Development of Acoustic and Linguistic Resources for Research and Evaluation in Interactive Vocal Information Servers

Giulia Bernardis[1] [2], Hervé Bourlard[1] [2], Martin Rajman[1], Jean-Cédric Chappelier[1]

[1] Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
[2] Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), Martigny, Switzerland
giulia@idiap.ch

Abstract

This paper describes the setting up of a resource database for research and evaluation in the domain of interactive vocal information servers. All this resource development work took place in a research project aiming at the development of an advanced speech recognition system for the automatic processing of telephone directory requests and was performed on the basis of the Swiss-French Polyphone database (collected in the framework of the European SpeechDat project). Due to the unavailability of a properly orthographically transcribed, consistently labeled and tagged database of unconstrained speech (together with its associated lexicon) for the targeted area, we first concentrated on the annotation and structuration of the spoken requests data in order to make it profitable for lexical and linguistic modeling and for the evaluation of recognition results. A baseline speech recognition system was then trained on the newly developed resources and tested. Preliminary recognition experiments showed a relative improvement of 46% for the Word Error Rate (WER) compared to the results previously obtained with a baseline system very similar but working on the unconsistent natural speech database that was originally available.

1. Introduction

In this paper, we will present the development of a resource database for research and evaluation in the domain of dialog-based Interactive Voice Response (IVR) systems. Since a correctly orthographically transcribed, consistently labeled and tagged database of unconstrained speech was not available for the targeted area, we concentrated on the annotation and structuration of a natural spoken requests database in order to make it profitable for lexicon and language modeling and for the evaluation of recognition results necessary for the assessment of our current speech recognition systems.

This resource set-up work took place in the framework of a research project aiming at the development of an advanced speech recognition system for the automatic processing of telephone directory requests. This project, referred to as INSPECT (INtegrating SPEech, acoustic and linguistic, ConsTraints for enhanced recognition systems), is a multi-faceted project involving (1) text processing (orthographically transcribing, labeling and tagging) of a large database of telephone-based natural voice requests, (2) development of robust acoustic models, (3) integrating advanced natural language constraints, (4) detecting and dealing with a large number of out-of-vocabulary words (proper nouns), and (5) testing of the resulting system on natural queries.

All the described work was performed on the basis of the **Swiss-French Polyphone** database (Chollet et al., 1996; Andersen et al., 1997), collected in the framework of the European SpeechDat project. Swiss-French Polyphone contains prompted (read) speech with a good phonetic coverage (**Polyphone** database) and (simulated) natural requests to information service (**"Appels 111"**) in Swiss French.

The paper first describes in details the database available at the beginning of the project, then focuses on the development of usable resources from this database.

In standard speech recognition systems, lexicon models are usually a priori defined on the basis of dictionaries containing all possible words with their phonetic transcription. In the considered case however, not all the words used by users for their requests were known (especially proper nouns). 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 the considered case, the Polyphone database (containing prompted speech) was not representative at all of the targeted application (telephone directory requests) and the "Appels 111" database had to be used instead.

Consequently, processing of the corresponding natural spoken requests (including numerous peculiarities that will also be briefly discussed in the paper) was required to come by a careful orthographic transcription, consistent labeling and tagging. The main goal was to get good enough text data to be suitable for lexical and linguistic modeling as well as for the evaluation of recognition results.

Finally, the setting up of a baseline speech recognition system on the newly developed resources (properly orthographically transcribed, consistently labeled and tagged) and the results of initial recognition experiments are described.

2. Swiss-French Polyphone and "Appels 111" Databases

The **Swiss-French Polyphone** database (Chollet et al., 1996; Andersen et al., 1997) 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 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 the Polyphone database);
- simulate a spontaneous query to the telephone directory (the name and the address of the queried person were given by the system), i.e., simulate a "111 information service call".

In particular, the subset of the Swiss-French Polyphone database consisting of the items related to the

"111 service calls" represents the application framework for our research and in the following we will 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 corresponding to the initial prompt and address request, and a data file containing the recorded speech in a-law format and the transcription of the speaker request.

Below is given an example of the initial prompt and address as well as of the orthographic transcription of the consequent request uttered by the speaker.

Prompt and address:

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

NEUKOMM ALBERT
APPLES

Transcription:

Oui bonjour j'aurais voulu savoir le numéro de téléphone de monsieur Neukomm Albert qui habite [\prononciation bizarre Apples] s'il vous plaît

Unfortunately, the text data available for the "Appels 111" database was not tagged, not always consistently labeled and sometimes not correctly orthographically transcribed.

We decided to use this database despite those serious drawbacks because it was the only unconstrained speech database currently available in Swiss French.

Consequently, as described in the next section, a significant amount of time was devoted to the proper processing of the "Appels 111" in order to create a good basis for the lexical and linguistic modeling as well as for the speech recognition system evaluation.

3. 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 were not tagged and very often not consistently labeled (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] Cousset [\hésitation euh] je pense c'est soit dans le canton de Vaud soit dans*", where the "*[\hésitation …]* " label is used in an inconsistent way), nor properly transcribed (e.g. available transcription: "*Bonjour j'aurais le numéro téléphone de Briguet-Duverney Hélène les Marécottes merci*"; correct transcription: "*Bonjour j'aurais aimé le numéro de téléphone de Briguet-Duverney Hélène Les Marécottes merci*").

Proper orthographic transcription and tagging of our database is however very important to the development and testing of our recognition system since "Appels 111" represents the only resource of unconstrained speech currently available in Swiss French and therefore will be used to:

- 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;
- automatically model syntactic constraints (grammar model) in terms of this lexicon on the basis of text data that reflects at best the conditions of use of the recognizer.

3.1. Address Prompt Reformatting and Information Extraction

After correction of miscellaneous errors (see (Bernardis et al., 1999) for details), the last three lines corresponding to the address of the queried person were extracted from the address prompt files of the "Appels 111" database and processed with the following heuristic:

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

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. 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 (see (Bernardis et al., 1999)), 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.
- “Noise” represented by non documented, non uniform annotations (e.g. “*[hésitation]*”, **“***[\inintelligible]*”, “*[\prononciation bizarre]*”, …) 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*”, …, although the word “*aujourd'hui*” was not split, …).
- Capital letters were lowercased (e.g. “*Je*” → “*je*”, “*Louis*” → “*louis*”, “*Camping TCS*” → “*camping tcs*”, …).
- Abbreviations and special signs were spelled out (e.g. “*ch*” → “*chemin*”, …).

Next, from the transcriptions a list of the words other than proper nouns was extracted (*general vocabulary list*) and, using the facilities of the Syntactical Language Processing ToolKit developed at the EPFL's Artificial Intelligence Laboratory (Chappelier & Rajman, 1998), several iterations through the following steps were done:

1. Creation of a lexicon from the *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 and written, as consequence, in capital letters.
2. String correction of orthographic transcription sentences based on this lexicon with :
   - very low penalty (= 0.005) for accent corrections and lowercase ↔ uppercase transformations;
   - maximal edit distance allowed for the lexical research D*max* = 0.5;
   - research mode enabling all the solutions at distance D <= D*max* (which corresponds to at most 100 accent/case corrections).
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.

Below is shown an example text segment of the speaker request orthographic transcriptions as available initially:

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

and after our text processing:

<s> bonjour mademoiselle puis-je avoir le numéro de téléphone de mademoiselle GALLAY-VIAL MIREILLE chemin de l' ENVOI treize à SION s' il vous plaît </s>

3.3. Text Data Statistics

The “Appels 111” database includes 4,293 orthographic transcriptions of spoken requests, but actually a few of them are cut off and one is empty.

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

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.

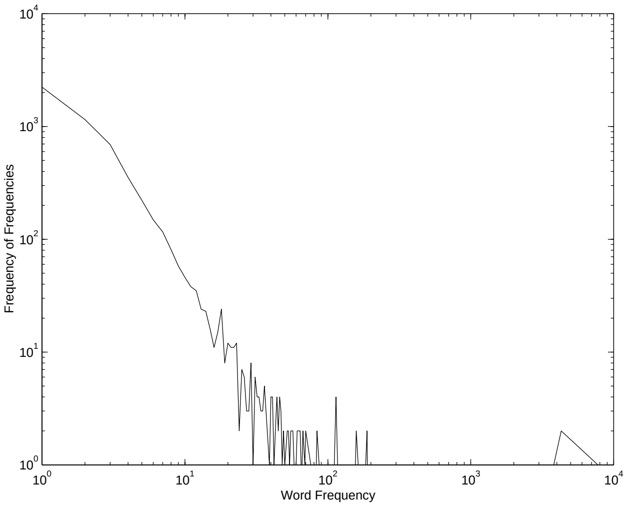

Figure 1: Number of occurrences of word frequencies (in log-log scale)

3.4. Named Entity Tagging

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

The accurate identification in spoken language of proper nouns or, more generally, **Named Entities (NE)** is likely to be an essential component of systems performing tasks such as speech understanding, information extraction and information retrieval. Furthermore the approaches based on the use of NEs have the potential to improve the performance of Large Vocabulary Continuous Speech Recognition (LVCSR) systems (Yoshihiko et al., 1999).

In consequence of these considerations, 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 version of the textual data with NE expressions marked up in SGML format was produced. On the previous example it gives:

<s> bonjour mademoiselle puis-je avoir le numéro de téléphone de mademoiselle <N>GALLAY-VIAL</N> <P>MIREILLE</P> chemin de l' <R>ENVOI</R> treize à <V>SION</V> s' il vous plaît </s>

3.5. 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 3593 remaining sentences on which learn lexicon and language models.

4. Baseline System Development

A baseline speech recognition system able to process 111 natural requests was developed to be used as a reference against which to compare future recognition results. That involved the training of acoustic models, as well as the extraction of lexical and linguistic models from the newly developed resources ("Appels 111" data properly orthographically transcribed and tagged) representative of the targeted application.

4.1. Acoustic Models

In our baseline system (Bernardis et al., 1999), initial acoustic models were trained on the Polyphone database containing prompted read speech (relatively well orthographically transcribed), because Polyphone was designed to cover all the phonemes in a large variety of contexts for the Swiss-French language.

The preprocessing of the speech signal consisted of a RASTA-PLP feature calculation. RASTA-PLP features are particularly robust to convolutional noise and additional noise, so they are well suited for telephone speech (Hermansky & Morgan, 1994).

The hybrid HMM/ANN paradigm, integrating Hidden Markov Models (HMM) and Artificial Neural Network (ANN), was chosen as speech recognition system. A particular form of ANN, referred to as Multi-Layer Perceptron (MLP) was trained to compute local emission probabilities of HMMs given the preprocessed acoustic data.

While yielding similar or better recognition performance than other state-of-the art systems, this approach has indeed been shown (Boite et al., 2000) to have several additional advantages particularly interesting in the framework of our research project, such as:

- A small set of HMM/ANN context-independent phone models is already yielding competitive results compared to a large set of HMM context-dependent phone models, making the system more flexible and better suited to research.
- Given the above, development of new tasks (involving different lexica) is easier. It has also been reported (Boite et al., 2000) 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 to focus research on those most interesting aspects.

4.2. Lexical Modeling

Building upon the general vocabulary list and the 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 (Bernardis et al., 1999).

Phonetic transcriptions of the general vocabulary words were obtained from the BDLex-50000 dictionary (Pérrenou et al., 1987). 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 phonetically transcribed automatically 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.

4.3. Linguistic Modeling

After the raw "Appels 111" textual data initially available had been transformed into a more usable format, its training part was processed to derive various back-off n-gram language models (Bernardis et al., 1999).

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*" → "*DENTS_DU_MIDI*").

5. Preliminary Recognition Results

Speech recognition experiments were done on a cross-validation set of 200 sentences from (properly orthographically transcribed and tagged) "Appels 111" to test our baseline system. They led to a Word Error Rate (WER) of 45.9%, which corresponds to a 46.1% relative improvement of the WER previously obtained on the unconsistent natural speech database that was originally available.

Although this is only an initial result and needs to be improved, it clearly illustrates the importance of a good orthographic transcription of the speech data.

A major source of error in recognizing spontaneously spoken utterances is represented by Out-Of-Vocabulary (OOV) words. Even when using a very large lexicon it is not possible to have complete coverage of the vocabulary (especially proper nouns) used by different speakers. Since this is one of the main problems related to natural speech recognition, we will focus on new ways of detecting and dealing with OOVs.

Moreover, while word-based n-gram language models are very flexible and adaptable, they also suffer from sparse data problem (even more serious in the current research framework because of the limited amount of text training data available to model unconstrained speech) and from the limitations of the statistical model (assuming short term dependencies). Consequently, ways of increasing the reliability of such language models will be investigated: our first step will be to exploit the information represented by Named Entity tags now available for "Appels 111" and to train a NE-tagged language model.

6. Conclusion

This paper described the setting up of acoustic and linguistic resources for research and evaluation in the domain of interactive vocal information servers.

As a properly orthographically transcribed, consistently labeled and tagged unconstrained speech database in Swiss French was currently not available for the targeted area, we concentrated on the annotation and structuration of a natural spoken requests database to make it profitable for lexical and language modeling and for the evaluation of recognition results.

A baseline speech recognition system was trained on the newly developed resources. Its initial testing led to a relative improvement of 46% for the WER compared to the results previously obtained with a baseline system very similar but working on the unconsistent original database. This result shows the importance of having a good orthographic transcription for the speech data.

6.1. Acknowledgements

We thank the Swiss National Science Foundation for their support in this work (INSPECT, INtegrating SPEech acoustic and linguistic ConsTraints for enhanced recognition systems, FNRS project no. 4030/145).

7. References

Chollet, G. et al. (1996). Swiss French Polyphone and Polyvar: Telephone Speech Databases to Model Inter- and Intra-Speaker Variability. Technical Report RR-96-01. IDIAP.

Andersen, J.M. et al. (1997). Advanced Vocal Interfaces Services Swisscom Project: Report for 1997. Technical Report COM-97-06. IDIAP.

Bernardis, G. et al. (1999). Integrating SPEech acoustic and linguistic Constraints: Baseline System Development. Technical Report 99-324. DI-EPFL.

Chappelier, J.-C. & Rajman, M. (1998). A generalized CYK algorithm for parsing stochastic CFG. In Proceedings of TAPD'98 Workshop (pp. 133--137). Paris, France.

Yoshihiko, G. et al. (1999). Named Entity Tagged Language Models. In Proceedings of ICASSP'99 (pp. 513--516). Phoenix, Arizona, USA.

Hermansky, H. & Morgan, N. (1994). RASTA Processing of Speech. IEEE Trans. on Speech and Audio Processing , vol. 2, no. 4 (pp. 578--589).

Boite, R. et al. (2000). Traitement de la Parole. Presses Polytechniques Universitaires Romandes.

Pérrenou, G. et al. (1987). Base de Donnés Lexicales du français écrit et parlé. http://www.irit.fr/ACTIVITES/EQ_IHMPT/bdlex.html.