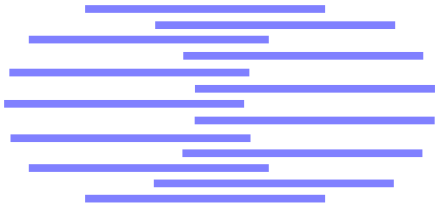
IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse



INTEGRATING SPEECH ACOUSTIC AND LINGUISTIC CONSTRAINTS: BASELINE SYSTEM DEVELOPMENT

Giulia Bernardis [a] [b] Hervé Bourlard [a] [b]
Martin Rajman [a] Jean-Cédric Chappelier [a]

IDIAP–RR 99-21

NOVEMBER 1999

SEE ALSO
TECHNICAL REPORT 99-324, DI - EPFL

Dalle Molle Institute for Perceptual 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] Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
[b] Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), Martigny, Switzerland

INTEGRATING SPEECH ACOUSTIC AND LINGUISTIC CONSTRAINTS: BASELINE SYSTEM DEVELOPMENT

Giulia Bernardis Hervé Bourlard Martin Rajman
Jean-Cédric Chappelier

NOVEMBER 1999

SEE ALSO
TECHNICAL REPORT 99-324, DI - EPFL

Abstract. In this report, we discuss the initial issues addressed in a research project aiming at the development of an advanced natural speech recognition system for the automatic processing of telephone directory requests. This multi-faceted project involves (1) text processing (labeling and tagging) of a large database of telephone-based natural voice requests (including all kinds of peculiarities), (2) development of robust acoustic models, (3) integrating advanced natural language (syntactic and semantic) constraints, (4) detecting and dealing with a large number of out-of-vocabulary words (proper names), and (5) testing of the resulting system on natural queries. All this work will be performed on the basis of a database containing prompted (read) speech and (simulated) natural requests to information service.
This report describes the initial steps that were required to set up a reasonable baseline system and a good research and evaluation framework. More specifically, a significant amount of time was devoted to proper text processing of speaker request transcriptions, in order to create the basis necessary for the lexical and linguistic modeling, as well as for the evaluation of recognition results.

Contents

1 Introduction

This activity report describes the research and development work performed in the framework of the "INSPECT" Swiss National Science Foundation project on "INtegrating SPEech (acoustic and linguistic) ConsTraints for enhanced recognition systems" and covers the period of February 1 - October 31, 1999.

The INSPECT project focuses on unconstrained speech recognition over the telephone line, in view of improving dialog-based Interactive Voice Response (IVR) systems. The application framework chosen for the evaluation of our work is the one that was initially developed in collaboration between EPFL, IDIAP, ISSCO and Swisscom for partial automation of the 111 information service of telephone-based phonebook inquiry.

In the framework of that collaboration, a large database, referred to as the **Swiss-French Polyphone** database [9, 2], had been collected[1] over the Swiss telephone network and contains numerous speakers pronouncing:

1. Prompted sentences: this part of the database is referred to as the **Polyphone** database and was designed to cover all the phonemes in a large variety of contexts for the Swiss-French language.

2. Unconstrained requests, referred to as **"Appels 111"**, where speakers were invited to simulate information requests (typically about telephone numbers or addresses of people).

The generic goal of the INSPECT project is to (initially) train acoustic models on the prompted speech and to test and perform research on the unconstrained requests. However, while the main aim of the project is to improve the integration of acoustic and linguistic (syntactic and semantic) constraints, this also required setting up a good research and evaluation framework. For example, one of the main problems we quickly faced was the unavailability of a properly tagged and consistent database of unconstrained speech, together with its associated lexicon. This was a major handicap for further research in the targeted area, and was even making the assessment of our current speech recognition systems impossible since there was not proper reference point. The reason why we decided to use this "Appels 111" database, in spite of these serious drawbacks, is that it is the only unconstrained speech database currently available in Swiss-French.

Consequently, during the first year of this project, we mainly focused our work on the following tasks:

- Acoustic modeling of telephone-based Swiss-French speech.
- Lexical modeling.
- Linguistic modeling of spontaneous requests.
- Improvement of interaction between acoustic, lexical and linguistic modeling layers.
- Setting up of a reference system for **Large Vocabulary Continuous Speech Recognition (LVCSR)**, with initial tests on the 111 service calls subset of the Swiss-French Polyphone database.
- Preliminary experiments with different recognizers (using different information) to be considered as a mixture of experts.
- Development of efficient modules for assigning confidence scores to hypothesized transcriptions and rejection of uncertain data.

This report first presents a description of the Swiss-French Polyphone database and then details further what has been achieved in the former items.

[1] Also in the framework of the European SpeechDat project.

2 Swiss-French Polyphone and "Appels 111" Databases

The **Swiss-French Polyphone** database [2] contains telephone calls from about 4,500 speakers recorded over the Swiss telephone network.

The calling sheets were made up of 38 prompted items and questions (see appendix A for a sample of a calling sheet) and were distributed to people from all over French speaking part of Switzerland.

Among other items each speaker was invited to:

- read 10 sentences selected from several corpora to ensure good phonetic coverage for the resulting database (in the following we will refer to them and more generally to the Swiss-French Polyphone subset of all the prompted sentences as to Polyphone database);
- simulate a spontaneous query to the telephone directory (giving the name and the address of the queried person), i.e., simulate a 111 information service call.

In particular, the Swiss-French Polyphone subset of the items related to the 111 service calls represents the application framework of the INSPECT project and in the following we refer to it directly as to "Appels 111" database.

The **"Appels 111"** database contains 4,293 recordings (2,407 female and 1,886 male speaker recordings), each consisting of 2 files: an ASCII file (`.txt`) corresponding to the initial prompt and address request, and a data file (`.alw`) of the recorded speech in a-law format along with a NIST header containing, among other information, the transcription of the speaker request. See appendix B for further description of data.

The main goal of our work is to train robust acoustic and language models that can be used to recognize the relevant information present in the "Appels 111" database.

3 Baseline System Development

The development of a baseline system (to be used as a reference against which to compare future recognition results) able to process 111 requests involved the training of acoustic models, as well as the extraction of lexical and linguistic models.

In our baseline system, initial **acoustic models** have been trained on the Polyphone database, which is (relatively) properly labeled. This will be briefly described in section 3.1.

In standard speech recognition systems, the lexicon models are usually a priori defined on the basis of a dictionary containing all possible words with their phonetic transcription. However, not all the words used in the "Appels 111" database are present in the Polyphone database. Furthermore, a language model is usually trained on a large amount of text corpora reflecting at best the conditions of use of the recognizer. In our case, the Polyphone database (containing prompted speech) is again not representative at all of the targeted application.

For both the **lexical modeling** (section 3.3) and the **language modeling** (section 3.4), it is thus important to have access to a large amount of properly labeled (orthographically transcribed) and tagged (by any additional information) "Appels 111" database. Obviously, this will also be necessary to assess the performance of our recognizer. Unfortunately, the text data that was available at the beginning of the present project for the "Appels 111" database was not properly tagged, and was often even inconsistent (e.g., "Bonjour mademoiselle j'aimerais le numéro de téléphone de madame groux-Fazan Anne-Lise vous voulez que je vous l'épelle non ah! très bien ça m'arrange j'ai déjà [\hésitation épeler] deux fois alors elle habite à Villarey [\hésitation euh] Cousset [\hésitation euh] je pense c'est soit dans le canton de Vaud soit dans"). Consequently, as described in section 3.2, a significant amount of time was also devoted to the text processing of the "Appels 111" to make them useful for lexical and linguistic modeling.

3.1 Acoustic Models

3.1.1 Feature Extraction

The preprocessing of the speech signal consisted of a RASTA-PLP feature calculation. RASTA-PLP features are particularly robust to convolutional and additional noise, so they are well suited for telephone speech.

12 RASTA-PLP coefficients along with their first derivatives (12ΔRASTA-PLP) as well as Δ-log- and $\Delta\Delta$-log-energy (which makes a total of 26 features) were calculated every 10 ms using a window of 30 ms.

3.1.2 Hybrid HMM/ANN Acoustic Modeling

Among the speech recognition systems available at IDIAP, the hybrid HMM/ANN, integrating Hidden Markov Models (HMM) and Artificial Neural Networks (ANN) is one of the most popular one. In a hybrid HMM/ANN speech recognition system, the ANN takes preprocessed acoustic data (features) as input and estimates posterior probabilities for different classes (e.g. phonemes) as output. In this project, a particular form of ANN, referred to as Multi-Layer Perceptron (MLP) was used to compute local emission probabilities of HMMs.

While yielding similar or better recognition performance than other state-of-the-art systems, this approach has indeed been shown to have several additional advantages particularly interesting in the framework of the present research, such as:

- A small set of HMM/ANN context-independent phone models is yielding similar performance than a large set of HMM content dependent phone models.
- Given the above, development of new tasks (involving different lexica) is easier. It has also been reported that generalization across tasks (training and test set containing different words) was more robust.
- As a consequence, hybrid HMM/ANN systems are usually easier to implement and to modify, allowing us to focus our research on those most interesting aspects.
- Less memory and CPU are required.

In the case of context-independent phonetic models as used here, the output layer of the MLP contained 36 units, corresponding to 35 basic phonetic units (from the SAMPA phoneme set) and a silence state. As usually done with MLPs in hybrid HMM/ANN systems, a context of 9 consecutive feature vectors (4 vectors before and after the current one) were used as input to the neural network, giving a total of 234 units in the MLP input layer. There were 600 units in the hidden layer.

Software developments were done in the framework of the STRUT (Speech Training and Recognition Unified Tool) toolkit [15].

3.1.3 MLP Training

The training of the MLP was performed on the Swiss-French Polyphone database (see [2]), using the 10 phonetically rich sentences read by each of 400 speakers (200 male and 200 female speakers). During the experiments, different kinds of irregularities (e.g. noise on the recording, strange utterances) were discovered, and the training set was finally reduced to $3,272$ sentences, corresponding to approximately 5 hours of speech.

In order to train an MLP, it is necessary to have the phonetic segmentation (and transcription) associated with each utterance, i.e., one of the 36 possible phones must be assigned to each of the acoutsic frame. For the training data only an orthographic transcription was available. As hand-segmentation is a very time consuming task, the orthographic text was first converted into a sequence of phones which was then matched to the speech signal using a forced Viterbi alignment. As a result,

to each frame was assigned one of the phones in the phonetic sequence. From this point, the MLP training was performed with the standard error back-propagation method.

3.2 Processing of "Appels 111" Text Data

As far as the address prompt files and the speaker request orthographic transcriptions are concerned, the data in the "Appels 111" database are unfortunately not properly tagged and even inconsistent (e.g., "Bonjour mademoiselle j'aimerais le numéro de téléphone de madame groux-Fazan Anne-Lise vous voulez que je vous l'épelle non ah! très bien ça m'arrange j'ai déjà [\hésitation épeler] deux fois alors elle habite à Villarey [\hésitation euh] Cousset [\hésitation euh] je pense c'est soit dans le canton de Vaud soit dans").

As discussed later, proper orthographic transcription and tagging of our database is however very important to the development and testing of our recognition system since this will be used to:

1. Automatically create the lexicon from the orthographic transcription of the speech utterances. In our case, each lexicon entry will then be (automatically) extracted as any character string between two empty spaces.

2. Automatically model syntactical constraints (grammar model) in terms of this lexicon. This will be one important aspect of the present INSPECT project.

3.2.1 Address Prompt Reformatting and Information Extraction

After correction of miscellaneous errors (see appendix C), the last three lines corresponding to the address were extracted from the address prompt files of the "Appels 111" database and processed with the following heuristic [8]:

line1 = name line2 = street, if line3 is not empty, and = town, otherwise line3 = town, if non empty

to obtain something like this example:

name: MOTTAZ MONIQUE
street: rue du PRINTEMPS 4
town: SAIGNELEGIER

Notice that the proper nouns in the (available) fields are written in capital letters and without accent informations.

Using the name fields, three lists were created:

- *family names* (e.g. "VON GUNTEN-BIGLER");
- *first names* (e.g. "ALAIN JEAN-LOUIS");
- *company/institution/organisation names* (e.g. "CHAMPOUSSIN SERVICES").

From the street fields a list of

- *street/building names* (e.g. "PRINTEMPS")

was extracted, as well as a list of street/building introduction expressions (e.g. "rue de l'", "place du", "bâtiment de la").
Finally, from the town fields, a list of

- *locality names* (e.g. "VAL-D'ILLIEZ")

was obtained.

3.2.2 Text Processing of Speaker Request Orthographic Transcriptions

The low quality level of transcriptions available for the "Appels 111" database does not allow to use them for lexical and/or linguistic modeling.

Speaker request orthographic transcriptions contain many peculiarities (for more details, see instances in appendix C), including:

- Numerous transcription errors (deletions, insertions, substitutions)
- Undocumented specific information annotations
- Spelling errors (typing or orthographic mistakes)
- Syntactic errors
- Lack of uniformity in the transcription style: specific information annotations, use of accents, capitalization and punctuation are not uniform at all nor objective, but really depend on the transcribed sentence
- Undocumented abbreviations.

Much work was thus required to correct spelling and syntactic errors, and to minimize transcription errors. Additional text processing was also necessary to convert the raw orthographic transcriptions into a format suitable to automatically extract the lexicon entries (defined as any character string between two spaces) and more profitable to estimate language models parameters and to be used as reference for the speech recognition results.

To do this, some preliminary text processing operations were required:

- Markers were inserted to indicate the beginning ("<s>") and the end ("</s>") of each speaker utterance (such context cues being important later on for a proper use of both the Carnegie Mellon University Statistical Language Modeling toolkit [13] and the Noway software [11, 12]).
- "Noise" represented by non documented, non uniform annotations (e.g. "[\prononciation bizarre]", "[hésitation]", "[\inintelligible]", ...) was filtered out.
- Punctuation signs ("!", """, ",", ".", ":", ";", "(", ")", ">", "?") were removed.
- Dashes in words were kept or added (e.g. "est-ce", "Jean-Pierre", ...) to include composed words as single lexical entries.
- Split on apostrophe (e.g. "c'est" → "c' est", ..., but "aujourd'hui" was not split, ...).
- Capital letters were lowercased (e.g. "Je" → "je", "Camping TCS" → "camping tcs", "Louis" → "louis", ...).
- Abbreviations and special signs were spelled out (e.g. "ch" → "chemin", ...).

Next, from the transcriptions a list of the not proper noun words was extracted (*general vocabulary list*) and, using the functionalities of the SLPToolKit software (Syntactical Language Processing ToolKit developed at the EPFL's Artificial Intelligence Laboratory [6, 7]), several iterations through the following steps were done:

1. Creation of a lexicon in SLPToolKit binary format from the ASCII *general vocabulary list* (above mentioned) and proper nouns lists (*family names*, *first names*, *company names*, *street names* and *locality names*) extracted from the address fields of prompt files (see section 3.2.1) and written, as consequence, in capital letters.
2. String correction of orthographic transcription sentences based on this lexicon with

- very low penalty ($d = 0.005$) for the correction of accents and the transformation lowercase–uppercase;
- maximal distance granted for the lexical research $D_{max} = 0.1$;
- research mode enabling all the solutions at distance $D <= D_{max}$.

3. Choice of one solution for each sentence between many possible corrections.

4. Integration of missing proper nouns in family names, first names, company names, street names, and locality names lists, and not proper noun words in general vocabulary list.

An example text segment of the speaker request orthographic transcriptions after such processing is shown below:

```
<s> j' aimerais le numéro de téléphone de MEIZOZ-MENDES MARIE-ANTOINETTE CHIPRES quatorze
au LANDERON </s>
<s> madame auriez-vous la gentillesse de me donner le numéro de téléphone de KREBS VINCENT
rue de l' AVENIR douze à COURT </s>
```

3.2.3 Text Data Statistics

The "Appels 111" database includes 4,293 speaker requests orthographic transcriptions, but actually a few of them are cut off and 1 is empty.

After the text processing described in 3.2.2, text data contain 77,702 words occurrences of 5,549 different word forms, which leads to an average length of 18 words per speaker request.

In the following, we give the most frequent word forms, together with their frequencies (word counts):

de 7753	avoir 780	a 287	voudrais 188
le 3804	oui 776	vingt 285	trente 188
numéro 3722	merci 720	l' 278	pouvez-vous 184
téléphone 3236	donner 698	r 273	quatre 180
à 2970	rue 523	route 272	veuillez 177
bonjour 2759	habite 453	connaître 248	et 174
j' 1971	qui 452	des 233	sept 168
aimerais 1863	euh 405	m' 232	poste 166
monsieur 1693	la 395	pourrais 223	c' 164
vous 1676	que 378	indiquer 212	dix 163
madame 1564	e 377	deux 210	six 158
il 1436	un 364	o 206	cent 158
s' 1409	est-ce 348	est 199	pourrais-je 156
plaît 1357	chemin 344	trois 195	l 155
mademoiselle 869	pourriez-vous 341	huit 192	t 152
me 836	savoir 323	n 190	au 148
je 811	du 319	i 189	adresse 144

Figure 1 shows the distribution of word frequencies. There are 2,236 hapax (i.e. 40%) and 74% of the words occur less than 3 times.

Figure 2 shows statistics for n-grams in the "Appels 111" text data. As it can be expected, the majority of n-grams occurs only once (70% of 2-grams, 80% of 3-grams and 86% of 4-grams) and there are very very few n-grams occurring at least 10 times (3% of 2-grams and 1% of 3-grams and 4-grams). This is a clear indicator of a text database too small to reliably estimate language model parameters.

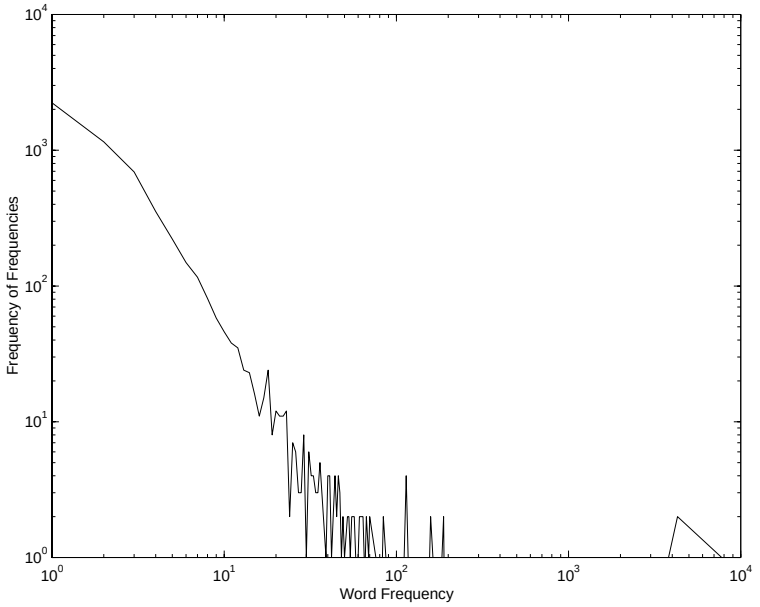


Figure 1: Number of occurrences of word frequencies (in log-log scale).

3.2.4 Named Entity Tagging

The largest proportion of words in the text data having very low unigram frequencies are proper nouns.

The accurate identification of proper nouns or, more generally, **Named Entities (NE)** in spoken language is likely to be an essential component of systems performing tasks such as speech understanding, information extraction and information retrieval. Furthermore approaches based on the use of NEs have the potential to improve the performance of LVCSR systems.

In consequence of those considerations, an operation of NE tagging was carried out on the "Appels 111" text data. The following 5 NE categories, and corresponding NE tags specified in brackets, were defined:

- Family Name (<N> standing for "Nom de famille");
- First Name (<P> standing for "Prénom");
- Company (<I> for "Institution") including company, institution, organisation names;
- Street (<R> for "Rue") including street, square, building names;
- Locality (<V> standing for "Ville") including town, village, canton, country names;

and a copy of the text with NE expressions marked up in SGML format was produced. An example segment is shown below:

```
<s> j' aimerais le numéro de téléphone de <N>MEIZOZ-MENDES</N> <P>MARIE-ANTOINETTE</P>
<R>CHIPRES</R> quatorze au <V>LANDERON</V> </s>
<s> madame auriez-vous la gentillesse de me donner le numéro de téléphone de <N>KREBS</N>
<P>VINCENT</P> rue de l' <R>AVENIR</R> douze à <V>COURT</V> </s>
```

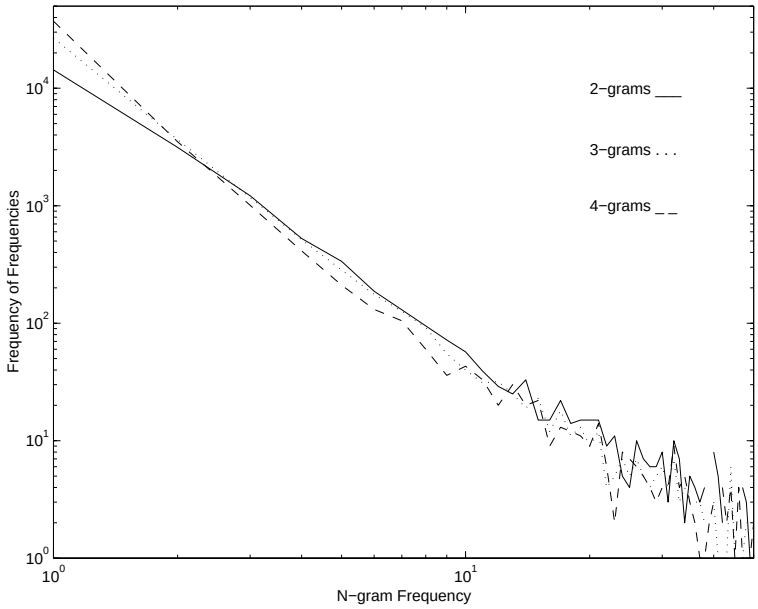


Figure 2: Distribution of frequencies of 2-grams, 3-grams, 4-grams (double logarithmic plot).

In addition to NE tagging, all the spelled items found in speaker requests orthographic transcriptions were labeled and the corresponding sentences (366 orthographic transcriptions) were marked. They will be treated in a separate framework, because the problem of spellings recognition represents a completely different and very difficult research topic.

3.2.5 Splitting

"Appels 111" data were split into three parts:

- A test set of 500 sentences, to be kept apart for final tests.
- A cross-validation set of 200 sentences on which doing recognition experiments and tuning system parameters.
- A training set consisting of the remaining sentences on which learning language models.

3.3 Lexical Modeling

Building upon the general vocabulary list and proper nouns lists extracted from "Appels 111" data, a lexicon was derived containing the words and their NE category information, together with their phonetic transcription.

3.3.1 Phoneme Models

The phonemes used in the phonetic transcriptions were those of the SAMPA phoneme set for French [14]. Their models were constructed building upon 35 basic phonetic units and 1 silence state corresponding to the 36 MLP outputs. The structure of all phone HMMs was left-to-right with a minimum duration constraint set to half the average duration of each phoneme in the training data, which has been shown to yield the optimal performance. For more details see appendix D.

3.3.2 Phonetic Transcription

The phonetic transcriptions of the general vocabulary words were obtained from the BDLex-50000 dictionary [3]. For these words, different phonetic transcriptions were introduced in the lexicon, to take into account the phenomenon of "liaison", very common in French, and to enrich the lexical modeling with pronunciation variants.

Proper nouns were automatically transcribed by a rule-based grapheme-to-phoneme transcription system available at IDIAP, or manually in case of failure of the automatic system. Since information concerning the correct accents of proper nouns was missing, several pronunciations of these words were generated, corresponding to different plausible accentuations.

3.4 Linguistic Modeling

After the raw Appels 111 textual data had been transformed into a more profitable format (see section 3.2.2), the text corresponding to the training part of speaker request orthographic transcriptions was processed into various back-off n-gram language models using the Carnegie Mellon University Statistical Language Modeling (CMU SLM) Toolkit [13].

Most speech recognizers today use n-grams (that calculate the probability of a word in a test set by considering the $n-1$ most recent words) as stochastic language models. More precisely, bi-grams and tri-grams were trained, both based on simple word units (that is, words delimited by space characters), and allowing proper nouns composed of more words as single units (e.g., "DENTS DU MIDI" $\rightarrow$ "DENTS_DU_MIDI").

Good-Turing discounting was applied, with a discounting range of 7 (which is a typical choice, [10]).

An open vocabulary model which covers situations where no out-of-vocabulary words (OOV) occurred in the training data, but some out-vocabulary words could occur in the test data was used. This situation, in fact, corresponds to our case because we had a limited amount of training data and we chose a vocabulary providing 100% coverage of the training set. Consequently, a proportion of the unigram discount probability mass of 0.5% was allocated to out-of-vocabulary (OOV) words.

Finally, to prune the LMs, several different cutoff values (0, 1 and 2) were specified for 2-grams and 3-grams. Language models were stored in the ARPA-standard format, because of its compatibility with the Noway LVCSR decoder [11, 12].

4 Related Research Work

In this section, we briefly describe some related developments that have already been documented elsewhere and/or that will be exploited in the future of the present project.

4.1 Using Multiple Recognizers

Natural speech recognition represents a very difficult task and traditional automatic speech recognition systems performance usually remains poor. This is due to several factors, including strong coarticulation and large phonological variability, but also to the fact that it is difficult to properly model the syntactical constraints underlying unconstrained speech. Indeed, spoken language usually does not follow the rules of read speech, the only case for which a lot of text training data is actually available. Furthermore, these syntactical constraints should be different and adapted according to the type of request (which is not know a priori) or the position in the dialog.

Consequently, it could be interesting to consider the possibility of using and combining different recognizers, working in parallel with different information or different syntactical constraints. In this framework, we started to study the possibility of using multiple recognizers used as a mixture of experts, providing solutions with complementary errors.

In our initialy study, we mainly focused on speech recognizers using syntaxes of different complexities. As already explained in section 3.4, most speech recognizers today use n-grams as stochastic

language models. Also, and depending on the amount of (text) training data available, tri-grams are yielding better performance than bi-grams and mono-grams. However, they also have a tendency to have a long term effect of mistakes. For example, it has often been observed that grammars of different orders were yielding different mistakes.

Consequently, some preliminary experiments were carried out on a small set of read sentences from the Swiss-French Polyphone database. Three recognizers with different language model information (recognizer 1: no language model, recognizer 2: 2-grams, recognizer 3: 3-grams) were used and the output was like:

```
reference:     ``déchausse toi avant d' entrer''
recognizer 1:  ``déchausse toit viande entrer''
recognizer 2:  ``déchausse toi avant de prix''
recognizer 3:  ``et ce au point avant d' entrer''
```

Namely, results indicated clearly that a lower Word Error Rate could be obtained by properly combining the different recognizers output hypotheses into a final solution. However, the recombination strategy to take advantage of this property still has to be investigated.

4.2 Confidence Measures

When a continuous speech recognition system is fielded to a large community of users, especially infrequent users (which, of course, is exactly the case of the callers of a telephone information service), it is common that many spoken inputs do not fall within the domain that the recognition system is able to handle. This may be due to speech disfluencies on the part of the user (e.g., hesitations, word fragments, corrections, sounds like um's and ah's, heavy breaths), utterances that contain out-of-vocabulary (OOV) words, incomplete language models (i.e., the system does not model all the word strings user say), or a poor understanding of the domain by the user (i.e., the user does not understand the range of inputs allowable in the interaction).

The ability to detect the breakdown of the speech recognizer or reject erroneous recognition answers is essential in the design of the system. Consequently, some research work focused on the derivation and evaluation of several confidence measures and led to very interesting results: see [4] and [5] for a detailed description.

It has been shown in section 3.2.3 that 74% of the words forms present in the "Appels 111" occur less than 3 times in the database and that 40% of them are hapax. It has to do especially with proper nouns. For an application like this, it would be not realistic to believe that the speech recognition system will be provided with a close vocabulary containing the lexical models of all possible words. On the contrary, we suppose that it will be unavoidable to deal with many out-of-vocabulary words, which will be detected by computing the confidence measure associated with each recognized word.

Consequently, in the sequel of the INSPECT project, our work on confidence measures could thus be exploited in different ways:

1. To detect out-of-vocabulary words, typically proper names, that will then be processed independently (e.g., by a recognizer more appropriate to the task, and dealing only with the identified OOV segment).

2. Finally, when using several recognizers working in parallel with different linguistic constraints to provide complementary information, confidence scores could be used to recombine the resulting hypotheses.

5 Conclusion

The INSPECT research project aims at developing advanced natural speech recognition systems for the automatic processing of telephone directory assistance. While improving the integration of acoustic

and linguistic constraints, this project should thus develop advanced speech recognition systems able to cope with large vocabulary (including out-of-vocabulary words and proper names) natural speech.

The first requirement was thus to set up a good research and evaluation framework. Indeed, one of the main problems we quickly encountered was the unavailability of a properly labeled and tagged database of unconstrained speech, together with its associated lexicon.

The present report described the initial steps in the development of a good baseline system. Initial acoustic models were trained on the Polyphone part of the Swiss-French Polyphone database, which contains prompted speech (relatively) properly labeled and phonetically rich (section 3.1).

However, Polyphone is certainly not representative enough of the targeted application, and the lexical and linguitsic models (sections 3.3 and 3.4, respectively) had to be extracted from the "Appels 111" part (containing the unconstrained speech) of the Swiss-French Polyphone database[2]. Unfortunately, this database was not properly labeled and tagged, and included numerous peculiarities.

Consequently, text processing of the orthographic transcriptions of natural queries (section 3.2) was required to standardize the database and make it useful to the training of lexical and linguistic models, as well as to the evaluation of the recognition results.

Finally, on top of the description of the baseline system, related research issues were briefly addressed, including the possibility of using several recognizers working in parallel with different linguistic constraints (section 4.1) as well as the use of advanced confidence measures for the detection of out-of-vocabulary word segments (section 4.2).

[2]Obviously, not all the "Appels 111" database will be used to train the language model, and a section will be left on the side for testing. For the lexical models, different possibilities will be considered, depending on whether we want to test our systems with or without OOV words.

Appendices

A Calling Sheet

Here is an example of calling sheet used during Swiss-French Polyphone data collecting [9, 2].

Votre appel téléphonique au N° gratuit 155 28 14

Tout d'abord, au nom de la recherche scientifique suisse, de l'économie en général et plus particulièrement des Télécom PTT et de l'institut LINK, nous tenons à vous remercier de votre précieuse collaboration.

Directives sur la façon de procéder

Lisez attentivement les quelques directives suivantes :

- **lors de l'appel :**
  - **la voix enregistrée du répondeur annonce les numéros en gras (colonne de gauche) de chaque rubrique**
  - **le répondeur enchaîne avec l'énoncé explicatif de la demande;** à titre purement informatif cet énoncé est mentionné dans la colonne du milieu (et aux rubriques 35-36 un choix de réponses est proposé entre parenthèses)
  - **un "bip sonore" vous indiquera que c'est-à-vous de parler: lisez alors le texte de la rubrique concernée**
- **avant de faire le N° gratuit** 155 28 14, parcourez une première fois les 38 rubriques de ce document
- si vous souhaitez une information complémentaire, vous pouvez téléphoner au N° 031 / 338 64 57 (MM. Mury ou van Kommer pendant les heures de bureau)

Ci-dessous commence votre tâche effective

N°	**Explication / texte annoncé par l'automate**	**Ce que vous devez lire (juste après le "Bip sonore")**
1	Lisez le nombre :	*13946*
2	Lisez la phrase :	*La fermière élève des oies.*
3	Lisez le(s) mot(s) :	*oui*
4	Lisez la quantité :	*4 032 mètres*
5	Lisez le numéro de carte de crédit :	*6493 3578 0602 1217*
6	Lisez la phrase :	*Déchausse-toi avant d'entrer.*
7	Lisez le nom puis épelez-le :	*Bexen* *B e x e n*
8	Lisez la somme :	*28 011 Dollars U S*

9	Lisez la phrase :	*L'épervier a saisi un moineau dans ses serres.*
10	Lisez le(s) mot(s) :	*informations consommateurs*
11	Lisez la phrase 11 :	*Le coupable a fait des aveux spontanés.*
12	Lisez la phrase 12 :	*Il nous a noyés dans des explications invraisemblables.*
13	Lisez l'heure :	*22:23*
14	Lisez le mot :	*Nicolas*
15	Lisez la phrase :	*Ce musicien a du génie.*
16	Lisez le(s) mot(s) :	*réservation*
17	Lisez la somme :	*14 811 Lires*
18	Lisez la date :	*Mercredi 4 octobre 2022*
19	Lisez la phrase :	*Les hameçons, les filets sont des engins de pêche.*
20	Lisez le nombre :	*12 028,711*
21	Lisez le(s) mot(s) :	*Le temps*
22	Lisez le nom puis épelez-le :	*Valaisan* *V a l a i s a n*
23	Lisez la phrase :	*Je sens une douleur à l'épine dorsale.*
24	Lisez le(s) mot(s) :	*l'heure*
25	Lisez un à un les chiffres et signes :	** 0 7 9 5 4*
26	Lisez le nom puis épelez-le :	*Scacco* *S c a c c o*
27	Lisez la phrase 27 :	*Cette maison a été le théâtre d'un crime.*
28	Lisez la phrase 28 :	*Il communie tous les dimanches.*

feuille d'appel n°13946

29	Lisez le nom de ville :	*schwytz*
30	Veuillez poursuivre en donnant un numéro de téléphone que vous connaissez par cœur :	.....
31	Indiquez votre langue maternelle :	.....
32	Répondez par Oui ou par Non à cette question : êtes-vous de sexe féminin ?	.....
33	Indiquez votre date de naissance :	.....
34	Indiquez la ville dans laquelle vous avez commencé votre formation scolaire, à savoir la 1ère année d'école primaire :	.....
35	Indiquez le niveau final de votre formation scolaire (école primaire, école professionnelle ou école supérieure) :	.....
36	Indiquez le type de téléphone que vous utilisez en ce moment précis (standard Tritel, téléphone sans fil, Natel C, Natel D, un appareil importé, ou une cabine téléphonique publique) :	.....
37	Veuillez maintenant faire comme si vous étiez en ligne avec le 111 ...pour demander le n° de téléphone de la personne imaginaire dont les coordonnées se trouvent en face :	*...PIPOZ-PILET NICOLE* *BOIS-NOIR 18* *CERNIER...*
38	Votre éventuel commentaire sur l'ensemble de cet entretien :	.....

B "Appels 111" Database

The data considered in the INSPECT project consist in a subset of the original Swiss French PolyPhone database restricted to the items related to the 111 service calls.

This database contains 4293 recordings, each consisting of 2 files: an ASCII file (`.txt`) corresponding to the initial prompt and address request, and a data file (`.alw`) of the recorded speech in a-law format along with a NIST header containing further information (among which the transcription of the speaker request) [1, 9].

Here is an example of the ASCII file:

```
Veuillez maintenant faire comme si vous étiez en ligne avec le 111
...pour demander le no de téléphone de la personne imaginaire dont
les coordonnées se trouvent ci-dessous:
BUCHWALDER-CUENAT PASCALE
MOULIN D'ALLERES
SEMBRANCHER
```

and of the corresponding NIST header in the data file:

```
NIST_1A
1024
database_id -s22 Swiss_French_Polyphone
recording_site -s17 Swiss_Telecom_PTT
sheet_id -i 13952
prompt -s168 Veuillez maintenant faire comme si vous étiez en ligne avec le
111 pour demander le no de téléphone de la personne imaginaire dont
les coordonnées se trouvent en face.
text_transcription -s114 Bonjour mademoiselle, j'aimerais le numéro de
téléphone de Buchwalder-Cuenat Pascale Moulin d'Allères Sembrancher
speaking_mode -s11 spontaneous
sample_begin -r 0.200000
sample_end -r 10.151875
sample_count -i 82815
sample_n_bytes -i 1
channel_count -i 1
sample_coding -s4 alaw
sample_rate -i 8000
sample_byte_format -s1 1
sample_sig_bits -i 8
sample_checksum -i 62637
database_version -s3 1.0
utterance_id -s8 m3212o06
end_head
```

The database is organised in blocks of less than 100 recordings. `block00` to `block24` contain 2407 female speaker recordings and `block30` to `block49` 1886 male speaker recordings.

Files are named according to their block number and number of the subdirectory in the block, as well as the sex of the speaker (for instance `f1234o06.txt` concerns a female speaker stored in `block12` subdirectory 34).

C Instances of Errors Encountered in the "Appels 111" Text Data

C.1 Errors in Address Prompt Files

1. Some address prompt files are empty.

2. As it can be seen in the following examples, there are many different formats for the not empty files, which made difficult to handle them:

   - ```
     Veuillez maintenant faire comme si vous étiez en ligne avec le 111
     ...pour demander le no de téléphone de la personne imaginaire dont
     les coordonnées se trouvent ci-dessous:
      MOTTAZ MONIQUE
      rue du PRINTEMPS 4
      SAIGNELEGIER
     ```
   - ```
     Veuillez maintenant faire comme si vous étiez en ligne avec le 111
     ...pour demander le no de téléphone de la personne imaginaire dont
     les coordonnées se trouvent ci-dessous:
      GROSJEAN MARTINE

      VAULION
     ```
   - ```
     Simulez une demande de renseignement au 111 concernant la personne
     correspondant aux coordonnées fictives ci-dessous:
     VOCAT STEPHANE
     RTE DE FENIL 40
      ST-LEGIER-CHIESAZ
     ```
   - ```
     Simulez une demande de renseignement au 111 concernant la personne
     correspondant aux coordonnées fictives ci-dessous:
     PERRIN-REYMOND FRANCOISE
     LES HAUDERES
     ```
   - ```
     Simulez une demande de renseignement au 111 concernant la personne
     correspondant aux coordonnées fictives ci-dessous:
     ```
   - ```
     Prononcez une demande ,au 111, du numéro de téléphone de DONZE GEORGES
     habitant  à  SENARCLENS.            (votre réponse)
     ```
   - ```
     Prononcez une demande ,au 111, du numéro de téléphone de ETHENOZ-LAVANCHY
     MONIQUE habitant route de COTEAU 36 à FRIBOURG.            (votre réponse)
     ```

   For instance, it was necessary to transform prompts similar to the last one into:

   ```
   Prononcez une demande au 111, du numéro de téléphone de:
    ETHENOZ-LAVANCHY MONIQUE
    route de COTEAU 36
    FRIBOURG
   ```

   Notice that:

   1. several address fields may be missing;

   2. the proper nouns in the available fields are written in capital letters and without accent informations.

Several kinds of error were found in the addresses. Here are some examples of improvements performed on the address fields:
   ```

- Street introduction expressions and street numbers were uncapitalized (e.g. “RUE DU” → “rue du”, “4 A” → “4 a”).
- Isolated cases of proper nouns in small letters were capitalized (e.g. “Favre Nicole” → “FAVRE NICOLE”, “SOUS la FORET” → “SOUS LA FORET”).
- All kinds of abbreviation were spelled out (e.g. “ch.” → “chemin”, “MAR.-ANTOINET.” → “MARIE-ANTOINETTE”, “LES 3 SAPINS” → “LES TROIS SAPINS”, “VD” → “VAUD”, “GD-ST-BERNARD” → “GRAND-SAINT-BERNARD”, “ROCHE S/FORON” → “ROCHE-SUR-FORON”).
- Characters like “””, “)”, “(”, “/” were replaced by a space character.
- Very many spelling errors were corrected (e.g. “GRAND’RUE” → “GRAND-RUE”, “LE COLLGE” → “LE COLLEGE”, “AVRY - SUR -MATRAN” → “AVRY-SUR-MATRAN”).

C.2 Errors in Speaker Request Orthographic Transcriptions

Examples of Transcription Errors

(f0155o06) Non défini Bircher-Schmid Elisabeth Vie-du-Haut trente à Buix
· insertion: “Non défini Bircher-Schmid” → “Bircher-Schmid”

(m3353o06) Bonjour j’aimerais le numéro de téléphone d’une abonnée il s’agit de madame i n e s i n e s Solioz-Walpen je vous épelle heinsolioz trait d’union w a l p e n et qui habite à la route Te du Curson dix huit
· insertions and substitution: “madame i n e s i n e s” → “madame Ines i n e s”
· substitution and deletions: “heinsolioz” → “eh s o l i o z”

(f0080o06) Oui bonjour j’aimerais le numéro de téléphone de Dominicis Françoise qui habite le le Tombex à Concise
· insertion: “habite le le Tombex” → “habite le Tombex”

(f0024o06) Bonjour j’aurais aimé connaître le numéro de téléphone de monsieur Oyvaert Steve je vous l’épelle o y v a e r t Steve habitant le Grand Clos je ne connais pas le numéro postal Pailly p a i l l y merci
· substitution: “p a i l l y” → “p a i deux l y”

(m4935o06) Bonjour j’aurais le numéro de téléphone de Briguet-Duverney Hélène les Marécottes merci
· deletion: “j’aurais le numéro” → “j’aurais aimé le numéro”

(f0167o06) J’aimerais bien avoir le numéro de téléphone à Semsales s’il vous plaît de monsieur Claude Wuillemin avec double w et la seule indication que j’ai c’est poste
· insertion: “avec double w” → “avec w”

(f0200o06) Oui bonjour j’aimerais avoir le numéro de téléphone de madame Anna Planchamp-Ruff à Boécourt l’adresse c’est le Chenois quarante huit r je vous épelle p l a c h a m p tiret r u f f merci
· deletion: “p l a c h a m p” → “p l a n c h a m p”

(m3539o06) Musy Valérie chemin Édouard Olivier trente six à Thonex
· substitution and insertion: “Édouard Olivier” → “Édouard-Olivet”

(f0378o06) Bonjour mademoiselle pourriez-vous me donner s’il vous plaît le numéro de la personne le numéro de téléphone pardon de madame Devaud tiret g Achet Marylise la Coie septante cinq c

Vendlincourt
· insertion and substitution: “g Achet” → “Gachet”

(f0455o06) Bonjour Gisèle Roduit Lausanne j’aimerais le numéro de téléphone de monsieur Félix Rime r i me chemin des Anémones huit à Saignelégier
· substitution and deletion: “r i me” → “r i m e”

(f0585o06) Oui bonjour je [\hésitation désirerais] le numéro de téléphone monsieur Pointet Gilbert à Panex merci
· deletion: “téléphone monsieur” → “téléphone de monsieur”

(f0817o06) Bonjour madame, j’aimerais connaître le numéro de téléphone de Veuthey-Aymon Marina en Désovuy euh! et des Pâsqier trait d’union Montbarry avec deux r
· substitutions: “Désovuy” → “Désovy”, “des Pâsqier” → “de Pâquier”

(f0887o06) Oui bonjour mademoiselle pourrais-je avoir le numéro de téléphone de mademoiselle Decosterd Nathalie Nathalie Prez-vers-Noréaz merci
· substitution: “Nathalie” → “Isabelle”

(f1323o06) Auriez-vous l’amabilité de me donner le numéro de téléphone de monsieur Hurni Walter Hurni h comme Henri u comme [\hésitation Ursule] r comme n comme Nicole i comme Ida Walter w comme William a comme Anna l comme [\hésitation Léon] t comme Thérèse
· deletion: “r comme n comme Nicole” → “r comme Robert n comme Nicole”

(f1647o06) Oui bonjour madame est-ce que je pourrais avoir le numéro de téléphone de madame Lienhard-Imhof Hélène à Yvonand l i e n h a r d - i
· wrong annotation modality: “l i e n h a r d - i” → “l i e n h a r d trait d’union i”

Undocumented, Not Uniform, and Not Objective Specific Information Annotations

· undocumented and not uniform annotations of “inintelligible”:
(f0280o06) Bonjour Catherine [\inintelligible Chvhorer] à Crissier j’aimerais savoir le numéro de téléphone de monsieur Billod Louis à Martigny-Croix s’il vous plaît
(f0297o06) Bonsoir madame auriez-vous l’obligeance de me donner le numéro de téléphone de Kurt Markus [\inintelligible] à Vétroz
(f0607o06) Oui bonjour c’est [\inintelligible .....] j’aimerais avoir un renseignement s’il vous plaît sur monsieur Piquilloud Henri-Daniel qui habite Haute-Nendaz j’aimerais son numéro de téléphone s’il vous plaît
(m4729o06) Euh je j’aimerais j’aimerais [\inintelligible mademoiselle] s’il vous plaît le le numéro de téléphone de monsieur Mettan Mettan Patrice rue des Vignerons à Vétroz s’il vous plaît merci

· undocumented and not uniform annotations of “hésitation”:
(f0007o06) Bonjour mademoiselle est-ce que vous pourriez me donner le numéro de téléphone qui correspond à Bernadette Lugon-Revaz la [\hésitation Chaussiaz] d un Echallens s’il vous plaît merci
(f0023o06) Hintermann-Grandjean [\hésitation Madeleine] Madeleine les Haudères
(f0037o06) Je désire [\hésitation euh] obtenir le numéro de téléphone de monsieur Chassot Fernand c h a deux s o t rue de la poste à Fully merci
(f0669o06) Oui bonjour s’il vous plaît j’aimerais un numéro de téléphone correspondant [\hésitation au numéro au nom] de monsieur Montandon Claude-Alain Avry-devant-Poste s’il vous plaît
(f0681o06) Oui bonjour madame euh c’est mademoiselle Sutter j’aimerais euh avoir le numéro de téléphone de madame Tornare-Rime Marie-Thérèse euh elle habite au chemin de Revatte quatre à Courtetelle

(f0843o06) Oui bonjour madame est-ce-que je pourrais s'il vous plaît avoir le numéro de téléphone de monsieur Renaud Henri-Louis case postale [\hésitation] cent trente cinq à Saxon
(f0980o06) J'aimerais savoir le numéro de téléphone de monsieur [hésitation Roethlisberger] de madame Roethlisberger-Haeberli Vivian grand-record vingt et un à Echallens

· undocumented and not uniform annotations of "prononciation bizarre"
(f0189o06) Bonjour j'aimerais avoir l'adresse de madame [\prononciation bizarre Monney Carine] à [\prononciation bizarre Biaufond] merci
(f0219o06) Oui bonjour j'aimerais le numéro et l'adresse de Descloux-Moullet Laurence à Frenières-sur-[\prononciation bizarre Bex] s'il vous plaît
(f0707o06) Bonjour, je voudrais savoir un numéro de téléphone [\hésitation de] Pichi-Ponto Sueli rue du Midi cinq [\prononciation bizarre [\hésitation Vallorbe]]
(f2171o06) Alors Victorio Maria de [prononciation bizarre] le grillon Leysin

(f0442o06) Bonjour je [\inintelligible] le numéro de Depaillens Roland [\hésitation Sornard] Haute-Nendaz
· "[\inintelligible]" → "désirais"
· "Depaillens" → "Depallens"
· "[\hésitation Sornard]": what does it mean?

(f1835o06) [\prononciation bizarre Veillez] me donner le numéro de téléphone de [\prononciation bizarre monsieur] Édouard Inderbitzin nom de famille Inderbitzin i n d e r b i t z i n il se trouve à Saint-Aubin dans le canton de Fribourg
· "[\prononciation bizarre Veillez]" → "Veuillez"
· "[\prononciation bizarre monsieur]" → "monsieur"

(m3065o06) [\hésitation numéro] de téléphone il s'agirait de monsieur Tercier Cyril au boulevard Saint-Germain à Savièse
· "[\hésitation numéro] de téléphone" → "Pour un téléphone"

(f0094o06) Auriez-vous la gentillesse de me communiquer le numéro de téléphone de monsieur Crettaz-[\prononciation bizarre Voide] Marie-Laurentin les Charbonnières
· "[\prononciation bizarre Voide]" → "Voïdé"

Examples of Spelling Errors

(m3909o06) Pourriez-vous me comuniquer le numéro de téléphone [\hésitation euh] de monsieur Rotenbuehler Albert route de Gy cent vingt six
· "comuniquer" → "communiquer"

(f2134o06) Bonjour mademoiselle veillez avoir l'obligeance de me donner le numéro de téléphone de mademoiselle Delabays Christiane les Cullayes [\hésitation j'épelle] le nom Delabays d comme Daniel é comme Émile l Lily a Astrid b comme Berthe a comme Astrid
· "veillez" → "veuillez"

(m4511o06) Leuba Willy place du vallon dix huit Lausanne je vais épeller l e e u b a v double v i l l y p l place d u v a l l
· "épeller" → "épeler"

(f1071o06) Bonjour pourriez-vous m'indiquer les coordonées téléphoniques de monsieur Fasel Jean-Jacques chemin de la colline trois à Orbe s'il vous plaît
· "coordonées" → "coordonnées"

(m4929o06) Bonjour madame auriez-vous l'obligence de me donner le numéro de téléphone de madame Juvet-Jeanneret Nadine qui habite chalet mirador à Château-d'Oex s'il vous plaît
· "obligence" → "obligeance"
(m3783o06) Bonsoir, j'aimerais connaître les coordonnées de monsieur Allemand Walter la nouvelle poste à Souboz-les- Écorches merci
· "Souboz-les- Écorches" → "Souboz-les-Écorches"

(f0000o06) Bonjour j'aimerais un numéro de téléphone à Saignelegier c'est Mottaz m o deux ta z Monique rue du printemps numéro quatre
· "ta" → "t a"

(f0572o06) Oui bonjour c'est Marianne Geiger qui appelle et ce que je pourrais avoir le numéro de téléphone de monsieur Kaenzig Daniel qui habite à la Lechire à [\hésitation Pontales] Pompaples pardon merci beaucoup
· "et ce que" → "est-ce que"

(m4334o06) Bonsoir mademoiselle je désirerais connaître ladresse de monsieur Schwitzguebel Georges la Léchère [\hésitation Villarlod]
· "ladresse" → "l'adresse"

(f0617o06) Bonjour mademoiselle j'aimerais s'il vous plaît le numéro de monsieur Grand René â Chavannes-de-Bogis
· "â" → "à"

(f0805o06) Bonjour j'aimerais avoir le numéro de téléphone de Bonfils-Wascher Stéphanie s'il vous plaît qui habite à l rue Billard Gimel
· "l" → "la"

(f0983o06) Bonjour veuillez me donner le numéro de téléphone de madame Aubry Aurélie à Chambydans le canton de Vaud s'il vous plaît merci
· "Chambydans" → "Chamby dans"

(m4674o06) Bonjour, je voudrais l'adressse monsieur Comby Frédéric au collège de Panex
· "adressse" → "adresse"

Examples of Syntactic Errors

(f0029o06) Bonjour je souhaiterai le numéro téléphonique de monsieur Genet Daniel route de Vétroz Conthey
· "je souhaiterai" → "je souhaiterais"

(f0320o06) Alors bonjour mademoiselle est-ce-que je peu avoir le numéro de monsieur Gendre Marcel le Relais à Bière merci beaucoup · "est-ce-que je peu" → "est-ce que je peux"

(f1460o06) Oui bonsoir, excusez moi, est ce que je pourrai avoir le numéro de monsieur Christ Jean-Pierre euh les Bambins en Collognes
· "est ce que" → "est-ce que"
· "je pourrai" → "je pourrais"

(f1395o06) Bonjour, je voudrai un renseignement pour savoir le numéro de téléphone de madame Clémence Dubois Germaine avenue du Moulin dix sept a Morges, merci

· "je voudrai" → "je voudrais"
· "a Morges" → "à Morges"

(f1614o06) Est-ce-qu'il serai possible d'avoir le numéro de Blanc Michel Fernand le Sepey
· "Est-ce-qu'il serai" → "Est-ce qu'il serait"

(f0975o06) Bonjour mademoiselle j'aimerais le numéro de téléphone de madame Groux-Fazan Anne-Lise vous voulez que je vous l'épelle non ah! très bien ça m'arrange j'ai déjà [\hésitation épeler] deux fois alors elle habite à Villarey [\hésitation euh] Cousset [\hésitation euh] je pense c'est soit dans le canton de Vaud soit dans
· "j'ai déjà [\hésitation épeler]" → "j'ai déjà épelé"
· "c'est soit" → "ce soit"

Errors in the Use of Accents, Capitalization and Punctuations

(f0011o06) Bonjour est-ce que vous pourriez me donner le numéro de téléphone de madame Martine Gerber-Burri celà s'épelle g e r b e r trait d'union b u deux r i elle habite probablement à Dommartin merci
· accents: "celà" → "cela"
· no punctuations

(f1984o06) Bonjour mademoiselle j'aimerais le numéro de madame Duperrut-Agassis Yvette sommtres dix sept Saignelegier
· capitalization: "sommtres" → "Sommtres"

(m4260o06) Oui bonjour j'aimerais avoir le numéro de téléphone de monsieur Georges Pierre-André rue de Madretsch cent vingt deux à Biel (bienne
· "("?
· capitalization: "bienne" → "Bienne"

(f0986o06) Mademoiselle bonjour, pourrais-je avoir le numéro le téléphone de madame Marie-Louise Paukovics p a u k o v i c s l'adresse est: la Ferme et le lieux : les Bioux
· not uniform punctuation

(m3259o06) Bonjour mademoiselle est-ce que je pourrai avoir le numéro de téléphone de madame Coudray-Doerfliger Yvette à la route de Bernex trois cent dix sept Bis à Bernex
· capitalization: "sept Bis" → "sept bis"

(f0086o06) Bonjour, c'est de Martigny qu'on téléphone. j'aurais aimé savoir où habite cette personne. Je vous épelle Fawer Bluette f a [\prononciation bizarre w] e r Bluette avec deux t rue de la Paix dix huit à Gland
· not uniform punctuation
· capitalization: "qu'on téléphone. j'aurais" → "qu'on téléphone. J'aurais"

(m3542o06) J'aimerais le numéro de téléphone de monsieur Borruat-Oeuvray Denise enfin madame plutôt qui habite au chemin Chasseral vingt huit à Saignelégier
(m4371o06) Alors bonjour pourriez-vous m'indiquer le numéro de téléphone de monsieur Piot Jean-Daniel chemin de Chasseral vingt huit à Saignelegier s'il vous plaît merci
· accents: "Saignelégier" or "Saignelegier"?

(f1049o06) [\hésitation euh] bonjour mademoiselle pourriez-vous me donner le numéro de téléphone de madame Glassey-Mariéthoz Agnès route de Lausanne seize à Mont-sur-Lausanne s'il vous plaît

(f1339o06) Vous me pouvez donner le numéro de téléphone de la famille Glassey-[\prononciation bizarre Mariethoz] Agnes route de Lausanne seize à Mont-sur-Lausanne
· accents: “Glassey-Mariéthoz” or “Glassey-Mariethoz”?, “Agnès” or “Agnes”?

Examples of Abbreviations

(f0739o06) Bonjour mademoiselle, seriez-vous assez aimable pour me donner le numéro de téléphone de madame Demierre-Guillaume Yolande Ch de Trembley douze à Prangins
· “Ch” → “chemin”

(f1520o06) Bonjour mademoiselle puis-je avoir le .numéro de téléphone de mademoiselle Gallay-Vial Mireille Ch. de l’envoi treize à Sion s’il vous plaît
· “Ch.” → “chemin”

D Phone Models

The file that defines the phone models specifies for each phoneme the number of states (including entry and exit null states), the model topology, the transition probabilities and the output probability distributions associated with each state.

The format of the file is as follows. The first line consists of the string "PHONE", and the second line contains an integer giving the number of phone models. The remainder of the file contains the description of each phone model. Within a phone model 0 indexes the entry null state, 1 indexes the exit null state and 2 onwards index the real emitting states. The format for a phone model is:

```
<id> <number of states> <label>
-1 -2 <probid-1> <probid-2> ...
<from state> <#out-trans> <to state> <prob> <to state> <prob> ...
<from state> <#out-trans> <to state> <prob> <to state> <prob> ...
...
```

Where -1 and -2 represent dummy phone numbers for the entry and exit states, and <probid-n> represents the element of the acoustic probability vector corresponding to that state (one for each state). The number of integers on this line equals the number of states. The remaining lines specify the transition probabilities giving the transitions out of each state.

Here are the phone models definitions used in our baseline system.

```
0 3 interword-pause
-1 -2 0
0 2 2 0.5 1 0.5
1 0
2 2 2 0.9 1 0.1

1 6 a
-1 -2 1 1 1 1
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

2 6 E
-1 -2 2 2 2 2
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

3 6 o
-1 -2 3 3 3 3
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

4 8 a~
-1 -2 4 4 4 4 4 4
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 7 0.5
7 2 7 0.5 1 0.5

5 6 b
-1 -2 5 5 5 5
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

6 10 S
-1 -2 6 6 6 6 6 6 6 6
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 7 0.5
7 2 7 0.5 8 0.5
8 2 8 0.5 9 0.5
9 2 9 0.5 1 0.5

14 7 i
-1 -2 14 14 14 14 14
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 1 0.5

15 7 e~
-1 -2 15 15 15 15 15
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 1 0.5

16 8 Z
-1 -2 16 16 16 16 16 16
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 7 0.5
7 2 7 0.5 1 0.5

17 6 k
-1 -2 17 17 17 17
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

18 5 l
-1 -2 18 18 18
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 1 0.5

19 6 m
-1 -2 19 19 19 19
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

20 6 n
-1 -2 20 20 20 20
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

27 7 s
-1 -2 27 27 27 27 27
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 1 0.5

28 6 t
-1 -2 28 28 28 28
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

29 6 y
-1 -2 29 29 29 29
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

30 6 H
-1 -2 30 30 30 30
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

31 8 9~
-1 -2 31 31 31 31 31 31
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 7 0.5
7 2 7 0.5 1 0.5

32 5 v
-1 -2 32 32 32
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 1 0.5

33 7 w
-1 -2 33 33 33 33 33
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 1 0.5
```

```
7 5 d
-1 -2 7 7 7
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 1 0.5

8 5 @
-1 -2 8 8 8
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 1 0.5

9 7 2
-1 -2 9 9 9 9 9
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 1 0.5

10 8 e
-1 -2 10 10 10 10 10 10
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 7 0.5
7 2 7 0.5 1 0.5

11 7 f
-1 -2 11 11 11 11 11
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 1 0.5

12 6 g
-1 -2 12 12 12 12
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

13 9 J
-1 -2 13 13 13 13 13 13 13
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 7 0.5
7 2 7 0.5 8 0.5
8 2 8 0.5 1 0.5

21 6 O
-1 -2 21 21 21 21
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

22 7 u
-1 -2 22 22 22 22 22
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 1 0.5

23 7 9
-1 -2 23 23 23 23 23
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 1 0.5

24 8 o~
-1 -2 24 24 24 24 24 24
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 6 0.5
6 2 6 0.5 7 0.5
7 2 7 0.5 1 0.5

25 6 p
-1 -2 25 25 25 25
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

26 5 R
-1 -2 26 26 26
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 1 0.5

34 6 j
-1 -2 34 34 34 34
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 5 0.5
5 2 5 0.5 1 0.5

35 5 z
-1 -2 35 35 35
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 4 0.5
4 2 4 0.5 1 0.5

36 4 N
-1 -2 20 12
0 1 2 1.0
1 0
2 2 2 0.5 3 0.5
3 2 3 0.5 1 0.5

37 3 6
-1 -2 8
0 1 2 1.0
1 0
2 2 2 0.5 1 0.5

38 3 #0
-1 -2 0
0 1 2 1.0
1 0
2 2 2 0.9 1 0.1

39 3 #1
-1 -2 0
0 1 2 1.0
1 0
2 2 2 0.9 1 0.1

40 3 #2
-1 -2 0
0 1 2 1.0
1 0
2 2 2 0.9 1 0.1

41 3 #3
-1 -2 0
0 2 2 0.5 1 0.5
1 0
2 2 2 0.9 1 0.1
```

References

[1] `readme.txt` on Polyphone Suisse Romand Appels 111 CD-ROM version 1.0a.

[2] J. M. Andersen, G. Caloz, and H. Bourlard. Swisscom "Advanced Vocal Interfaces Services" project – Technical report for 1997. Technical Report COM-97-06, IDIAP, December 1997.

[3] Base de Données Lexicales du français écrit et parlé. http://www.irit.fr/ACTIVITES/EQ_IHMPT/bdlex.html, 1987.

[4] Giulia Bernardis and Hervé Bourlard. Confidence measures in hybrid HMM/ANN speech recognition. In *Proceedings of Workshop on Text, Speech and Dialog (TSD'98) Brno, Czech Republic*, pages 159–164, September 1998.

[5] Giulia Bernardis and Hervé Bourlard. Improving posterior based confidence measures in hybrid HMM/ANN speech recognition systems. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'98) Sydney, Australia*, pages 775–778, 1998.

[6] J.-C. Chappelier and M. Rajman. A generalized CYK algorithm for parsing stochastic CFG. In *TAPD'98 Workshop*, pages 133–137, Paris (France), 1998.

[7] J.-C. Chappelier and M. Rajman. A practical bottom-up algorithm for on-line parsing with stochastic context-free grammars. Technical Report 98-284, DI – EPFL, July 1998.

[8] J.-C. Chappelier, M. Rajman, P. Bouillon, S. Armstrong, V. Pallotta, and A. Ballim. ISIS project – Final report. ISSCO and EPFL, September 1999.

[9] G. Chollet, J.-L. Cochard, A. Constantinescu, C. Jaboulet, and Ph. Langlais. Swiss French PolyPhone and PolyVar: Telephone speech databases to model inter- and intra-speaker variability. Technical Report RR-96-01, IDIAP, April 1996.

[10] Slava M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages 400–401, March 1987.

[11] S. Renals. Noway's manual page. University of Cambridge. http://www.clsp.jhu.edu/ws96/ris/man/noway.doc, 1994.

[12] S. Renals and M. Hochberg. Efficient evaluation of the LVCSR search space using the NOWAY decoder. In *Proceedings of ICASSP'96*, pages 149–152, Atlanta, May 1996.

[13] R. Rosenfeld. Carnegie Mellon University Statistical Language Modeling Toolkit. http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/11761-s97/toolkit/CMU-Cam_Toolkit_v2.tar.gz, 1994.

[14] Speech Assessment Methods Phonetic Alphabet. http://www.phon.ucl.ac.uk/home/sampa/home.htm, 1989.

[15] Speech Training and Recognition Unified Tool. http://tcts.fpms.ac.be/speech/strut.html, 1996.