New resources for Brazilian Portuguese: Results for grapheme-to-phoneme and phone classification

Conference Paper · October 2006

DOI: 10.1109/ITS.2006.4433322 · Source: IEEE Xplore

Luiz Alberto Novaes Baptista

1 PUBLICATION **12** CITATIONS

SEE PROFILE

Tales Imbiriba

Northeastern University

95 PUBLICATIONS **1,194** CITATIONS

SEE PROFILE

Aldebaro Klautau

Federal University of Pará

253 PUBLICATIONS **3,512** CITATIONS

SEE PROFILE

New Resources for Brazilian Portuguese: Results for Grapheme-to-Phoneme and Phone Classification

Chadia Hosn, Luiz Alberto Baptista, Tales Imbiriba and Aldebaro Klautau

Abstract— **Speech processing is a data-driven technology that relies on public corpora and associated resources. In contrast to languages such as English, there are few resources for Brazilian Portuguese (BP). Consequently, there are no publicly available scripts to design baseline BP systems. This work discusses some efforts towards decreasing this gap and presents results for two speech processing tasks for BP: phone classification and grapheme to phoneme (G2P) conversion. The former task used hidden Markov models to classify phones from the Spoltech and TIMIT corpora. The G2P module adopted machine learning methods such as decision trees and was tested on a new BP pronunciation dictionary and the following languages: British English, American English and French.**

Index Terms— **Grapheme-to-phoneme, letter-to-sound, decision trees, phone classification, hidden Markov models.**

I. Introduction

Speech processing is a data-driven technology and researchers rely on public corpora and other resources to expand the state-of-art frontier. In contrast to languages such as English, there are few resources for BP. This can be noted by browsing the catalog of corpora distributed by the Linguistic Data Consortium [1]. Consequently, there are no publicly available scripts (or software *recipes*) to design *baseline* systems, which considerably speedup the learning process. For example, the software HTK has a complete procedure to design a baseline system for the Resource Management corpus [2]. The international community also benefits from programs that promote cooperation or even open competition among research groups.

In summary, there are three major factors that drive the speech processing community:

- Public **corpora** distributed by institutions such as the Linguistic Data Consortium [1] and CSLU / OGI [3];
- Public and in some cases, free **software** with recipes for building baseline systems: HTK [2] (in C language), Sphinx 4 [4] (Java), ISIP [5] (C++), Festival [6], etc.;
- **Evaluations** organized for specific tasks, such as the ones for speech and speaker recognition by NIST [7] and the recent Pronasyl Letter-to-Phoneme Conversion Challenge [8].

In Brazil, the speech area has not been catalyzed by such factors. The research in BP has been conducted in a relatively isolated way and it is difficult to reproduce the results among different sites. For example, considering the *large vocabulary continuous speech recognition* (LVCSR) [9] task, with few exceptions the Brazilian community does not get involved in the tasks faced by the international community (e.g., [10]) nor concentrates in the development of resources specific to the BP language. In contrast, the Brazilian Natural Language Processing (NLP) community, strongly focus on the development of BP resources [11] and has been obtaining funding from private and governmental institutions.

This work discusses some efforts within the *FalaBrasil* initiative [12]. The overall goal is to develop and deploy resources and software for BP, aiming to establish baseline systems and allow for reproducing results across different sites. More specifically, the work presents resources and results for phone classification and grapheme to phoneme (G2P) conversion (also called letter-to-phoneme and letter-to-sound). The former task used hidden Markov models (HMM) to classify phones from the Spoltech (BP) and TIMIT (English) corpora. The G2P module used machine learning methods such as decision trees and was tested on a new BP pronunciation dictionary. The G2P module was validated with the following languages: British English, American English, French and BP. The paper also presents scripts for designing a baseline system for speaker recognition based on the IME corpus.

Throughout the paper, phones and phonemes are distinguished as follows (see, e.g., [13] for more details). A *phoneme* is the basic theoretical unit for describing how speech conveys linguistic meaning. The variants of the phonemes, i.e., the actual sounds that are produced in speaking are called *allophones*. In linguistics, allophones are considered to be generated as a result of applying phonological rules to the underlying phonemes [13]. On the other hand, in most engineering applications there are no concerns on keeping a strict relationship between phonemes and sounds through phonological rules, and the (lousy) term *phone* substitutes allophones.

This paper is organized as follows. In Section II the new pronunciation dictionary is presented and results for grapheme-to-phoneme are discussed. Section III deals with the phone classification problem, constrasting the TIMIT and Spoltech corpora. A set of scripts for creating a speaker recognition baseline system is then briefly discussed in Section IV, which is followed by the conclusions.

II. Grapheme-to-Phoneme

An important prerequisite for services involving speech recognition and/or speech synthesis is information about the correspondence between the orthography and the pronunciation(s). For example, in order to develop LVCSR for BP, one

The authors are with the Signal Processing Laboratory - LaPS, Federal University of Pará, C.P. 479, CEP: 66075-110, Belém-PA, Brazil. email: {chadia,novaes,tales,aldebaro}@ufpa.br.

TABLE I

NUMBER OF PHONEMES AND WORDS IN THE TRAINING AND TEST SETS OF EACH DICTIONARY.

		Phonemes	Words
Nettalk	Train	88,112	12,000
American English	Test	58,831	8,008
Brulex	Train	141,332	16,500
French	Test	93,921	10,973
Beep 1.0	Train	1,413,479	154,200
British English	Test	941,369	102,780
UFPAdic 1.0	Train	64,799	8,300
Brazilian Portuguese	Test	27,497	3,527

needs a *pronunciation* (or phonetic) dictionary, which maps each word in the lexicon to one or more phonetic transcriptions (pronunciation).

The G2P systems can be organized in *rule-based* and *data-driven*. In [14], the rule-based approach for letter-to-phone conversion was compared with two self-learning methods, one based on a multi-layered neural network and another based on table look-up. More recently, rule-based, data-driven and hybrid approaches have been implemented as *finite state transducers* [15]. Of importance is the transformation-based learning (TBL) discussed in [16]. Some studies have attempted to use learning algorithms to incorporate pronunciation by analogy [17], [18]. Specifically for Portuguese, important work has been developed at INESC, Portugal [19] and Unicamp, Brazil [20], both aiming the development of a G2P module to be part of a *text-to-speech* system.[1]

A. The UFPAdic dictionary

A contribution of this work is the release of a hand-labeled pronunciation dictionary with 11,827 words in BP. The phonetic transcriptions adopts the SAMPA alphabet. The transcriptions were obtained from several sources. Most of them were obtained from an electronic version of a commercial dictionary. The version 1.0 of the dictionary can be found at [12].

Some other publicly available pronunciation dictionaries are:

- NETtalk, 20,008 words, American English [21];
- Brulex, 27,473 words, French [22];
- Beep 1.0, 256,980 words, British English available as file beep.tar.gz from [23].

A comparison of the size of these dictionaries is provided in Table I. It can be noted that UFPAdic version 1.0 is the smallest among them, but its size is comparable to the dictionaries used in other recent studies (e.g., [16]).

The next subsections will describe how to generate larger pronunciation dictionaries using machine learning. It is important to note that the experiments conducted in this work used the hand-labeled transcriptions, which provide the “ground truth”. If one tries to evaluate G2P using automatically-generated transcriptions, the results will be biased towards imitating the algorithm used to generate the transcriptions.

TABLE II

ERROR RATES FOR THE UFPADIC DICTIONARY AS A FUNCTION OF THE CONTEXT SPAN.

UFPAdic1.0 - Error rates in %				
	Phone		Word	
	J48	Naive Bayes	J48	Naive Bayes
Context=1	2.77	**5.53**	18.66	**33.82**
Context=3	2.77	8.55	11.40	48.20
Context=5	**1.60**	12.52	**10.92**	62.63

B. G2P Modules based on Machine Learning

This work adopts a two-step self learning approach that automatically derives algorithms for G2P conversion from training data. In the first step, corresponding grapheme and phoneme strings in the training data are aligned according to the method described in [24]. Lexicon alignment is an important and critical step of the whole training scheme of such G2P systems, as it builds up the data on which the learning methods extracts the transcription rules. This alignment can be done manually, but this is time-consuming, error-prone, and limits the size of datasets that can be used for training. In the second step, the Weka machine learning tool [25] is used to build a classifier. In this work only decision trees (Weka’s J4.8 algorithm) and Naive Bayes were used, but there are many other options in Weka [26].

Transcription rules are extracted in an automatic way by machine learning techniques. In particular, inductive learning techniques are able to find out the common characteristics in the data and generalize them. A sliding grapheme window moves over the word. The window takes into account a subsequence of the word including a focus (the central grapheme to be transcribed) and 1, 3 or 5 symmetric contexts, indicating the left and right context of the focus, respectively.

Using a context of 1, means here that a sliding window passes three (1 left + 1 focus + 1 right) graphemes to the classifier and obtains the phoneme , that could be a null symbol, corresponding to the focus grapheme.

C. Simulation Results

The dictionaries were split into two disjoint sets, as indicated by I, which were used for training and testing the algorithms. After automatically aligning graphemes and phonemes [24], a file in Weka’s format was created for the three possible symmetric context spans: 1, 3 and 5. Cross-validation was used to perform model selection (pick the best parameters of a given classifier). For example, the decision tree requires specifying the minimum number of instances per leaf and such number was automatically found through 3-fold cross-validation on the train set. This way the test set is never used during the training stage.

Tables II to V present the results obtained for the four dictionaries. The Naive Bayes classifier was used only to indicate the level of performance that can be achieved with such simple algorithm.

The obtained results are compatible to the ones found in the literature for Beep, Brulex and Nettalk. The results for Brazilian Portuguese compare well with the ones for other

[1] To the best of the author’s knowledge, the resources for Portuguese G2P are not publicly available.

TABLE III

ERROR RATES FOR THE NETTALK DICTIONARY AS A FUNCTION OF THE CONTEXT SPAN.

Nettalk - Error rates in %				
	Phone		Word	
	J48	Naive Bayes	J48	Naive Bayes
Context=1	16.65	23.41	72.18	83.85
Context=3	**10.60**	**22.41**	**52.45**	**81.38**
Context=5	10.94	26.58	53.66	84.62

TABLE IV

ERROR RATES FOR THE BRULEX DICTIONARY AS A FUNCTION OF THE CONTEXT SPAN.

Brulex - Error rates in %				
	Phone		Word	
	J48	Naive Bayes	J48	Naive Bayes
Context=1	7.81	16.09	48.57	77.31
Context=3	2.12	**14.75**	13.19	**71.30**
Context=5	**2.06**	19.19	**12.56**	81.30

languages, given that written Portuguese is more "homogeneous" in terms of pronunciation than American English, for example. However, such experiments are rather simple. More elaborated simulations take in account syllables, stress, etc., and will be incorporated into the system to be presented in the Pronasyl Challenge [8].

The next section discusses a different problem. While G2P systems deal only with text (graphemes and phonemes). Phone classification can be seen as a simplified acoustic modeling problem [9] and deals with digitized speech files.

III. PHONE CLASSIFICATION USING SPOLTECH AND TIMIT

Discussing TIMIT and Spoltech together allows for drawing comparisons that can guide future designs of resources for BP. There are several ways of conducting experiments based on these two corpora. One can use the provided phonetic transcriptions and perform *segmental classification*, for example, with HMMs. Another possibility is to convert the phone segments into a fixed-length vector, which leads to the *conventional classification* problem. A third scenario corresponds to performing *phone recognition*, where each phone is treated as a word in continuous speech recognition. In other words, classification means that the phone boundaries are available during the test stage. This is the setup used in this work. The next two subsections will briefly describe each corpora.

TABLE V

ERROR RATES FOR THE BEEP DICTIONARY AS A FUNCTION OF THE CONTEXT SPAN.

BEEP1.0 - Error rates in %				
	Phone		Word	
	J48	Naive Bayes	J48	Naive Bayes
Context=1	14.65	23.82	72.92	89.81
Context=3	**5.66**	**23.70**	**35.95**	**89.70**
Context=5	-	26.71	-	91.99

A. The Spoltech Corpus

The Spoltech [27] was created by UFRGS, Brazil and OGI, USA. Its construction was funded by CNPq, Brazil and NSF, USA.[2] The corpus is currently distributed by LDC [1] (LDC2006S16) and OGI [3] (the version used in this work is the latter). It consists of waveform speech files (WAV), orthographic (TXT) and phonetic transcriptions (PHN). Although useful, Spoltech has several problems. Some WAV files do not have their corresponding TXT and PHN files, and vice-versa. Another problematic aspect is that the phonetic and orthographic transcriptions have many errors.

For this work, a pre-processing stage tried to find pairs of valid WAV and PHN files. It found only 5,479 pairs (files with both PHN and WAV). These files were split into two disjoint sets with 4,105 and 1,374 files, for the training and test sets, respectively. Care was exercised to avoid having a given speaker participating in both sets. The work considered only phonetic transcriptions, for which one can find 183 different symbols in the 5,479 files. These symbols are part of the Worldbet phonetic alphabet, which is adopted at OGI. Many of these symbols have only few occurrences and some of them are not valid Worldbet symbols (probably typos). Because of that, a histogram of the phones was calculated and only the 64 most frequent phones were used in this work. The speech was resampled to 16 kHz, the same sampling frequency used by TIMIT (Spoltech uses 16 bits per sample).

B. The TIMIT Corpus

The TIMIT acoustic-phonetic continuous speech corpus is the most popular among the corpora distributed by the Linguistic Data Consortium. It was recorded at Texas Instruments (TI), transcribed at the Massachusetts Institute of Technology (MIT), verified and prepared for CD-ROM production by the National Institute of Standards and Technology (NIST). The corpus includes the speech waveform files with corresponding time-aligned orthographic and phonetic transcriptions. TIMIT is a valuable resource: it took 100 to 1000 hours of work to transcribe each hour of speech, and the project cost over 1 million dollars [28].

TIMIT was annotated using a relatively narrow phonetic transcription. In fact, in most speech recognition experiments the 61 TIMIT symbols are collapsed into 39 classes for scoring purposes [29]. The more detailed the transcriptions are, the smaller the percentage of agreement among phoneticians. When developing TIMIT, five phoneticians agreed on 75% to 80% of the cases [28]. This clearly indicates that the number of symbols used in Spoltech is too large.

TIMIT contains speech from 630 speakers representing 8 major dialect divisions of American English, each speaking 10 phonetically-rich sentences. There are 438 male speakers and 192 female speakers.

Figure 1 shows the number of occurrences of each phone in the training set. The most frequent phone is [sil], corresponding to silence in the begin and end of each sentence, and the

[2] Through "Advancing Human Language Technology in Brazil and the United States Through Collaborative Research on Portuguese Spoken Language Systems," U$199,481, granted to R. Cole.

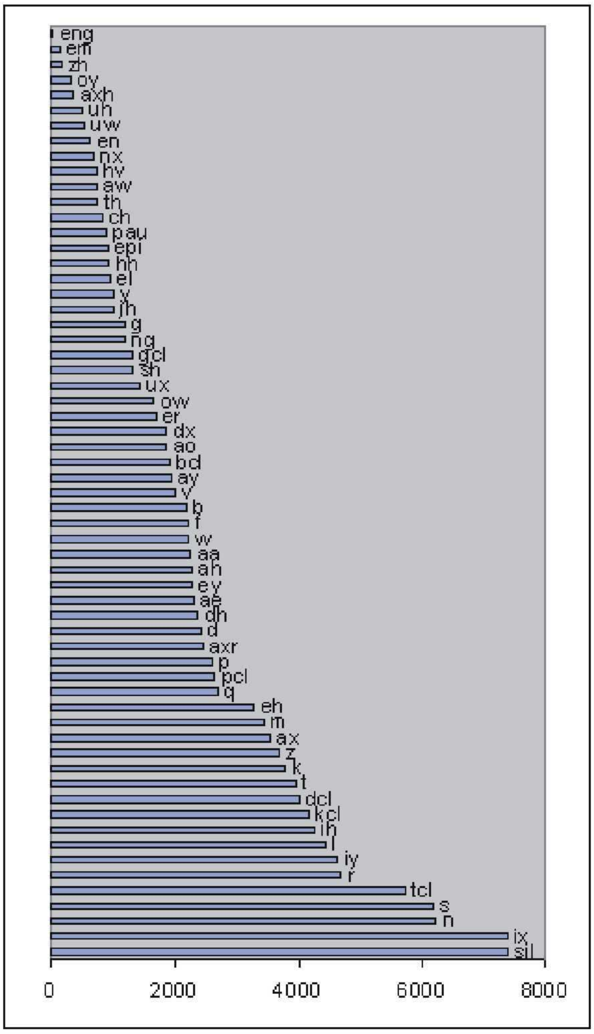

Fig. 1. Histogram of the 61 phones in TIMIT training set.

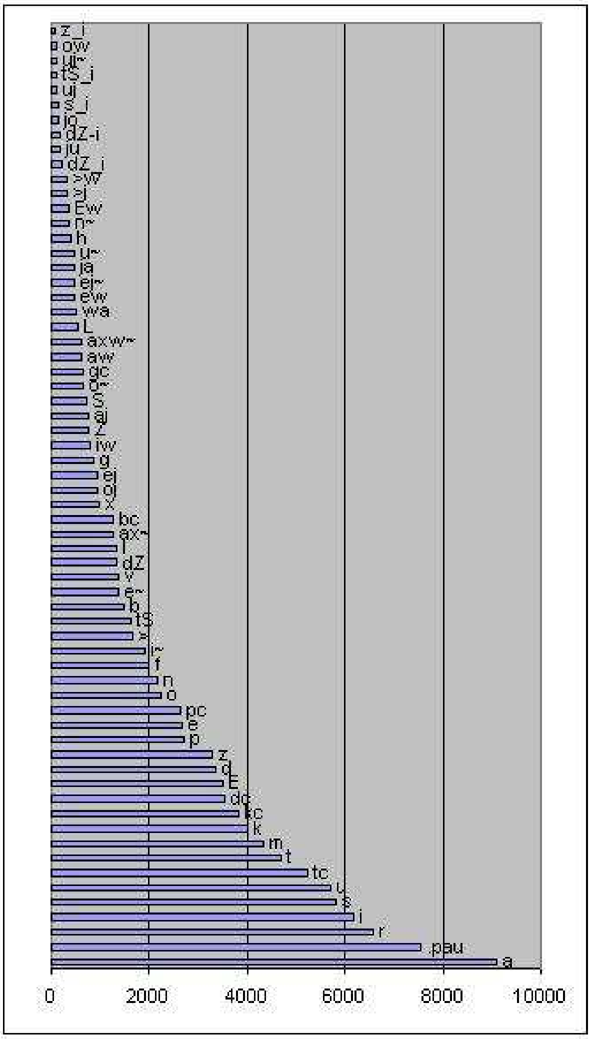

Fig. 2. Histogram of the 64 most frequent phones in Spoltech (5,479 files).

second is [ix]. The least frequent phone is [eng], with only 26 occurrences. Figure 2 shows a similar plot, but for the 5,479 files of Spoltech, in which the most frequent phone is [a].

C. Simulation Results

The experiments used popular front ends: mel-frequency cepstrum coefficients (MFCC) [30], perceptual linear prediction (PLP) coefficients [31] and RASTA [32]. For each phone an HMM with 3 states, left-right topology with no-skips was estimated, using the Baum-Welch algorithm [9]. The results for Spoltech are shown in Figure 3. The identifier of each front end should be interpreted as follows: a suffix "e" indicates the energy was incorporated to the static coefficients, "d" and "a" corresponds to first and second derivatives, respectively, while "c" corresponds to *cepstrum mean subtraction*. The number before the "w" informs the total number of parameters per frame, while the numbers after "w" and "s" indicates the window length and shift, in samples, respectively. For example, mfccedac39w400s160 represents the popular front end with 13 static coefficients (13 MFCC coefficients and energy), and their first and second derivatives. The window length is 400 samples and the shift is 160 samples, in this case.

Figures 4 and 5 shows the misclassification error for the phones with lowest and highest rates, respectively, when assuming the mfccedac39w512s160 front end. Most of the phones in Figure 5 do not have many occurrences in the training sets and should be potentially grouped with others, as done for TIMIT [29].

Figure 6 shows a comparison between the results obtained with TIMIT and Spoltech for the mfccedac39w512s160 front end, as the number of Gaussians per HMM state is increased from 1 to 10. It can be seen that Spoltech leads to a higher error, which can be caused by the many errors in its transcriptions and a less controlled recording environment. Also, Figure 6 illustrates that the HMM classifier is *overfitting* the Spoltech training set, while this does not happen so severely for TIMIT. This situation points towards a redefinition of the size of the training and test sets of the Spoltech corpus.

The next section presents a publicly available baseline system for speaker recognition, which is another contribution of this work.

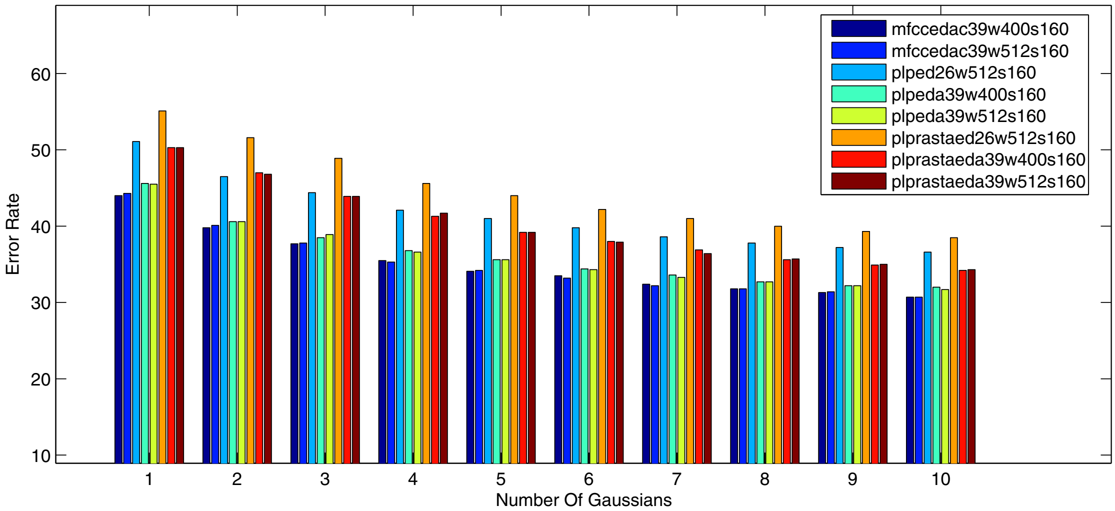

Fig. 3. Error rate for the Spoltech test set using several front ends and number of Gaussians per HMM state.

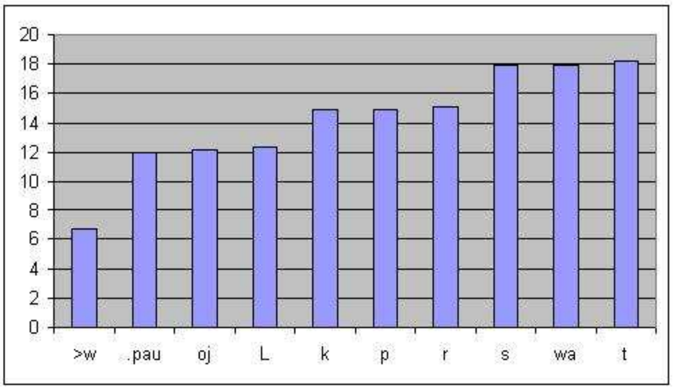

Fig. 4. Phones with lowest error rates for Spoltech.

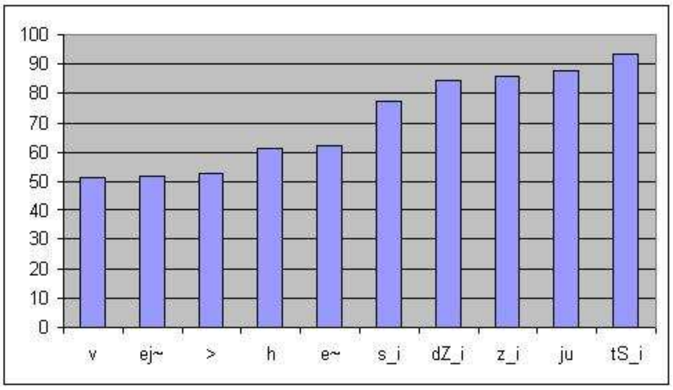

Fig. 5. Phones with highest error rates for Spoltech.

IV. Speaker Recognition with the IME Corpus

The IME 2002 corpus is a Brazilian Portuguese corpus for speaker recognition that has been made available on the web [33]. The IME corpus is composed by 468 files, corresponding to 21.9 hours of recorded signal. For the sake of comparison, the popular NIST-2001 [7] is composed by 2350 (shorter) files, which correspond to 26.4 hours of speech. The utterances in the IME corpus were collected from cellular and wired phone calls made by 75 speakers.

The contribution presented in this work is to provide a set of scripts to design a Gaussians mixture model (GMM)-based system for the the IME corpus. The scripts must be downloaded from [12] and placed in the same directory of the unzipped corpus. HTK must be installed and in the PATH. For the sake of simplicity, only the files corresponding to the wired calls are used. Running the scripts will perform the following tasks:

- organize the files and their names, discarding the wired phone files that are too short or corrupted. This leads to 75 speakers, each one with two files: for training and test;
- HCopy (the HTK front end) is invoked to extract the features, which are by default plpeda39w640s320;
- A silence extraction is performed based on two Gaussians estimated in an unsupervised way and using only the trainig files;
- The GMM training starts by dividing the files in two groups of almost the same size A and B (37 and 38 speakers, respectively). Files from group A use speakers from B as impostors when designing the *universal background model* (UBM). During the test stage, for each speaker, the impostor are the other speakers from its own group. This procedure assures the models trained for each espeaker have never "seen" the test impostors, which leads to a more realistic result.

The final result for this baseline, which adopts 128 Gaussians per mixture is an equal-error rate (EER) of 2.7% and is competitive with previously published results.

V. Conclusions

This paper presented some resources and software for BP. The goal was to establish baseline systems and allow for

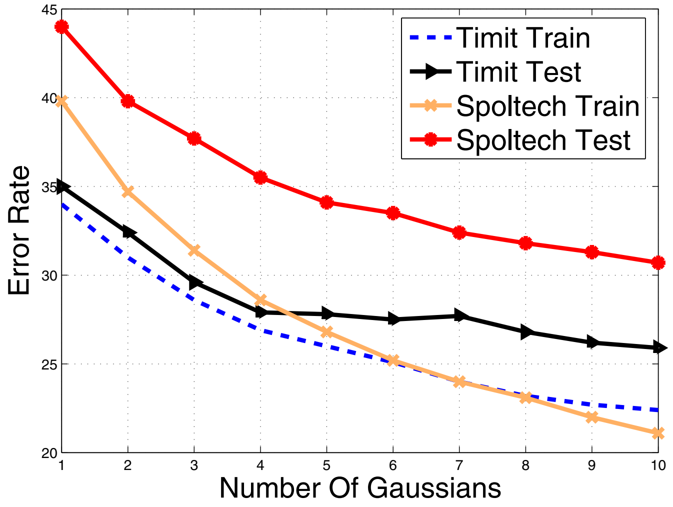

Fig. 6. A comparison of Spoltech and TIMIT results: error rate for the training and test sets of each corpus.

reproducing results across different sites. Results for phone classification and grapheme to phoneme conversion were presented. The G2P experiments used decision trees, Naive Bayes and were validated with the following languages: Brazilian Portuguese, British English, American English and French. The phone classification experiments were based on three popular front ends: MFCC, PLP and RASTA, and constrasted the Spoltech and TIMIT corpora. The paper also discussed publicly available scripts for designing a speaker recognition baseline system based on GMMs.

The overall philosophy behind this work is to promote the distribution of resources and baseline scripts for BP. It is clear from the comparisons between TIMIT and Spoltech, for example, that there is a long way ahead, but the strategy is to emphasize the creation of necessary resources even if they are not of "high quality". For example, attempts to build a comprehensive speech corpus for BP LVCSR face scarse funding opportunities. The authors have been developing a "cheap home-made" corpus by digitalizing and transcribing at the word-level some TV shows. This corpus will have a relatively low signal-to-noise ratio, a non-balanced coverage or Portuguese phonemes, dialects, etc., and will not be adequate for developing commercial applications. However, any speech corpus that allows to put together a *trigram language model*, *cross-word triphone* models, a pronounciation dictionary and a faster decoding scheme to build a LVCSR has importance given the lack of resources for BP.

ACKNOWLEDGEMENTS

This work was partially supported by SECTAM / Pará and CAPES.

REFERENCES

[1] "http://www.ldc.upenn.edu." Visited in March, 2005.
[2] "http://htk.eng.ac.uk," Visited in March, 2006.
[3] "http://cslu.cse.ogi.edu/corpora," Visited in March, 2006.
[4] "http://cmusphinx.sourceforge.net/sphinx4/," Visited in March, 2006.
[5] "http://www.isip.msstate.edu."
[6] "http://www.cstr.ed.ac.uk/projects/festival," Visited in April, 2006.
[7] "http://www.nist.gov/speech," Visited in April, 2006.
[8] "http://www.pascal-network.org/challenges/pronalsyl/," Visited in April, 2006.
[9] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing*. Prentice-Hall, 2001.
[10] "http://www.nist.gov/speech/publications/tw00," Visited in April, 2006.
[11] "Graca Nunes, invited talk at TIL'2005, http://www.unisinos.br/congresso/sbc2005/?sessao=til," 2005.
[12] "http://www.laps.ufpa.br/falabrasil," Visited in April, 2006.
[13] P. Ladefoged, *A Course in Phonetics*, 4th ed. Harcourt Brace, 2001.
[14] I. Trancoso, M. Viana, and F. Silva, "On the pronunciation of common lexica and proper names in European Portuguese," in *2nd Onomastica Res. Colloq*, 1994.
[15] G. Bouma, "Comparison of two tree-structured approaches for grapheme-to-phoneme conversion, http://citeseer.ist.psu.edu/article/bouma00finite.html," in *1st Meeting of the North-American Chapter of the Association for Computational Linguistics, Seattle*, 2000.
[16] A. Teixeira, C. Oliveira, and L. Moutinho, "On the use of machine learning and syllable information in european portuguese grapheme-phone conversion," in *7th Workshop on Computational Processing of Written and Spoken Portuguese (to be presented) - Itatiaia, Brazil*, 2006.
[17] M. J. Dedina and H. C. Nusbaum, "PRONOUNCE: A program for pronunciation by analogy," *Computer Speech and Language*, pp. 5:55–64, 1991.
[18] R. I. Damper, Y. Marchand, J. Marsters, and A. Bazin, "Can syllabification improve pronunciation by analogy of English?" *Natural Language Engineering*, pp. 1–25, 2005.
[19] N. J. Mamede, J. Baptista, I. Trancoso, and M. das Graças Volpe Nunes, "Computational processing of the portuguese language, 6th international workshop, propor 2003, faro, portugal, june 26-27, 2003. proceedings," in *PROPOR*. Springer, 2003, pp. 23–30.
[20] P. Barbosa, F. Violaro, E. Albano, F. Simões, P. Aquino, S. Madureira, and E. Françozo, "Aiuruetê: a high-quality concatenative text-to-speech system for brazilian portuguese with demisyllabic analysis-based units and hierarchical model of rhythm production," in *Proceedings of the Eurospeech'99, Budapest, Hungary*, 1999, pp. 2059–2062.
[21] T. J. Sejnowski and C. R. Rosenberg, "Parallel networks that learn to pronounce english text," *Complex Systems*, vol. vol. 1, pp. 145–168, 1987.
[22] A. Content, P. Mousty, and M. Radeau, "Brulex: Une base de données lexicales informatisée pour le français éecrit et parlé," *L'Année Psychologique*, pp. 551–566, 1990.
[23] "ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries," Visited in February, 2006.
[24] R. I. Damper, Y. Marchand, J. Marsters, and A. Bazin, "Aligning letters and phonemes for speech synthesis," in *5th ISCA Speech Synthesis Workshop - Pittsburgh*, 2004, pp. 209–214.
[25] "http://www.cs.waikato.ac.nz/ml/weka," Visited in April, 2006.
[26] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.
[27] "Advancing human language technology in Brazil and the United states through collaborative research on portuguese spoken language systems," Federal University of Rio Grande do Sul, University of Caxias do Sul, Colorado University, and Oregon Graduate Institute., 2001.
[28] J. Picone, "Talk at SRSTW'02, http://www.isip.msstate.edu," 2002. [Online]. Available: www.isip.msstate.edu
[29] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–8, Nov. 1989.
[30] S. Davis and P. Merlmestein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on ASSP*, pp. 357–366, Aug. 1980.
[31] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–52, Apr. 1990.
[32] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–89, Oct. 1994.
[33] "http://larso.ime.eb.br/tools.htm," Visited in April, 2006.