vocabulary of 39 application words. Recently the recognizer has been tested on the much more difficult operator calls where an error rate of 60.7% was observed on an independent test set and 59.6% on the training set. In the report below, it is discussed why the recognition is so low. Probably this is mainly due to a poor language model as well as a non-optimal dictionary especially for proper names.

To allow tests on the operator calls, 3500 proper names were transcribed by means of a rule-based grapheme-to-phoneme converter followed by manual corrections.

1.2.3 Software

This year has been largely dominated by the installation and debugging of the software needed for state-of-the-art (continuous) speech recognition. Given in-house expertise and the current projects at IDIAP, we also wanted to be able to use hybrid systems, using artificial neural networks (ANN) together with hidden Markov models (HMM) [1] for speech recognition (see later).

This section briefly describes the different software components that have been installed and used in our experiments. To develop a continuous speech recognition system, we need:

1. An acoustic (training and test) database: in our case, the Polyphone database has been used.

2. Training of the acoustic models: we used the STRUT software (briefly described below), as well as HTK (in the case of preliminary tests on GSM speech).

3. A dictionary containing the list of words to be recognized, together with their phonetic transcription. While this is easy to get for well defined (medium size vocabulary) tasks, it can be a serious limitation in the case of open tasks like 111 calls containing proper names and other not predefined words. In the work reported here we mainly used BDLex, occasionally extended manually.
4. A grammar model: in all modern speech recognition systems, this grammar is represented in terms of bigrams and trigrams that are estimated from very large text corpora, ideally representative of the targeted task. The software tool used here to estimate those bigrams and trigrams was the CMU Toolkit. Since no specific text corpora were available, we mainly used the word transcriptions of the Polyphone acoustic data to train the language model (although this was known to be largely undersampled).
5. A recognition system: as briefly described below, we used NOWAY, an efficient large vocabulary continuous speech decoder developed by Steve Renals at Sheffield University in the framework of the THISL project.

STRUT: Speech Training and Recognition Unified Tool

The majority of the work has been based on the STRUT (Speech Training and Recognition Unified Tool) software, initially developed by FPMS. (Faculté Polytechnique de Mons, Belgium). After some preliminary installation and testing of STRUT at the end of 1996, the first part of 1997 has been used on learning how to use STRUT as well as installing and debugging new versions of this software. STRUT was chosen because of its flexibility and because it will enhance the collaboration with FPMS. Another important factor is that STRUT enables us to use hybrid systems for speech recognition in which IDIAP has a very good expertise (and active collaboration with ICSI, Berkeley).

BDLex

Besides setting up STRUT, IDIAP also installed and tested a new version of the BDLex dictionary. BDLex is a French dictionary specifically designed for speech recognition, and phonetically transcribes 50,000 words and their inflected forms. The use of this dictionary required the adaptation of the software modules accessing that database. Additional preliminary software has also been developed to deal with liaisons and other phonological phenomena in French. Of course, this is just a preliminary version and it is expected that more research will be done in the future in the field of phonological variations and multiple pronunciations.

CMU Toolkit

The CMU (Carnegie Mellon University) Toolkit is a set of utilities to create bigram and trigram language models from large text corpora. This software was installed as a step towards continuous speech recognition.

We later describe the text used for training these language models. The Toolkit can be obtained for free by *ftp* to *ftp://ftp.cs.cmu.edu/project/fgdata/CMU_SLM_Toolkit_V1.0_release.tar.Z*

NOWAY Decoder

The NOWAY decoder (developed by Steve Renals at Sheffield University in the framework of the THISL EC project in which IDIAP is an active partner) was compiled and installed at IDIAP. NOWAY is a large vocabulary continuous speech decoder which takes the output from an ANN and returns a best string of words or a lattice of words.

The major advantage of NOWAY is that it is written to work well with ANNs and uses a very flexible scheme of pruning (taking full advantage of artificial neural networks). This allows to make an appropriate compromise between speed and accuracy which is important for real-time applications.

1.2.4 Algorithms and approaches

Hybrid Speech Recognition System

All the experiments reported in this section were based on hybrid speech recognition. Such a recognition system is composed of an artificial neural network (ANN) and a hidden Markov model (HMM). The ANN takes preprocessed acoustic data (features) as input and estimates posterior probabilities for different classes (e.g. phonemes) as output. In our case, a particular form of ANN, referred to as multilayer perceptron (MLP) was used to represent acoustic models and compute emission probabilities of standard HMM systems as used by the vast majority in the speech community.

Hybrid systems is a relatively new trend in speech recognition. However, they have been shown (in the framework of several international evaluation tests) to yield competitive performance with additional advantages in terms of CPU and memory requirements. There are also many particularly promising extensions that are currently investigated at IDIAP (as well as in other laboratories).

Feature Extraction

The preprocessing of the speech signal consists of a RASTA-PLP feature calculation[2, 3]. RASTA-PLP features were designed to be particularly robust to noise and different transmission channels, and as such is well suited for telephone speech. 12 RASTA-PLP coefficients along with their first derivatives (delta's) as well as delta and delta-delta values of the log-energy, which makes a total of 26 features. These features were calculated every 10 ms (a frame) using a window of 30 ms (thus an overlap of 20 ms.) As is usually done with MLPs in hybrid systems, we have used a context of 9 consecutive feature vectors (4 vectors before and after the current frame), giving a total of 234 input to the MLP.

Artificial Neural Networks

The Artificial Neural Network (ANN) used in our experiments is of the type Multi-Layer Perceptron (MLP.) MLPs have been shown to be able to approximate any function, and in particular they can learn a posteriori probabilities.

The output of the MLP (and input to the decoder) in theory can be any phone like units. However, phonemes are most often used, and in our experiments we have used the phoneme set defined by the SAMPA standard which is used by BDLex - a total of 36 MLP outputs.

In order to train an MLP, it is necessary to have a segmentation for each utterance, that is for each frame one of the 36 phonemes must be assigned as the one being present. Hand-segmentation is a very time consuming task, but never-the-less, we have a considerable amount of hand-segmented speech. At the time when the database was recorded at IDIAP, almost 5 hours of speech had been hand-segmented. This part served to train a so-called boot-net.

For the training data used in the experiments reported here, only an orthographic transcription was available. To obtain a segmentation, it is necessary to first convert the orthographic text into a sequence of phonemes. Then this phoneme sequence is matched to the speech signal, using a dynamic programming method often called forced Viterbi. The result is that each frame is assigned one of the phonemes in the phonetic sequence. From this point, the MLP training can be performed by the well-known error back-propagation method.

1.2.5 Experiments on Isolated Words

For initial testing of the speech recognition system, an isolated word task was chosen.

The Swiss French Polyphone database that has been used in this work, was split into different training set, development set, and test set. The same split of the data is now also used at FPMS (Faculté Polytechnique de Mons), which was provided the same data for comparison purposes (in return to access to their results and some of their software).

Training Set

A series of experiments was performed on the Swiss French Polyphone. A subset of 400 speakers were used (200 male and 200 female speakers) out of the full training set of 2000 speakers. The database contains about 10 phonetically rich sentences for each speaker. However during the experiments different kinds of irregularities as e.g. noise on the recording or strange utterances, were discovered, and so the number of training set were reduced to a final 3272 sentences or approximately 5 hours of speech. Although this is just a subset of this database, training of MLPs is still a time and CPU demanding task, and a net with 234 inputs, 600 hidden units, and 36 outputs takes more than three days to train on an Sun Ultra-SPARC workstation.

Test Set

The test set consisted of 39 different application words as “l'heure”, “le temps” and was selected so that each word has about the same frequency of occurrence. The words were also selected from several test speakers in the Swiss French Polyphone database. A total of 1871 utterances were selected with each word occurring about 48 times.

This isolated word test is thus speaker independent and to some extend vocabulary independent in the sense that the training was not optimized for this particular vocabulary, although some of the words might have occurred in a few of the training sentences.

Results

The results of different experiments are shown in Table 1.

train1.good			
hidden units	1-state	3-state	mean/2
60	79.3%	85.3%	87.7%
200		90.6%	92.4%
600	87.8%	92.4%	94.0%
2000			94.7%

Table 1: Performance on isolated word task using various MLP sizes and different minimum duration. Correct transcriptions were used.

Multi-Layer Perceptrons of different sizes were trained and afterward tested with different minimum durations. The column *1-state* is without any minimum duration and *3-state* is with a minimum duration of 3 frames for all phoneme, implemented as HMMs with three states. To model the durations more closely, histograms were calculated for the duration of each phoneme in the training set. Based on these statistics, the minimum duration for each phoneme was set to half the average duration, and the transition probabilities were calculated so that the HMMs had the same average duration as observed in the training data. In general this means that all transition probabilities are 0.5 which were also used in the first two types of HMMs. Results using the histogram derived durations are marked *mean/2* in the table.

It is seen that the state-dependent minimum duration outperforms a fixed minimum duration of three frames, which again is better than no minimum duration.

Also it is observed that with the current training set the performance saturates as the MLP size is increased. Only one experiment was conducted with a MLP with 2000 hidden units, because the training takes about two weeks on a Sun Ultra SPARC workstation.

1.2.6 Pronunciations and Dictionary

The Importance of Correct Transcriptions

To illustrate the importance of a good transcription, Table 2 is included. The results correspond to those in Table 1, but in Table 2 a transcription error occurred. During the phonetic transcription, approximately 4.7% of the sentences contained comments of the form "[bruit]", "[inspiration]", etc., which by mistake were transcribed as words, that is as if they had been uttered. As is seen, the best results for 600 hidden units dropped from 94.0% to 92.6% Off cause the experiments were done in the opposite order. Here the faulty results serve to illustrate that a correct phonetic transcription is important to obtain good acoustic models (MLPs.)

train1.good, transcription bug			
hidden units	1-state	3-state	mean/2
60	79.5%	85.7%	87.7%
600	87.6%	90.7%	92.6%

Table 2: Performance on isolated word task using various MLP sizes and different minimum duration. Faulty transcriptions were used.

New version of BDLex

In earlier experiments, only an old (unofficial) release of BDLex has been used at IDIAP. Later on IDIAP acquired the last (official) version of that dictionary. Unfortunately, comparing recognition performance showed using the new BDLex dictionary was yielding a small degradation compared to the old version of BDLex. For a 600 hidden unit MLP performance dropped from 94% to 93.4%. For a smaller MLP, however, an improvement of 1.2% was observed, so it is difficult to conclude whether the new BDLex is better or worse than the old one.

Multiple Pronunciations

One of the studies was concerned with the effect of multiple pronunciation. Instead of considering one single possible pronunciation for each word, several pronunciations were allowed for each word. This was implemented by augmenting the phoneme set with artificial phonemes, such as optional phonemes and multiple choice phonemes.

After one MLP training, 93.3% word recognition rate was obtained on the same isolated word task. From this MLP we proceeded with a so-called embedded training, where the acoustic data is re-segmented with the previous MLP and then a new MLP is trained. By repeating this process three times, performance was improved to 94.8% still using 600 hidden units.

1.2.7 Experiments on Operator Calls

Finally, a first series of tests was performed on continuous speech. The acoustic models were the same as in the isolated word case (i.e., same training set).

Training of Language Model

To train a trigram *language model*, 1455 operator calls from the 2000 train speakers were selected. All utterances that were obscured in some way were left out, e.g. sentences with hesitations and unfinished words.

To avoid dependency on specific proper names, which are less likely to generalize, all proper names were treated as one single lexical item in the language model.

The proper names is what makes this task difficult. Each time a proper name occurs in a sentence, there is a choice of more than 3000 proper names.

Dictionary

The dictionary used in these experiments contained 4574 different words. Out of these words approximately 1100 words were obtained from the BDLex dictionary and the remaining words were proper names. As mentioned earlier the proper names were transcribed by a rule-based system followed by a manual correction.

Test Set

The test set were defined: a "true" test set named *test-set* in Table 3 and a cheating test set named *train-set*.

The set marked *train-set* is 294 operator calls drawn from the set of speakers used for training of the MLP and the language model. The acoustic models were not directly trained on these sentences but other sentences spoken by these speakers were used to train the acoustic models. The language model, however, was directly trained on these operator calls as well as other operator calls not used for this test.

The set marked *test-set* is 321 operator calls drawn from the test speakers, which were neither used for training of the acoustic models nor for training the language models.

Results

Language scaling or acoustic scaling is a widely used technique to scale the relative confidence one has in each of the models. Either the language model probabilities are raised to a power greater than one, which is equivalent to multiplying the corresponding log-probabilities or, alternatively, the acoustic log-likelihoods are divided by the same factor. In the experiments reported here, acoustic scaling of 0.2 and 0.15 were used. These values were found to be "optimal" based on a few experiments on a small independent subset.

Table 3 shows performance on the training and test set.

W.E.R. on Operator Calls		
ac.scale	train-set	test-set
0.15	68.8%	60.7%
0.20	59.6%	67.1%

Table 3: Performance on continuous speech: operator calls. Word error rate (W.E.R.) is shown for the training set ant the test set. The set-up uses an MLP with 600 hidden units, trigram language models, and state-dependent minimum durations.

As can be seen in Table 3, the operator call task is much more difficult than the isolated word recognition. It is expected, however, that fine-tuning of the acoustic and language parameters can improve these results. The presented results have to be seen as pilot experiments in this difficult field.

The results show that the recognition system does not perform significantly worse or better on the test-set than on the train-set. However, the best performance on the train-set and test-set is obtained with different acoustic scaling factors. Thus one has to put more emphasis on the language model (acoustic scale = 0.15) for the test-set than for the train-set (acoustic scale = 0.20). This is quite surprising because the language model was in part trained directly on the train-set of table 3 and as such should be expected to cover these sentences quite well. This is especially true because the number of sentences for training the language model was rather limited. However, handling proper names as a single lexical item seems to make the language model cover the test-set as well as the train-set.

Though, it is still difficult to explain why performance on the test-set is better than (not just as good as) the train-set for an acoustic scaling factor of 0.15. More experiments are needed to explore the effect of different parameter settings.

Finally it should be emphasized that the large number of proper names makes this task particularly difficult. In a word sequence without proper names, the language model can help reduce the set of possible words to follow. For proper names the choice is almost unlimited as in "Je voudrais le numéro de NAME" where NAME can be *any* proper name in the dictionary.

2 SWISS-POLYPHONE DATA

IDIAP is responsible for the development and management of Swisscom speech databases. In 1997, two main activities have taken place in the framework of our contract with Swisscom: contract with Swisscom:

1. Management of the Swiss Polyphone databases
2. Recording of a medium size Mobile (GSM) Swiss French database, followed by early recognition experiments.

2.1 Task 2.1: Swiss Polyphone databases management

Actually, Swisscom is the owner of two main databases, the *Swiss French* and *Swiss German* Polyphone.

2.1.1 Swiss French database

As this database was already recorded before 1997, the main work at IDIAP was to adapt the format and to modify the annotation in relation to the European Speechdat project and the STRUT software.

In the version 0.0, the database was stored on CD-ROMs using NIST format and all the files were compressed with the SHORTEN algorithm. In order to simplify the use of the database, it was decided to adapt the format to a more simple and practical one. The new version 1.0, is also using a NIST format, but this time, the files are stored in an a-law speech format. The choice of the a-law format, was decided in order to minimize disk place and to avoid the amplification problem, we had in the 0.0 version.

The structure of the database has also been changed, we have now another file nomenclature. The main goal was to identify the speech files contents, by just looking at the filenames. To avoid confusion problems, all the files have now an unambiguous name. For example, the file *F0054S03.ALW* correspond the female speaker number 0054 and the sentence number three.

In the framework of the European SpeechdatII project, 3000 sessions of Polyphone have been re-annotated. The main goal was to correct errors of annotation and to add specific noises. Four different noises were checked: speaker noises, non-speaker noises, stationary noises and intermittent noises. The new annotation is available in an ASCII format and these 3000 sessions will be available soon within the context of the SpeechdatII project. Now 1000 sessions have been already delivered to Speechdat.

In 1997, IDIAP was also responsible for the diffusion of this database. Work was done for Entropic, ENST and FPMS (Faculté Polytechnic de Mons, Belgium). In March 97, IDIAP sent to Entropic, the first CD of polyphone. A phonemic transcription and a dictionary were also prepared in the same way. For bootstrap models building, we also delivered a set of 625 segmented sentences in ESPS format. For ENST, we delivered a copy of all the spelled items found in this databases. This material was stored on a DAT and sent to ENST in May 1997. In summer, all the Swiss French Polyphone database has been sent to FPMS.

2.1.2 Swiss German database

As this database was still under recording in 1997, IDIAP continued to generate the call sheets within the Swiss German Polyphone project. For the SpeechdatII project, 1000 sessions have been put in SAM format in accordance with the SpeechdatII specifications. These 1000 sessions have been already validated by Spex.

2.2 Task 2.2: Recording of the Mobile Swiss French database

As mentioned in the technical proposal, IDIAP was responsible to specify and record a Mobile Swiss French database. The specifications were elaborated together with Swisscom. A pre-study and initial version of the specifications was released as **T0+6 deliverable** and is given in **Annex B**. Final specifications as used for the collection of the final database (as briefly described below) are given in **Annex C**.

2.2.1 Final specifications and topology of the database

The content of the database is composed of 7 items. We have 4 items with a sequence of 6 digits - including $\star$ - and # and 3 yes/no questions. All the 12 digits are represented twice for each call as in the following example:

1	Lisez la séquence de chiffres	7 1 6 $\star$ 3 #
2	Lisez la séquence de chiffres	2 8 0 5 4 9
3	Lisez la séquence de chiffres	1 3 8 4 0 #
4	Lisez la séquence de chiffres	9 2 7 5 $\star$ 6
5	Êtes-vous de sexe masculin ?	
6	Êtes-vous âgé de plus de 45 ans ?	
7	Êtes-vous âgé de moins de 30 ans ?	

Table 4: Call sheets content

Given that IDIAP did not have much experience in recording mobile databases, we specified different objectives. The main goal was to record 600 calls from all the Swiss French speaking part of Switzerland with a good coverage over age, calling places and environments. Table 5 summarizes the final characteristics of the collected data compared to the targeted objectives.

As we can see on this table, different problems were underestimated, we can resume these problems as following:

- Difficulty to persuade people more than 30 to call the server. Young people do accept it more easily.
- In Geneva, people were not really cooperatives.
- Most appropriate place to ask people for calling is in the street. Difficulty to have calls from train or from inside places.
- November was not a really good period for this kind of work, it is preferable to do it in summer, when wheater is more pleasant and people have more time.

To complete this database, 50 another calls will be recorded at IDIAP. These calls will be also recorded on DAT platform in the same way. Annotation will be done next year.

	Obectives	Results
Number of calls	600	558
Repartition male/female	50%/50%	47%/53%
Age coverage :		
- less than 30	25%	61%
- between 30 and 45	25%	17%
- more than 45	25%	18%
Geographical covarage:		
- Vaud	25%	36%
- Genève	20%	5%
- Neuchâtel	15%	25%
- Valais	15%	13%
- Fribourg	15%	13%
- Jura	10%	8%
Environment topology:		
- calls quiet	40%	19%
- calls from street	20%	64%
- public places	20%	10%
- noisy calls	15%	5%
- from cars, train	5%	1%

Table 5: GSM database typology

2.2.2 Recording procedure

The database was recorded on a SUN ISDN platform. To have variability on GSM quality, two mobile phones were bought and used to collect the data: 1 ERICSSON GA 628 and 1 NOKIA 8110.

To collect this database, a person was recruited by IDIAP and was responsible to go to different places all over swiss french speaking part of Switzerland. This person had to persuade people for calling the server.

First recording tests began in September 1997 and we noticed that the task was more difficult than expected. In general, the recruiting person need between 10 and 20 minutes to find and persude people for calling the server. The first recruited person give up to do this job after a few days. Finally, another person accepted to do these recordings in November 1997 and the work was done in one month. In the average, we recorded 22 calls per day and during 25 days.

For further information, the person who collected (Sophie Loye – thank you to her) the data and recruited people was asked to write a little report, which is given in Annex D.

2.2.3 Preliminary GSM Recognition Experiments

As planned as T0+12 deliverable, first recognition tests were performed at IDIAP on a subset of this GSM database. The main goal of these tests was to assess the robustness of automatic recognition of isolated digits on GSM network in comparison with regular (fixed) telephone network.

In all the preliminary experiments reported below, HTK was used for training and recognition. In the future, hybrid models will also be used for training and recognition and will be compared to the performance achieved with HTK.

Feature extraction – The speech features used are the mel-frequency cepstral coefficients parameters (MFCC). For each frame we calculated 12 MFCC coefficients, with the delta and acceleration coefficients. We also used the energy coefficient, as well as delta and acceleration of the energy, which make a total of 39 coefficients. These features were calculated every 10 ms using a window size of 25 ms with a pre-emphasis coefficient of 0.98.

Training set – The models were trained on a subset of the Swiss French Polyphone database, including 521 items (264 male and 257 female speakers). The vocabulary used for the training set is composed of the 10 french digits, included the # and $\star$ symbols. For each digit we created Hidden Markov Models (HMM) with between 10 and 14 states (depending on digits) with 5 gaussian density functions for each state. The frequency and topology for each digit model is describing in Table 6.

Digits	Frequency	HMM states	Digits	Frequency	HMM states
un	232	10	deux	225	12
trois	225	10	quatre	225	12
cinq	249	12	six	227	12
sept	259	12	huit	393	12
neuf	337	12	zéro	241	14
dièse	281	12	étoile	232	14

Table 6: Training set and HMM topology

Two sets of models have been tested and compare for the recognition of real GSM speech:

- Train set A: Models trained on regular telephone data (without GSM coder).
- Train set B: Models trained on GSM encoded/decoded data. In this case, all the training data (from fixed network) was first encoded and decoded with a GSM sotware in order to better reflect GSM quality.

For both training sets, we used exactly the same segmentation and the same HMM topologies.

Test Set – The recognition test was performed on a subset of the GSM database. This subset is composed of 68 speakers and 272 items. All the french digits and # and $\star$ symbols are included.

Preliminary Recognition Results – Based on the results reported in Table 7, we have the following conclusions:

1. Models trained on training set B are yielding slightly better results.
2. On average, results are not too bad compared to regular telephone data performance, although being a little bit on the "low end" (which could also be explained by the fact that the database has not been cleaned up and properly labeled yet).

It is expected that these preliminary results will be significantly improved in the future (e.g., by using efficient noise cancellation techniques and more robust models, such as hybrid HMM/ANN systems).

	Train set A	Train set B
Word	90.48 %	91.53 %
Sequence of 6 digits	56.99 %	61.40 %

Table 7: Performance on digits sequence

The recognition confusion matrices resulting of the two training sets A and B are given below and provides us with more detailed results.

Confusion Matrix for Train Set A :

```
------------------------ Overall Results -------------------------
SENT: %Correct=56.99 [H=155, S=117, N=272]
WORD: %Corr=90.48, Acc=85.72 [H=1464, D=63, S=91, I=77, N=1618]
------------------------ Confusion Matrix ------------------------
       d   e   u   d   t   q   c   s   s   h   n   z
       i   t   n   e   r   u   i   i   e   u   e   e
       e   o       u   o   a   n   x   p   i   u   r
       s   i       x   i   t   q       t   t   f   o
       e   l           s   r                           Del [ %c / %e]
dies  131  0   0   0   0   0   0   1   0   0   0   0    2 [99.2/0.1]
etoi   1  131  0   1   0   0   1   1   0   0   0   0    1 [97.0/0.2]
  un   0   0  117  0   1   0   2   0   0   2   1   0   12 [95.1/0.4]
deux   2   0   2  117  0   0   1   0   0   1   0   1    9 [94.4/0.4]
troi   0   4   1   0  122  0   1   1   0   0   2   0    4 [93.1/0.6]
quat   0   0   1   0   0  121  0   1   0   0   8   0    4 [92.4/0.6]
cinq   1   1   1   0   0   0  124  2   0   1   1   0    5 [94.7/0.4]
 six   1   0   0   0   0   0   0  116  2   4   0   2   10 [92.8/0.6]
sept   5   0   0   3   0   2   0   2  111  1   7   0    5 [84.7/1.2]
huit   0   0   0   0   0   0   0   9   1  119  0   0    6 [92.2/0.6]
neuf   0   1   2   0   0   0   0   1   0   0  125  0    4 [96.9/0.2]
zero   0   0   4   0   0   0   0   0   0   0   0  130   1 [97.0/0.2]
Ins    4   4   8   5   1   4   5  25   4   7   9   1
==================================================================
```

Confusion Matrix for Train Set B

```
------------------------ Overall Results -------------------------
SENT: %Correct=61.40 [H=167, S=105, N=272]
WORD: %Corr=91.53, Acc=88.26 [H=1481, D=39, S=98, I=53, N=1618]
------------------------ Confusion Matrix ------------------------
       d   e   u   d   t   q   c   s   s   h   n   z
       i   t   n   e   r   u   i   i   e   u   e   e
       e   o       u   o   a   n   x   p   i   u   r
       s   i       x   i   t   q       t   t   f   o
       e   l           s   r                           Del [ %c / %e]
dies  130  0   0   0   0   0   0   0   0   0   0   1    3 [99.2/0.1]
etoi   1  129  1   1   0   0   0   2   0   1   0   0    1 [95.6/0.4]
  un   0   0  117  0   2   0   5   1   0   0   2   0    8 [92.1/0.6]
deux   9   0   2  110  0   0   1   1   0   1   5   0    4 [85.3/1.2]
troi   0   3   1   0  125  0   1   0   1   0   0   0    4 [95.4/0.4]
quat   1   0   2   1   0  119  0   0   1   0   7   0    4 [90.8/0.7]
cinq   3   0   2   0   0   0  126  4   0   0   0   0    1 [93.3/0.6]
 six   1   0   0   0   0   0   0  125  0   3   0   1    5 [96.2/0.3]
sept   7   1   1   2   0   1   1   1  112  0   5   0    5 [85.5/1.2]
huit   0   0   0   0   0   0   0   3   1  130  0   0    1 [97.0/0.2]
neuf   1   0   0   0   0   0   0   1   0   0  129  0    2 [98.5/0.1]
zero   0   0   2   0   1   0   0   0   0   0   2  129   1 [96.3/0.3]
Ins    3   1   8   4   0   2   0  18   5   5   6   1
==================================================================
```

References

[1] Bourlard, H. & Morgan, N., *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Press, 1994.

[2] Hermansky, H. (1990). Perceptual Linear Predictive (PLP) Analysis of Speech, *Journal of the Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752.

[3] Hermansky, H., Morgan, N., & Hirsch, H., 1993. Recognition of speech in additive and convolutional noise based on RASTA spectral processing, *Proc. IEEE Intl. Conf. Acoustics, Speech & Signal Processing*, vol. 2, pp. 83-86

[4] STRUT: Speech Training and Recognition Unified Tool, `http://tcts.fpms.ac.be/`