A Brazilian Portuguese Speech Database

Carlos Alberto Ynoguti & Fábio Violaro

***Abstract*— The construction of large speech databases is generally only achieved with the cooperation among research centers, universities, private companies and the government. This is the model that the USA and countries of the European comunity use. In less resourced countries, like Brazil, such consortiums are not even mentioned, and the researchers have to work with small, locally developed databases. To face this problem, a corporative approach to develop a speech corpus for Brazilian Portuguese is proposed. In this approach, every researcher that is part of the project should contribute with some speakers. With contributors in various regions of the country, it should be possible to collect a reasonable number of speakers at a low cost, in a relatively short time. This article reports the database specification details and the current status of the work.**

***Keywords*— speech database, Brazilian Portuguese.**

I. INTRODUCTION

Speech technologies have now evolved from laboratory prototypes to practical systems in the great majority of countries all over the world.

One of the main reasons for the current state of art of such technologies is the existence of large speech corpora, such as the TIMIT, WSJ, SpeechDat and others [1]. In fact, such databases allowed researchers to compare their results in a statistically consistent way, focusing on the ideas that really work, instead of ideas that work only for a small group of speakers.

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 (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, have speech corpus in Brazilian Portuguese, but unfortunately they are intended for private use only.

To fullfill this gap, the present work proposes a low cost approach for the task of building those speech databases. The idea is to form a consortium of researchers interested in this kind of material, and divide the total work among participants: if each member of such consortium contributes with some speakers, it would be possible to construct a relatively large database, with a low cost, and in a short space of time.

Carlos Alberto Ynoguti, Instituto Nacional de Telecomunicações, Santa Rita do Sapucaí, Minas Gerais, Brazil, E-mail: ynoguti@inatel.br. Fábio Violaro, Departamento de Comunicações, Faculdade de Engenharia Elétrica e Computação, Universidade Estadual de Campinas, São Paulo, Brasil, E-mail: fabio@decom.fee.unicamp.br. Este trabalho foi parcialmente financiado pela FAPEMIG.

II. METHODOLOGY

The philosophy of the project is to share the workload among the participants. Therefore, 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 (or almost all) regions of the country, it is expected to be possible to quickly construct a large corpus at a very low cost. Also, the methodology used to generate the database allows it to be in continuous expansion.

A. Acquisition software

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

- Before starting a recording session, the speaker fills a register with his/her name, surname, age, gender, education level, profession, city where he was born, name of his father and his mother and cities where he has lived in. 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 checks 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 (Instituto Nacional de Telecomunicações, Santa Rita do Sapucaí, MG, Brasil), the institution where this database implementation is centralized, so that the data can be stored and organized.

B. Utterance types

The most promising technologies in speech recognition (artificial neural networks and hidden Markov models [2][3][4][5]) 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 work properly, the training database should be large enough.

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

In order to cover all of these applications, the following set of sentences was designed, to be pronounced by each speaker:

- **Continuous speech**. 20 utterances of different sentences per speaker (Ex.: "Meu filho está doente, e eu vou levá-lo ao médico"). 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 is 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. To satisfy this requirement, 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 by another one that contains the absent phoneme. The counting and verification of phonemes is executed through an automatic ortographic-to-phonetic transcription software, developed by the researchers of the Laboratório de Fonética e Psicolingüística in the Instituto de Estudos da Linguagem at the Universidade Estadual de Campinas (LAFAPE/IEL/UNICAMP)[6].

- **Connected digits**. 5 utterances of different digit sequences per speaker (Ex.: "cinco sete oito zero seis meia dois"). For this part, a software was developed to generate sequences of 4 to 8 digits in a random fashion.

- **Numbers in full**. 5 utterances of different numbers per speaker (Ex.: "dois mil, trezentos e quarenta e sete"). As in the previous case, a software was developed to generate random numbers and to transcribe them for the speaker to read.

- **Isolated words**. 25 utterances of different words per speaker (Ex.: "abrir", "imprimir", "esquerda", etc.). 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 isolated words database is set to be equally distributed.

- **Spelled words**. 5 utterances of different spelled words per speaker. Usually, the application of this kind of utterance refers to the user giving his/her name for a given service (e.g. banking service, air travel reservation). When dealing with not common or foreign names, it's usual to spell them. 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**. 5 utterances of different sentences per speaker (Ex.: "Leões verdes jorram dos porões de Java."). 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. The idea of having the speech files associated with these uterances is the possibility of making subjective tests on low bit rate speech coding algoritnms.

- **Sentences for prosodic study**. 4 to 8 utterances of different sentences per speaker.
  1) "Eu vejo o mar"
  2) "Eu vejo o mar azul"
  3) "O mar azul é o que eu vejo."
  4) "Eu vejo que você quer ir ao mar."

  The sentences exemplified 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"

C. Task assignment

For each participant a task of recording utterances of 20 speakers was assigned. According to their possibilities, it is allowed (and desirable) the contribution of other sets of 20 speakers.

For each set of speakers, a different set of material was prepared, to maximize the linguistic coverage of the database.

D. Recordings

Contributors were suggested to perform the recordings in an office environment (low noise environment), using a good quality dynamic microphone. The sampling frequency was selected at 22.05 kHz, the speech signal is digitized with linear 16 bits A/D converter using any available audio card and the

files stored in standard Windows PCM Wave format (.wav). Each recording section was designed to last around 20 minutes.

III. 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 from a given region, in order to balance the database.

Until the moment, utterances from 72 speakers were collected, from some regions of the country, as shown in Figure 1.

Fig. 1. Geographical distribution of recorded speakers until the moment.

As can be seen in Figure 2, there is a great concentration of speakers in the states of São Paulo (SP) and Minas Gerais (MG). We hope that in the near future, with more collaborators, this scenario changes to a more balanced distribution.

The speakers ages range from 17 to 59 years, and the distribution is shown in Figure 3.

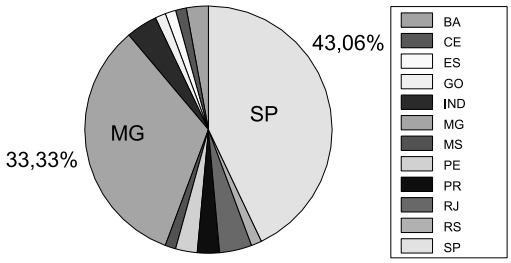

Fig. 2. Relative amount of speakers per state.

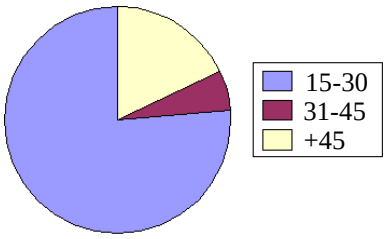

Fig. 3. Age distribution of recorded speakers until the moment.

Finally, all the speakers have completed at least the high school. The distribution of this variable is shown in Figure 4.

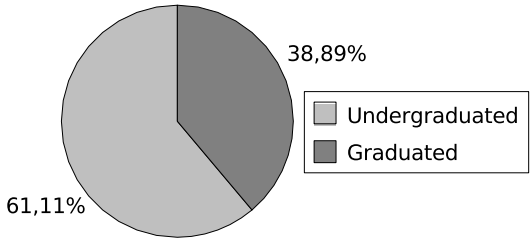

Fig. 4. Education level of recorded speakers.

IV. Phonetic transcription

The ideal would be that the utterances could be manually transcribed by a linguistic specialists team. But this involves costs and a lot of time and, at the moment, due to lack of financial resources, this task will be carried out automatically, using an ortographic-to-phonetic transcriber software (*ortofon3*) developed by LAFAPE/IEL/UNICAMP [6]. This software makes a wide transcription using 51 phonological symbols. Although it was primarily designed to provide the transcription of isolated words, some rules are now being implemented in order to take into account the coarticulation between adjacent words.

The phonetic units used by the *ortofon3* system are listed in Table I.

V. Drawbacks

Everyone that has been involved in the construction of a large speech corpora knows that the most difficult part is the recruitment of speakers (see for example [7]): from the initial contact to the completion of the recordings, most of candidates are missing.

Furthermore, although all speech scientists agree that large speech databases are necessary for the evolution of speech technology, the great majority of them don't want to be involved in such projects, due to the great amount of work involved.

At the beginning of the project, several researchers were contacted to help in this effort, but only two of them actually contributed with some speakers. The conclusion is that maybe there must be an additional incentive for the researchers to participate in the project.

VI. Conclusions and Future Work

In this work, a initiative to build a large speech corpora for the Brazilian Portuguese was presented. The approach is based on a collaborative effort from the researchers around the country to collect utterances from speakers in their regions.

In theory, this approach would serve as a way to quickly construct a relatively large speech corpora with low cost. However, the researchers didn't participate actively in the project, and only 72 speakers were recorded.

Others researchers are being contacted at this time, and maybe, in the near future, more contributors would join this effort.

Manual transcription of recorded utterances are being considered also.

The database is publicly available. For informations, please send an e-mail for ynoguti@inatel.br.

Referências

[1] http://www.ldc.upenn.edu (accessed in 14/05/2008)

[2] 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).

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

[4] 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).

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

[6] Albano, E. C. and Moreira, A. A., Archisegmented-based letter-to-phone convertion for concatenative speech synthesis in Portuguese, Proceedings ICSLP'96, 1996, v.3, pp. 1708-1711.

[7] Lindberg, B. et. al. Speaker recruitment methods and speaker coverage - experiences from a large multilingual speech database collection. Proceedings of the ICSLP 1998.

TABELA I

List of phones used by the ortofon system [6] for the phonetic transcriptions of the utterances.

Phone	Example
a	**a** ç *a* frão, p *a* ta, *a* ta
A	pat *a*, at *a*
e	*e* l *e* vador
E	lent *e* , el *e*
eh	p *e* le, f *e* sta
i	s *i* no, part *i* , ca *í*
I	fu *i* , ca *i*
o	b *o* lo, *o* vo
O	bol *o* , ov *o*
oh	b *o* la
u	l *u* a
U	canto *u*, glób *u* lo
aN	maç *ã*, pl *an* ta
AN	ím *ã*
eN	s *en* ta
EN	híf *em*
iN	p *in* to
IN	ínter *im*
oN	s *om* bra, t *on* ta
ON	mórm *on*
uN	um, m *un* do
UN	fór *um*
b	*b* ela
B	su *b* mete, o *b* soleto
d	*d* a *d* o, *d* ia
D	a *d* ministrador
f	*f* eira
g	*g* orila
G	co *g* nome, ma *g* neto
zh	*j* iló, ca *j* u
k	*c* achoeira, *c* asa
K	te *c* nologia, aspe *c* to
Ks	fi *x* o, tá *x* i, inde *x* ar
l	*l* eão
L	p *l* anta, c *l* aro
lh	*lh* ama, ca *lh* a
m	*m* ontanha
M	*m* nemônico, a *m* nésia
n	*n* évoa
nh	i *nh* ame, ma *nh* ã
p	*p* oente, *p* ata
P	ade *p* to, sino *p* se
r	ce *rr* ado, ca *rr* o, *r* ato
R	ce *r* a, ca *r* o, ca *r* ta, ama *r*
s	*s* apo
S	e *s* tar, casa *s*
t	*t* empes *t* ade, *t* ia,
v	*v* erão
sh	*ch* ave, li *x* o
z	*z* abumba, ca *s* a