AUTOMATIC SPEECH RECOGNITION SYSTEM FOR UNDER-RESOURCED LANGUAGES BASED ON SPEERAL : APPLICATION TO BERBER LANGUAGE

Z. Benkhellat[1], *E. Ferreira*[2], *P. Nocera*[2], *M. Guerti*[1]

(1) Informatic Department, University of Bejaia, Bejaia, Algeria
(2) Laboratoire Informatique d'Avignon, LIA, Avignon, France

ABSTRACT

The ability to collect and process a large amount of resources (vocabularies, text corpora, transcribed speech corpora, phonetic dictionaries) constitutes a critical prerequisite of systems based on statistical methods. This problem becomes crucial for languages presenting a lack of computer resources, also known as under-resourced languages, such as African ones. Our work consists in finding an efficient methodology which can improve Speech recognition systems for this kind of languages. This article presents a possible solution proposed for the Berber Language and describe the set of tools used in our study. Namely, we dealt with the problem of insufficient amount of resources by taking into account linguistic specificities of the Berber language and using innovative methods in the building process of ASR resources (acoustic model, lexicon and language model).

Index Terms— Speech recognition, berber language, speeral, under-resourced language

1. INTRODUCTION

Nowadays, in the domain of spoken language technologies, the application of state-of-art techniques and methods is an expensive process which requires a considerable amount of computational resources. That is why only a small part of the world's languages (less than 1%) can benefit of this kind of technologies and related tools [1]. The aim of the PI project [2] consists in proposing viable innovative methods to develop them. This project especially explores strategies intending to exploit all the available resources to improve the accuracy of these technologies.

As a project member, the LIA faced the challenge and currently is in charge of developing an Open Source Toolkit for building a language independent Large Vocabulary Continuous Speech Recognition (LVCSR) system in order to process under-resourced languages. Concerning the Berber language, an African under-resourced language, speech recognition research is only at the beginning of its development. In the current work, we focus on the building process and statistical models improvement of a Berber ASR system based on SPEERAL [3] using the first version of the LIA toolkit.

Our paper is organized as follows:

- In Section 2, some basic elements of the Berber language are presented followed by a description of the available corpus contents, in Section 3.
- Then, in Section 4, we describe the building process of the LVCSR for Berber and the baseline results obtained.
- Next, some improvements and PI utilities used to increase the baseline system performance are explained in Section 5
- Finally, in section 6, we conclude by our future related research concerning the Berber language and the PI project.

2. THE BERBER LANGUAGE

Tamazight (name of the Berber language) belongs to the family of Hamito-Semitic languages. It is spoken in a vast geographical area of North Africa : from Egypt to Morocco. Concerning speakers, they mainly come from Algeria, Morocco, Niger and Mali (Tuareg). Moreover, this language is diversified into several varieties [4].

2.1. Spelling system used by the Berber Language

The Berber, although it has been mainly transmitted through an oral tradition, has its own consonantic spelling system called "Libyan-Berber" (or Tifinagh) which has been established two and a half millennia ago. The figure 1 illustrates the alphabet of this system. Currently it is still used by the Tuareg and the Kabyle (with some variants) communities, but there are two other more recent spelling systems, widely used by available electronic text resources, which correspond to a Latin one and to an Arabic one (Morocco). In order to facilitate the different steps of the acoustic model building and training, we have chosen the Latin spelling system which is also closer to the Berber phonetic transcription [4].

Fig. 1. The Tifinagh Alphabet

2.2. Berber phonology

Modern Tamazight has 43 phonemes : three are vowels, two semi-vowels and 38 consonants. Thus, this language is considered to have a rich consonantism and a poor vocalism [4].

3. CORPUS

3.1. Speech corpus

The acoustic corpus is composed of 2261 recorded audio files with their transcripts. The speech data are digitalized in Wave format with 16 KHz sampling rate and A/D conversion precision of 16 bits. There is a total of 56 Kabylian speakers : 15 men and 41 women corresponding to approximatively 19 hours of read speech. This corpus covers all Berber phonemes, although it is considered tiny. For our study, we divide the corpus into two parts: 14 hours for the training corpus (51 speakers) and 5 hours for the test corpus (5 speakers).

3.2. Text corpus

According to [5], the text corpus was collected from newspapers and web sites which were using the Latin writing system. Subsequently, all the superfluous data (menus, references, etc.) were removed and non-fulltext data, like number or date, were transliterated using language-independent generic automatic script which was developed during this work. The current corpus is exclusively composed of cleaned parts of two Berber newspapers, the "Tamazight tura" and the "Ayamun", corresponding to 445169 words and 67434 sentences.

4. BASELINE ASR SYSTEM

The aim of an ASR system is, from an observed audio signal, to find the most relevant(probable) sequence of words (hypothesis) respecting to all the assumptions made by the system.

4.1. System Description

The ASR system decision process is based on statistical approaches representing the acoustic and linguistic constraints of a language. Thus, we need to build at least three components : an acoustic model, a linguistic model and a pronunciation dictionary. For this purpose, we developed and applied some tools in order to facilitate both building and testing them with the Berber corpus.

4.1.1. Pronunciation Dictionary

The pronunciation dictionary for the Berber language is built by applying an especially designed rule-based tool on a classic lexicon (list of words). This program, called LIA_SCAL_PHON, uses a stochastic approach to determine which rule to apply according to a graphemic context. Another particularity of it consists in the fact that it is language-independent. In that way, the tool provides a standard UTF-8 XML-oriented interface intended to custom the way it performs this task. Thus, the user may configure his own phonetization process by adding rules and exceptions.

As it will be shown in the example below, the user can define a phonetization rule creating a "rule" XML element with a specific structure which can be read as follows : *the grapheme "a", whatever the left (lc tag corresponding to a wilcard symbol) and the right context (rc), can be phonetized as "aa" (specific phone alphabet), like in the word "azul".*

$< rule\ id = "1" >$
$< lc > * < /lc >$
$< graph > a < /graph >$
$< rc > * < /rc >$
$< phon > aa < /phon >$
$< ex > azul < /ex >$
$< /rule >$

Furthermore, the regularity of the Berber language in terms of correspondence between "grapheme" and "phoneme" facilitates the building process. Thus, only 56 rules were created in order to perform the lexicon phonetization task.

4.1.2. Acoustic model

The acoustic model represents the acoustic properties of a language and should be able to cope with the variability of its characteristics. Thus, the acoustic signal is associated with a sequence of units corresponding to the elementary sounds of the language : the phonemes. So, each phoneme of the Berber language is represented by a 5-states Hidden Markov Model (HMM) architecture : 3 emitting states and 2 non-emitting ones as entry and exit which provide the glue needed to join models of HMM units together in the ASR system. Each emitting state consists of 64 Gaussian mixtures trained on vectors of 39 features (13, Δ, $\Delta\Delta$ MFCC) which are extracted from the signal using a 30ms Hamming window with 10ms overlap between frames.

Because of the lack of available resources, a bootstrapping method was employed for training a context independent Berber acoustic models according to [6]. In this method, initially demonstrated in [7], acoustic models are initialized using models from a donor language (or languages) and then they are rebuilt using target language data only. Thus, the most recent LIA French acoustic model is used in order to make the correspondence between the French phoneme states

and the Berber ones. Furthermore, a graphical user interface was designed to facilitate this process. The system was built using the embedded training capability of the LIA acoustic modelling toolkit [3]. The table 1 presents the results of the Acoustic Phonetics Decoding (APD) carried out on the test corpus for each iteration of the training process. In this evaluation, decoding task only took into account the acoustic aspect of the language (no lexicon and langage model are employed). That is to say that the ASR results produced consist in phonetic sequences which are scored using the phonetic transcriptions of the audio test files sample.

Acoustic model	Error (%)
BerberAM0.hmm	67.5
BerberAM1.hmm	44.4
BerberAM2.hmm	43.1
BerberAM3.hmm	41.9
BerberAM4.hmm	41.5
BerberAM5.hmm	41.1

Table 1. The relationship between APD error rate and iterative learning process

4.1.3. The language model

The N-gram language model describes the grammatical and semantic language constraints, assigning on one hand, low probabilities on sequence of words that never occur, on the other hand, high ones on those that appear frequently.

In our study, to train a 3-gram language model from the text corpus detailed in 3.2, we use the SRILM toolkit[1] with Good-Turing discounting and Katz back-off for smoothing. The language model produced obtained a perplexity of 118,83 on the development text sample (extracted from the test corpus) and the resulting lexicon containing 45783 words.

4.2. Results

Considering the entire test corpus, we reached a WER of 32.8% using the context-independent acoustic model and the language model previously built. The table 2 presents a detailed report for each test speakers.

SPKR	Snt	Wrd	Corr	Sub	Del	Ins	Err(%)
Sonia	90	4225	71.1	21.1	7.8	4.3	33.1
Soraya	50	2424	66.9	20.2	12.9	1.1	34.2
Soso	46	2743	65.5	26.7	7.8	8.9	43.4
Thachawith	14	875	63.0	21.5	15.5	0.7	37.7
Zahira	465	20126	70.3	19.7	10.0	1.2	30.9
Sum/Avg	665	30393	69.5	20.6	9.9	2.3	32.8

Table 2. Baseline system results as a function of speakers

[1]http://www.speech.sri.com/projects/srilm/

5. IMPROVEMENTS

5.1. Spelling standardization

According to the baseline system results (detailed analysis and percentage of substitution), we realized that many decoding errors came from the use of alternative spellings of a same word. These differences are found when comparing the reference transcriptions and the results obtained with the ASR system. An example of such variation is given below :

REFERENCE: *yecfa* FELLAS *yighzer si tarda zgan nes1len waccaren is maca di tin n leqdic ur telli ara d tamxaleft gar cwit1 d wat1as kan wamma tamet1t1ut i* DYERNAN FELLAS

ASR RESULT: *yecfa* FELL_AS *yighzer si tarda zgan nes1len waccaren is maca di tin n leqdic ur telli ara d tamxaleft gar cwit1 d wat1as kan wamma tamet1t1ut i* D_YERNAN FELL_AS

Consequently, the computation of the WER was impacted by this kind of non-real errors. Based on this observation we were motivated to work on the spelling standardization of the whole text corpus (train + test) for building an updated language model and lexicon.

In the first place, we manually identified the different kinds of errors (see table 3) and then we automatically performed these changes on the corpus. Nevertheless, it was not sufficient to solve all the standardization problems. Therefore, we also explored an unsupervised Levenshtein distance guided technique which produced some reports which were used to identify possible spelling discrepancies as well as to propose ways to correct them automatically. However, this step also requires a human checking in order to ensure the accuracy of corrections.

unexpected spelling utterance	correction
ittwergel	ittwargel
lewlaya	lwilaya
twacultis	twacult_is
wulis	ul_is

Table 3. Example of different kinds of spelling standardization errors

The table 4 shows that the words' normalization obtained by combining the two approaches previously described significantly improved the performance of the system. Indeed, the WER was enhanced from 32.8% to 25.4%.

5.2. Adaptation to speakers

Another improvement of these results was reached in proceeding to speaker adaptation using the MLLR technique. The aim of this method was to adapt the acoustic model to each speaker considering audio segments belonging to him. Thus, as shown in table 5 the WER increased from 25.4% to

24.1%. Nevertheless, the reasons of the slight gain observed require some further research on inherent characteristics of the Berber language.

5.3. Context dependant acoustic model using an unsupervised states clusterisation

A known way to improve the ASR system performance is to use monophone (context-idependent) HMMs to create triphone (context-dependant) HMMs.

For our work, instead of applying the classical tree based clustering method in order to share the parameters of similar triphone HMMs, we used an innovative approach based on unsupervised state clustering, previously introduced in [8]. This method is based on a low-dimension vectorized representation of states obtained by the use of factor analysis and k-means method. In that way, any additional acoustic knowledge other than the phonemes themselves are no longer mandatory. Applying this method on an under-resourced language is an interesting perspective considering the fact that we were often hampered by a lack of sufficient language knowledge, or of experts, for formulating accurate acoustic questions (used in the decision tree) in order to obtain relevant clusters of context-dependant HMMs parameters.

In this perspective, we carried out some experiments which obtained some interesting results. However, we will need to compare this approach with other related methods (like automatic question generation for decision tree based state tying) before proceeding to any presentation of our study.

6. CONCLUSION

In this paper, we have presented some characteristics of the Berber Language and the available electronic resources. Then we described the building process of a LVCSR conceived to work on an under-resourced language and some tools used to obtain the Berber baseline system. We have also talked about performance enhancement, using the spelling system standardization and the speaker adaptation. The system showed an encouraging WER of 24,1%, although it is still higher than the state-of-the art French one. For future research, our building method of a context-dependent acoustic model will be compared to other related works and new approaches concerning the language model training will be explored (semi-semantic approach).

SPKR	Snt	Wrd	Corr	Sub	Del	Ins	Err(%)
Sonia	90	4280	79.8	12.5	7.6	2.0	22.2
Soraya	50	2399	72.4	15.1	12.5	1.2	28.8
Soso	46	2884	74.1	17.1	8.7	4.2	30.1
Thachawith	14	866	68.5	17.6	14.0	1.0	32.6
Zahira	465	19896	76.5	14.1	9.5	1.2	24.7
Sum/Avg	665	30325	76.2	14.3	9.5	1.6	25.4

Table 4. Performance improvement through standardization of spelling

SPKR	Snt	Wrd	Corr	Sub	Del	Ins	Err(%)
Sonia	90	4314	80.8	11.2	8.0	1.7	21.0
Soraya	50	2413	74.4	14.1	11.5	1.2	26.8
Soso	46	2906	75.8	15.6	8.6	3.9	28.1
Thachawith	14	868	68.9	17.3	13.8	1.0	32.1
Zahira	465	19977	77.7	13.1	9.2	1.2	23.5
Sum/Avg	665	30478	77.5	13.3	9.3	1.5	24.1

Table 5. Performance improvement proceeding speakers adaptation

7. REFERENCES

[1] Vincent Berment, *Méthodes pour informatiser les langues et les groupes de langues peu dotées*, Ph.D. thesis, Université Joseph Fourier, Grenoble 1, France, March 2004.

[2] Pascal Nocera, Laurent Besacier, and Eric Castelli, "The PI Project : Spoken Language Technologies for p-languages," in *African HLT 2010*, Djibouti, jan 2010.

[3] Pascal Nocera, Georges Linares, Dominique Massonié, and Loïc Lefort, "Phoneme lattice based a* search algorithm for speech recognition," in *Text Speech and Dialogue*. 2002, p. 83111, Springer.

[4] Salem Chaker, *Textes en linguistique berbère: introduction au domaine berbère*, Ed. du C.N.R.S., Paris, 1984.

[5] Viet Bac Le, Brigitte Bigi, Laurent Besacier, and Eric Castelli, "Using the web for fast language model construction in minority languages," *Eurospeech 2003 Geneva*, pp. 3117–3120, September 2003.

[6] Viet Bac Le and Laurent Besacier, "First steps in fast acoustic modeling for a new target language : application to vietnamese.," *ICASSP Philadelphia USA*, pp. 821–824, March 2005.

[7] Louise Osterholtz, Charles Augustine, Arthur McNair, Ivica Rogina, Hiroaki Saito, Tilo Sloboda, Joe Tebelskis, and Alex Waibel, "Testing generality in janus: A multilingual speech translation system," *ICASSP'92*, vol. 1, pp. 209–212, 1992.

[8] Mohamed Bouallegue, Driss Matrouf, and Georges Linares, "A simplified subspace gaussian mixture to compact acoustic models for speech recognition.," in *ICASSP'11*, 2011, pp. 4896–4899.