IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 5, May 2016.
www.ijiset.com
ISSN 2348 – 7968

The design of an Albanian large vocabulary continuous speech recognition system

Ervenila Musta[1], Ligor Nikolla [2] Alvin Asimi[3]

[1] Department of Mathematics, Polytechnic University of Tirana/ Mathematics and Physics Engineering Faculty, Tirana/Albania

[2] Department of Mathematics, Polytechnic University of Tirana/ Mathematics and Physics Engineering Faculty, Tirana/Albania

[3] Department of Mathematics, Polytechnic University of Tirana/ Mathematics and Physics Engineering Faculty, Tirana/Albania

Abstract

This paper presents an overview of speech recognition technology. This report presents an overview of speech recognition technology, software, development for Albanian language. It begins with a description of how such systems work, and the level of accuracy that can be expected. Present work is aimed at developing suitable speech recognition for Albanian language. In this paper we are using a HMM (hidden Markov model) to recognize speech samples to give excellent results for isolated words. It consists of isolated words that are separated by silences. The advantage of discrete speech is that word boundaries can be set exactly while with continuous speech; words will be spoken without silences.

The main objective of this thesis was to develop a speaker-independent large-vocabulary continuous speech recognition system for Albanian, a under-resourced language. This system should be able to recognize general Albanian continuous speech produced by any speaker with a decent performance

Keywords: - *Speech Recognition, HMM., Acoustic modeling*

1. Introduction

Speech is the primary means of communication between people. Speech recognition, generation of speech waveforms, has been under development for several decades [10]. Automatic speech Recognition is a process by which a computer takes a speech signal and Converts it into words [1]. It is the process by Which a computer recognizes what a person Said. Keyboard, although a popular medium, is not very convenient, as it requires a certain amount of skill for effective usage .A mouse on the other hand requires a good hand eye co-ordination. Physically challenged people find computer difficult to use. Partially blind people find reading from a monitor difficult. All these constraints have to be eliminated. Speech interface help us to tackle these problems. The objective is to trap human voice in a digital computer and decode it into corresponding text. Speech recognition can be defined as the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. When two people speak to one another, they both recognize the words and the meaning behind them. Computers, on the other hand, are only capable of the first thing: they can recognize individual words and phrases, but they don't really understand speech in the same way as humans do. Computer recognizes the command and software tells the computer what to do when that command is recognized.

One of my targeted languages was Albanian, a language with poor resources..These languages are spoken by a large number of people, but so far too few acoustic resources (speech data bases) and linguistic resources (text corpuses) were acquired in order to develop an unconstrained continuous speech recognition system. The Baum-Welch training paradigm requires speech audio clips along with their textual transcriptions in order to estimate the models parameters. Thus, speech databases are critical resources along with their characteristics, such like the number of hours of speech, number of speakers, etc, in developing a speech recognition system.

2. Classification of Seech Recognition System

There is a large variety in the speech recognition technology and it is important to understand the differences. One can classify speech recognition systems according to the type of speech, the size of the vocabulary, the basic units and the speaker dependence[2]. The position of a speech recognition system in these dimensions determines which algorithm has to be used.

A. Type of speech There are basically two types of speech:

1. Continuous speech
2. Discrete speech.

Discrete speech consists of isolated words that are separated by silences [3]. The advantage of discrete speech

ISSN 2348 – 7968

is that word boundaries can be set exactly while with continuous speech; words will be spoken without silences.

B. Size of the vocabulary

The size of the vocabulary is the second typical aspect of a speech recognition technology. The vocabulary is a set of words that have to be recognized. A small vocabulary is one, which contains less than about 30 words. A 500-word vocabulary is average size. A vocabulary with more than 25000 words generally will be seen as very big, although these definitions tend to depend on the application field [6].

C. Speaker dependence

- Speaker dependent system
- Speaker independent system
- Speaker adaptable system

Some speaker-dependent systems require only that the user record a subset of system vocabulary to make the entire vocabulary recognizable. A speaker-independent system does not require any recording prior to speaker-dependent system requires that the user record an example of the word, sentence, or phrase system use. A speaker independent system is developed to operate for any speaker of a particular type (e.g., American English). A speaker adaptive system is developed to adapt its operation to the characteristics of new speakers[5].

3. Design of the System

The prepared system if visualized as a block diagram will have the following components: Sound Recording and word detection component, feature extraction component, speech recognition component, acoustic and language model.

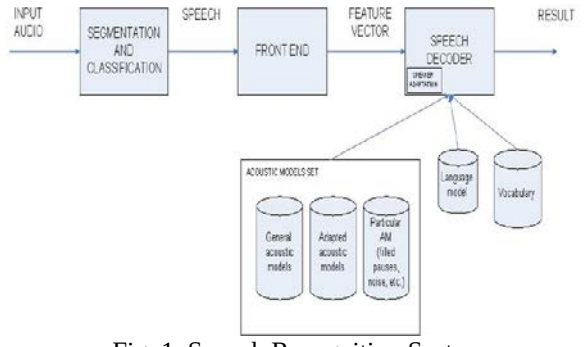

Fig. 1 Speech Recognition System

A. Sound Recording and Word detection component.
The component is responsible for taking input from microphone and identifying the presence of words. Word detection is done using energy and zero crossing rate of the signal. The output of this component can be a wave file or a direct feed for the feature extractor

B. Feature Extraction component:
The component generated feature vectors for the sound signals given to it. It generates Mel Frequency Cepstrum Coefficients and Normalized energy as the features that should be used to uniquely identify the given sound signal [5].

C. Recognition component:
This is a Continuous, Multi-dimensional Hidden Markov Model based component. It is the most important component of the system and is responsible for finding the best match in the knowledge base, for the incoming feature vectors

D. Knowledge Model:
The components consist of Word based Acoustic. Acoustic Model has a representation of how a word sounds. Recognition system makes use of this model while recognizing the sound signal

Once the training is done, the basic flow can be summarized as the sound input is taken from the sound recorder and is feed to the feature extraction module. The feature extraction module generates feature vectors out of it which are then forwarded to the recognition component. The recognition component with the help of the knowledge model and comes up with the result. During the training the above flow differs after generation of feature vector. Here the system takes the output of the feature extraction module and feeds it to the recognition system for modifying the knowledge base[11].

4. Speech database acquisition for Albanian ASR

A complete speech database is formed from :

- ❖ a set of speech signal samples.
- ❖ a set of transcriptions, which must be perfectly synchronized with what is spoken in each speech sample.
- ❖ additional information regarding speech type (isolated words, continuous, spontaneous).

ISSN 2348 – 7968

We have started to build speech databases by extracting fragments from website news. These audio clips also had correspondent transcriptions, which, in most of the cases, were related. We had access at 4 news databases : www.balkanweb.tv, www.topchannel.tv, www.topchannel2.tv and www.vizionPlus.tv. SpeeD gave us access to each database's .php files. By processing the .php files specific for each database, we searched for the URL in each file and created two lists : one containing the fileids of the files, and the other one the correspondent link. 3

5. Data preparation for training and testing an ASR

CMU Sphinx project offers the possibility to create acoustic models for a new language. The trainer learns the parameters of the models of the sound units using a set of sample speech signals. This is called a training database. The trainer needs to be told which sound units he has to learn the parameters of, and at least the sequence in which they occur in every speech signal in your training database. This information is provided to the trainer through a file called the transcript file, in which the sequence of words and non-speech sounds are written exactly as they occurred in a speech signal, followed by a tag which can be used to associate this sequence with the corresponding speech signal. The trainer then looks into a dictionary which maps every word to a sequence of sound units, to derive the sequence of sound units associated with each signal. The are two dictionaries, one in which legitimate words in the language are mapped sequences of sound units (or sub-word units), and another in which non-speech sounds are mapped to corresponding non speech or speech-like sound units. They are referred as the language dictionary and the latter as the filler dictionary.[21] The file structure for the database is:

- etc
  - database.dic - Phonetic dictionary
  - database.phone - Phoneset file
  - database.lm.DMP - Language model
  - database.filler - List of fillers
  - database.fileids - List of files for training
  - database.transcription - Transcription for training
- wav
  - fileID.wav – Recording of speech utterances

6. Testing

It is critical to test the quality of the trained database in order to select best parameters, understand how application performs and optimize parameters. To do that, a test decoding step is needed. The decoding is now a last stage of the training process. After testing the models built on the small database, the next step was to try to extend them, in order to obtain an accurate speech recognition system.

7. Final Results for Albanian ASR

The total amount of training speech data summed up to about 8 hours of speech from 23 different speakers and 6 hours of audio clips extracted from the websites' databases. SD2 contains audio clips which are not of god quality, being filtered at 5.5 kHz. SD4 contains audio clips extracted from www.topchannel2.tv in which several speakers are present. Even though we have the transcriptions for all the spoken text, the clips are too long. Consequntly, Sphinx toolkit fails to perfectly align the audio to the transcription.

Table 1.1 Albanian Speech database

ID	database	duration	Filtered low-pass[kHz]	type	comments
SD1	Pc recorder	2	8	recordings	Native speaker
SD2	Chunk1-6	3.20	5.5	broadcastnews	expert
SD3	Chunk7-10	2	8	broadcastnews	Expert
SD4	Topchanel 2	5.40	8	broadcastnews	loose

Table 1.1 summarizes the data regarding the text corpora that were collected and further used in the experiments. The numbers are computed on the clean corpora (after the processing operations).

Table 1.2 Albanian Text copora

ID	database	#total words	#unigrams	#phrases
TD1	Pc recorder	14063	1165	969
TD2	Topchannel+vizionplus+balkanweb	48560551	377170	4841049
TD3	Pc recorder+chunk1-10	74063	11342	3340
TD4	Chunk7-10	22003	11342	845
TD5	Topchannel2	57129	12310	187

The acoustic models were created using the CMU Sphinx toolkit and the default training strategy. We employed (– states) HMMs to model context-dependent phones (triphones) using MelFrequency Cepstral Coefficients (MFCCs). The total number of HMM states (called senones) was limited to 1000. Every senone was modelated

ISSN 2348 – 7968

with a Gaussian Mixture Model (GMM) with 8 Gaussian components.

Table 1.3 Albanian Acoustic Models

ID	Trained on
AM01	SD1
AM02	SD1+SD2
AM03	SD1+SD2+SD4
AM04	SD3
AM05	SD3+SD4

Table 1.4 presents the various experiments made during a period of one year, consequently revealing the evolution of our ASR system.

Table 1.4 Albanian experimental results

exp	ASR system	Used models	Evolution set	accuracy	Error ate
1	ASRS-1	AM05+LM05	SD1	14.56	85.54
2	ASRS-2	AM02+LM02	SD1+SD2	45.63	54.37
3	ASRS-3	AM04+LM01	SD3	38.88	61.62
4	ASRS-4	AM02+LM01	SD3	25.21	74.79
5	ASRS-5	AM04+LM03	SD3	72.39	27.61
6	ASRS-6	AM02+LM03	SD3	66.14	33.86
7	ASRS-7	AM04+LM04	SD3	40.87	59.13

Several conclusions can be drawn from Table 1.4 . First, ASRS-1 had a low accuracy percentage. From these results, we drawn the conclusion that topchannel2 was not a trust worthy speech database. The audio clips were too long, and the Sphinx toolkit failed to align the whole clip to the audio. Moreover, since the clips were raw news broadcasts, the environment in which they were recorded presented different disadvantages. These wavs contained high additive noice, multiple acoustic sources, like music or other people talking in the back, or reverberant environments. Another issue worth mentioning is the difference between ASRS-4 and ASRS-6. Even though they are tested on the same database, that is SD3, and use the same acoustic model, that is AM02, their performance is very different. Obviously, the difference lies in the language model. ASRS-6 has a greater accuracy percentage because the language model is trained on the same database on which it is tested. Even though ASRS-6 would seem a good ASR at a first glance, these are called artificially improved results. In the end, the best configuration seems to be ASRS-5. But this configuration uses models that are trained and tested on the same database, that is SD3. In order to properly evaluate an ASR, it must be tested on unseen data. Consequently, we trained an interpolated language model LM04 (MediaEval2013 + topchannel2 (90%) + all news corpora (10%)) in order to obtain some real results. ASRS-7 reflects the real performances of our continuous recognition system.

8. Conclusion

The main objective of this paper was to create an automatic speech recognition system for language of Albanian. Albanian is a so called low-resourced language, a language that is spoken by a large number of people, but no prior work of collecting and organizing speech and/or text resources has been done. At this part, our contribution was to acquire a speech database in order to train and evaluate the ASR.

The paper describe explicitly the steps in designing an ASR from zero. It presents the problems encountered when gathering resources for building a speech database. After the required resources are described, together with the cleaning tools for the text corpora, several experiments are presented. As a conclusion, we have managed to build a 14 hours speech database which can be further used to design a large vocabulary-speech recognition system. The results presented in this thesis approach the reality, since all the tests are done on unseen data. It comes by default with the CMU Sphinx toolkit, my contribution being the GUI interface. It is a user-friendly application through which one can test our automatic speech recognition systems.

.

.

References

[1] Anne Johnstone Department of Artificial Intelligence Edinburgh University Hope Park Square, Meadow Lane Edinburgh EHB 9LL, (GB) Gerry Altmann “AUTOMATED SPEECH RECOGNITION: A FRAMEWORK FOR RESEARCH”.

[2] Reddy, D.R. & Ermann, L.D. 1975. “Tutorial on System Organisation for Speech Understanding.” In D.R. Reddy (ed) Speech Recognition, Academic Press.

[3] Rumelhart, D.E. & McClelland, J.L. 1982. “An Interactive Activation Model of Context Effects in Letter Perception: Part II. The Contextual Enhancement Effect. Some Tests and Extensions of the Model. In Psychological Review”

[4] Tetsuya Matsumoto, Kazuhito Hagio, and Masayuki Takeda Department of Informatics, Kyushu University, Fukuoka 819-0395, Japan {tetsuya.matsumoto, kazuhito.hagio, takeda “More Speed and More Compression: Accelerating Pattern Matching by Text Compression”

[5] C. H. Lee; F. K. Soong; K. Paliwal "An Overview of Speaker Recognition Technology", Automatic Speech and

ISSN 2348 – 7968

Speaker Recognition: Advanced Topics. Kluwer Academic Publishers 1996, Norwell, MA.

[6] "Pattern matching for large vocabulary speech recognition systems" available at www.freepatentsonline.com/6879954.html

[7] "Isolated-word automatic speech recognition (iwasr) design" available at www.dspace.fsktm.um.edu.my/xmlui/bitstream/handle/1812/111/Chapter%205.pdf?sequence=7

[8] R. Rodman, "Computer Speech Technology". Artech House, Inc. 1999, Norwood, MA 02062.

[9] M. J. Castro; J. C. Perez, "Comparison of Geometric, Connectionist and Structural Techniques on a Difficult Isolated Word Recognition Task.", Proceedings of European Conference on Speech Comm. and Tech., ESCA, Vol. 3, pp 1599-1602, Berlin, Germany, 1993

[10] Rabiner Lawrence ,Juang Biing-Hwang, "Fundamental of speech recognition ",AT & T,1993.

[11] "Speech recognition for Hindi language", C-DAC India, available at http://www.cdacmumbai.in/design/corporate_site/override/pdfdoc/speech_ recognition_for_hindi.pdf