A LARGE SPEECH CORPUS DEVELOPMENT

Carlos Alberto Ynoguti[1], Thiago Pereira Raposo[2] and Fabrício Vital[3]

Abstract *— Speech recognition systems use statistical methods based algorithms, and therefore need several training samples to perform properly. Consequently such systems require huge databases for training and testing. Large speech corpora development are hard to construct, and in Europe and in the USA, their development was possible only with the cooperation among research centers, universities, private companies and the government. In these countries, the availability of such databases provided the resources for the great improvement in speech technologies observed in the last years. Here in Brazil, such consortiums are not even mentioned, and the researchers have to work with small, locally developed databases. In this article we report an effort to develop a large speech corpus for the Brazilian Portuguese to fulfill this important lack for research in this area.*

Index Terms — speech processing, speech recognition, speech corpus.

INTRODUCTION

Spoken language is the natural way human beings use to communicate each other. Its structure is based on phonological, syntactic and prosodic structures of the language, on the acoustical environment, on the context in which the speech is being produced (e.g. people speak differently in noisy and silent environments), and on the channel through which it is transmitted (telephone channel, microphone, etc.)

Each person produces the speech in a different way, and the differences are due to differences in accent, shape and size of vocal tract, rhythm, among other factors. Furthermore, speech patterns are modified by the physical environment, social context, and speaker physical and emotional conditions.

The most promising technologies in speech recognition (Artificial Neural Networks and Hidden Markov Models [6][9][12][13]) use statistical modeling techniques that learn by examples. To provide this training samples, the training database must be large enough to cover all the phonetic, linguistic and acoustic phenomena encountered in spoken language. In fact, bad modeled variables (such as channel or microphone differences, out of vocabulary words, bad trained subunits) cause a devastating effect in the system's overall performance. So, in order to provide sufficient training samples for the statistical methods to work properly, the training database should be large enough.

Speech synthesis and coding do not require such large databases but need some material for evaluation and testing.

Unfortunately, such databases are very expensive to construct. These high costs can only be accomplished by a joint effort of private institutions, research centers and public funding agencies, in order to distribute tasks and avoid doubling efforts. Also, to involve more people in this process, this material should not be specific to one area or task, but instead, serve to as many groups and research areas as possible, in several knowledge areas (speech coding, synthesis and recognition, phonetic and linguistic studies, etc.)

In Brazil, due to the disinterest of the private sector and the lack of government incentives, there is no such speech corpus available in public domain. Some private companies, such as IBM, had speech corpus in Brazilian Portuguese, but unfortunately they are for private use only.

In Brazil, such consortiums were not even contemplated, and the researchers here have to develop their research using small, locally developed databases, which try to cover the most significant aspects of the spoken language, unsuccessfully in the great majority of the cases.

For this reason, a large speech corpus is of great importance for this area to achieve the great development verified in Europe and in the USA. A 500 speakers database seems to be a reasonable goal, comparing to European databases [2][3].

The final point is that this work has a purely scientific approach, that is, it's not intended to earn money

SPEECH DATABASES AROUND THE WORLD

In Europe, France, Portugal, Italy, Germany, Greece, England, Denmark, Spain, Norway, Sweden and the Netherlands, joined in the EUROM_1 project [2], there is a multinational effort to create a speech corpus in the languages spoken in that countries. For each country, a database consisting of 60 speakers (30 males and 30 females), selected in the same way and recorded in the same conditions with common file formats was created.

In Portugal, it was also created a database called BD-PUBLICO (Base de Dados em Português eUropeu, vocaBulário Largo, Independente do orador e fala

[1] Carlos Alberto Ynoguti, INATEL National Institute of Telecommunications, Av. João de Camargo, 510, 37540-000, Santa Rita do Sapucaí, MG, Brazil, ynoguti@inatel.br
[2] Thiago Pereira Raposo, INATEL National Institute of Telecommunications, Av. João de Camargo, 510, 37540-000, Santa Rita do Sapucaí, MG, Brazil, thiagop@inatel.br
[3] Fabrício Vital, INATEL National Institute of Telecommunications, Av. João de Camargo, 510, 37540-000, Santa Rita do Sapucaí, MG, Brazil, fvital@inatel.br

COntínua), with about 10 million words in approximately 156 thousands sentences, pronounced by 120 speakers (60 males and 60 females). This speech corpus was developed by a joint effort of research centers, the government and some private institutions [3].

In the USA a great effort was made in this sense and there are available in the public domain several speech databases for development and evaluation of speech processing systems.

The Linguistic Data Consortium [8] is an open consortium of universities, companies and government research laboratories. It creates, collects and distributes speech and text databases, lexicons and other resources for research and development purposes. The University of Pennsylvania is the LDC's host institution. The LDC was founded in 1992 with a grant from the Advanced Research Projects Agency (ARPA), and is partly supported by the Information and Intelligent Systems division of the National Science Foundation. Most of the well known speech corpora (TIMIT, TI-DIGITS, SWITCHBOARD, Wall Street Journal, etc.) are available at LDC.

Also, research centers as the Center for Spoken Language Understanding from Oregon Graduate Institute are making solo efforts to construct particular databases [4][5].

In the United States there are also many speech corpora available in public domain for speech systems development and evaluation.

The availability of large databases allowed an expressive improvement in the speech technologies, not only by alleviating the task of corpus development, a hard and expensive task, but also by providing a way to compare the results achieved by different researchers in a statistically significant way.

METHODOLOGY

The goal of this work is to be an initial step to the construction of a large speech database. The idea is to distribute an acquisition software among the researchers interested in this kind of material and ask them to contribute with some speakers. With contributors in all regions of the country, it would be possible to quickly construct a large corpus at a very low cost. In this sense, researchers of many universities are being contacted to help in this effort.

It is intended to collect recordings that contemplate the majority of the applications in speech technology, such as machine commands, continuous speech, connected digits and others.

Eventually, this database could became a reference in the speech area, allowing researchers to implement and test their ideas, and to compare their results in a statistically consistent way. Also, the methodology used to generate the database allows it to be in continuous expansion.

To achieve this goal, the whole project was divided in the following stages:

- Study of dialects and geographic distribution of the speakers.
- Determination of the types of locution
- Acquisition software development
- Recordings
- Phonetic transcription
- Organization of the database

Next, each topic will be described in more details.

Study of dialects and geographic distribution of the speakers.

For a database to be representative, it's necessary that it has utterances from people representing all the accents found in the country. It's not an easy problem and, in fact, neither the number of different accents nor the accents themselves are determined for Brazilian Portuguese.

Another problem one has to face is how many people to record from each region/accent? Which criterion should be used? The first idea that comes to mind is: the percentage of speakers of one determined region must be proportional to the number of inhabitants of that region. However, other factors have to be taken into account:

- Percentage of the people who will really use the speech technology;
- Economic importance of each region;

Clearly, these issues are not easy to handle and further study must be done in order to have a truly representative speech corpus. Despite of these issues, it's necessary to define the number of speakers to be collected from each region, and for the first approach, the intention is to collect as many speakers as possible. Afterwards, when (and if) these linguistic studies become available, it's always possible to collect more utterances in order to balance the database.

Determination of utterance types

The goal here is to cover all the possible applications of speech technology. After long research and exhaustive discussion, one decided to contemplate the following topics:

- **Continuous speech.** ("My son is sick and I'll bring him to the doctor.") – 20 utterances per speaker. In order to model all the phonetic and grammatical variations, it is interesting that this database should be as assorted as possible. Good sources of such kind of sentences are the newspapers, the internet, magazines, books and others. Up until the moment, about 10000 sentences were collected, and it's expected to collect material enough for 1000 speakers (20000 sentences). Of course, it's a hard and tedious work, and an acquisition software was developed for this purpose. Also, sentences were limited to have 8 to 12 words each, so that there are not too short or too long ones.It is desirable that the set of sentences sent to each speaker has at least one sample of each phoneme. Being thus, the verification of the

phonetic content of the sentences assigned to each speaker becomes necessary. In the case of absence of some phoneme, one of the sentences is substituted for one that contains the absent phoneme. The counting and verification of phonemes is executed through an automatic transcription software, developed by the researchers of the Institute of Studies of the Language of the University of Campinas.

- **Connected digits.** ("five seven eight oh six zero two one") – 5 utterances per speaker. For this part, a software was developed to generate sequences of 4 to 8 digits in a random fashion.
- **Numbers in full.** ("two hundred thousands, three hundreds and forty") – 5 utterances per speaker. As in the previous case, a software was developed to generate random numbers and to transcribe them for the speaker to read.
- **Isolated words.** ("open", "print", "left", etc.) – 5 sets of 5 utterances per speaker. These words were chosen to meet applications such as computer operation, machine operation, banking services, etc. For each speaker, a set of 25 words is chosen in a random manner. For balance, it was not allowed for the same speaker to utter the same word twice, and the number of occurrences of each word in the whole database is set to be equally distributed.
- **Spelled words.** ("S-H-A-K-E-S-P-E-A-R-E") 5 utterances per speaker. Usually, the application of this kind of utterance is to give a name for a given service (e.g. banking service, air travel reservation), specially for foreign names. So, the speakers are asked to spell their first name, their last name, the first name of their father and mother, and the last name of the city they live.
- **Semantically unpredictable sentences.** ("Blue lions fall from Java's basements.") - 5 utterances per speaker. Semantically unpredictable sentences like the one from the example above are used for speech synthesis systems evaluation: when the listener cannot predict which word will be pronounced next, it's necessary to really understand what was spoken.
- **Sentences for prosodic study**. 4 to 8 utterances per speaker.
  1. "I **see** the **sea**"
  2. "I **see** the blue **sea**"
  3. "The blue **sea** is what I **see**."
  4. "I **see** that you want to go to the **sea**."

  The sentences listed above (and similar sets) are intended to evaluate the prosodic aspects of words uttered in different positions inside the utterance. Further, each sentence should be uttered in three different ways (slow, normal and fast) so that one can construct rhythm models of speech.
- **Spontaneous speech.** 1 utterance per speaker. The application for this topic is human-machine interface, and word spotting methods. For example, for utterances like "I'd like to know my credit card number", "Please tell me my credit card number", the system must understand that the information required is the credit card number. The idea here is to create situations in which the speaker is asked to formulate a question or make a comment about some topic. Examples of motivating questions are:
  "Ask for information about the movies for tonight."
  "Make a comment about the weather"
  "Ask for a pizza on a delivery service"

Construction of the acquisition software

The acquisition software is one of the most important part of this project because it will make it possible the acquisition of the utterances in a fast and low cost way. This software performs the following tasks:

- Before starting the recording session, speakers fill a register with their name, surname, age, sex, education level, profession, city where he was born, name of the father and the mother and cities where he lived. This last item has great importance because accent is defined until the fourteen years of age. The father and mother's names are asked for the spelled words section.
- After the registration part, the speaker goes to the recording session. The acquisition system shows the sentence to be uttered in the computer screen, together with recording controls, so that the recording can be made in an easy way. Also, the system check for recording saturation and, if this occurs, the speaker is asked to repeat that sentence.
- Recordings concluded, the software sends the information via ftp protocol to Inatel, so that the data can be stored and organized.

Recordings

The recording task will be distributed among the people who are interested in collaborate with this effort. An instruction manual was elaborated and will be distributed together with the software, with directions about the microphone type, recording environment, sound card type, registration and sending procedures.

For this first version of the database, it's intended to collect utterances in a office environment, with a good quality microphone.

Phonetic transcription

The ideal would be that the utterances could be manually transcribed by linguistic specialists team. But this involves costs and, at the moment, due to lack of financial resources, this task will be carried out automatically, using a software developed at the Institute of Language Studies (IEL) of the State University of Campinas (UNICAMP).

Database organization

As the recording stage ends, it is necessary to organize them in order to facilitate their manipulation. It is intended to divide the speakers into training and test sets. Also, the speakers can be clustered in terms of age, gender and accent. For this purpose a SQL based software is being developed.

CONCLUSION AND FUTURE WORK

In this article, a large speech corpus development effort is described. Researchers frequently report the difficulty to work in speech processing area without large enough databases. Also, with this database available, results can be compared in a statistically significant way.

Most of the applications of speech technology were contemplated, and further ones can be added along the time. Furthermore, the recording process can continue throughout, and the database can be in continuous expansion.

For the future, it is intended to build a telephone quality speech corpora by passing this database through a telephone line.

REFERENCES

[1] Jelinek, F. "Statistical methods for speech recognition", *MIT Press,* 1998.

[2] "EUROM_1 : a multilingual European speech database". http://www.icp.grenet.fr/Relator/multiling/eurom1.html#PortugCorpus (31/03/1999)

[3] "BD-PUBLICO (Base de Dados em Português eUropeu, vocaBulário Largo, Independente do orador e fala Contínua)" http://www.speech.inesc.pt/bib/Trancoso98a/bdpub.html (31/03/1999).

[4] Cole, R. et. al., "Corpus development activities at the Center for the Spoken Language Understanding", *Proceedings of the ARPA Workshop on Human Language Technology*, April 7-11, 1994.

[5] Cole, R. et. al., "Telephone speech corpus development at CSLU". *Proceedings of ICSLP*, Yokohama, Japan, September, 1994.

[6] Cole, R., ed.., "Survey of the State of the Art in Human Language Technology", http://cslu.cse.ogi.edu/publications/index.htm, (26/10/98).

[7] Haykin, Simon, "Neural Networks - A Comprehensive Foundation", *MacMillan Publishing Company*, New York ,1994.

[8] http://www.ldc.upenn.edu/Catalog/index.html(14/08/2001)

[9] Lee, K. F., Hon, H. W., Reddy, R., "An overview of the SPHINX speech recognition system", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol 38, No 1, April, 1990,pp. 35-45.

[10] Pallet, D. S. et al, "1997 broadcast news benchmark test results: English and non-English", *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, February 8-11, 1998, Lansdowne, Virginia

[11] Rabiner, L., "Fundamentals of speech recognition". *Prentice Hall Press*, 1993.

[12] Rudnicky, A. I., Hauptmann, A. G., and Lee, K. F., "Survey of Current Speech Technology", http://www.lti.cs.cmu.edu/Research/cmt-tech-reports.html, (22/11/1998).

[13] Tebelskis, J., "Speech recognition using neural networks", *Ph.D. Thesis*, School of Computer Science, Carnegie Mellon University, 1995.