Recognition of Classroom Lectures in European Portuguese

Isabel Trancoso[(1)], *Ricardo Nunes*[(2)], *Luís Neves*[(2)],
Céu Viana[(3)], *Helena Moniz*[(3)], *Diamantino Caseiro*[(1)], *Ana Isabel Mata*[(3)]

[(1)] L^{2}F INESC-ID/IST, [(2)] L^{2}F INESC-ID, [(3)] CLUL
Lisbon, Portugal
Isabel.Trancoso@inesc-id.pt

10.21437/Interspeech.2006-88

Abstract

Classroom lectures may be very challenging for automatic speech recognizers, because the vocabulary may be very specific and the speaking style very spontaneous. Our first experiments using a recognizer trained for Broadcast News resulted in word error rates near 60%, clearly confirming the need for adaptation to the specific topic of the lectures, on one hand, and for better strategies for handling spontaneous speech. This paper describes our efforts in these two directions: the different domain adaptation steps that lowered the error rate to 45%, with very little transcribed adaptation material, and the exploratory study of spontaneous speech phenomena in European Portuguese, namely concerning filled pauses.
Index Terms: spontaneous speech recognition, Portuguese.

1. Introduction

The goal of the national project LECTRA is the production of multimedia lecture contents for e-learning applications. Nowadays, the availability on the web of text materials from University courses is an increasingly more frequent situation, namely in technical courses. Video recording of classes for distance learning is also a more and more frequent possibility. Our contribution to these contents (text books, slides, exercises, videos, etc.) will be to add, for each recorded video, the synchronized lecture transcription. We believe that this synchronized transcription may be specially important for hearing-impaired students. Moreover, the synchronized transcription opens the possibility of browsing through the audio recordings and the corresponding course material.

From a research point of view, the lecture transcription domain is very challenging, mainly due to the fact that we are dealing with spontaneous speech, characterized by strong coarticulation effects, non-grammatical constructions, hesitations, repetitions, filled pauses, etc. [1]. Moreover, for e-learning purposes, a plain transcription may not be intelligible enough, and may need "enrichment" with punctuation, capitalization, marking of disfluencies, etc. The fact that the lectures are taught in European Portuguese (EP) adds further challenges, because very often the text material for the course is in English (in fact, that is probably the most common scenario in computer science and electrical engineering courses in the Technical University of Lisbon). Hence we have to deal with the problem of very little text material to do domain adaptation, a problem which we could not yet counterpart with a large number of manually transcribed lectures.

Lecture transcription has been the target of much bigger research projects such as the Japanese project described in [2], the European project CHIL (Computers In The Human Communication Loop) [3], and the American iCampus Spoken Lecture Processing project [4]. In some of these projects, the concept of lecture is different. Our classroom lectures are almost 90 minutes long, and they involve mostly a single speaker (the teacher) who tried to create a very informal atmosphere. This contrasts with the 20 minute seminars used in [3], where a more prepared speech can often be found. Unfortunately, as explained, the amount of material for adapting our recognizer to the lecture domain is also very different from the very large amounts collected in other projects.

Section 2 summarizes the first task of the project - corpus collection, which started with two very different courses. Section 3 describes our baseline recognizer and the corresponding results. Section 4 is dedicated to the adaptation of the recognizer modules to the domain of the 2 courses. The following Section reports on different strategies for vocabulary reduction. Section 6 summarizes our exploratory efforts in terms of dealing with fillers and edit disfluencies [5], in particular with filled pauses.

2. Corpora Collection

Two very different courses have been selected for our pilot study: one entitled "Economic Theory I" (ETI) and another one entitled "Production of Multimedia Contents" (PMC). The ETI course (17 classes) and the first 6 classes of the PMC course were recorded with a lapel microphone. The last part of the PMC course (14 classes) was recorded with a head-mounted microphone. The two recording types presented specific problems. The lapel microphone proved inadequate for this type of recordings given the very high frequency of head turning of the teacher (towards the screen or the white board) that caused very audible intensity fluctuations. The use of the head-mounted microphone clearly improved the audio quality. However, 11% of the recordings were saturated, due to the increase of the recording sound level during the students' questions, in the segments that were recorded right after them. The classes had variable duration, ranging from 40 to 90 minutes. Both professors were male speakers, with Lisbon accent. Table 1 shows the duration of the very limited training, development, and test sets selected from different classes, and the number of words in each. Segments from students were not transcribed, as most were not intelligible enough, due to the distance to the microphone. These segments correspond to around 170s in each of the test sets. The ETI course included very frequent references to mathematical variables and expressions (e.g. $P1'$). The PMC course, on the other hand, included much computer jargon, usually derived from English (e.g. *e-mail*, *software*), and a heavy use of spelt or partially spelt acronyms (e.g. *http*). The computer jargon was generally pronounced very close to their English pronunciation, even including xenophones that are not part of the phone inventory for European

Table 1: Duration and number of words in each manually transcribed set.

	ETI			*PMC*		
	Train.	*Dev.*	*Test*	*Train.*	*Dev.*	*Test*
Duration [min.]	62	46	36	73	52	42
# words	8k	7k	5k	10k	8k	6k

Portuguese [6]. The percentage of technical terms in English in the PMC test corpus was 2.1%, a fact that will affect the lexical model, as we shall see below.

In order to adapt the language models to the domain of each course, we tried to get additional course materials. For the ETI course, we had a textbook and viewgraphs. Given the extension of the text book (452k words), we discarded the viewgraphs. For the PMC course, the textbook was in English. So, in order to train language models in Portuguese we only had viewgraphs (25k words), exam questions (2k words) and student reports (23k words). Viewgraphs are typically characterized by specific grammatical constructions which clearly differentiates this material from other textual sources. By analyzing a small set of sentences from the PMC viewgraphs (around 2k words), the percentage of verbs that was obtained (9.1%) was much smaller then the one observed in a similar set of sentences from PMC reports (17.0%). The percentage of nouns, on the other hand, was much higher (42.2% in viewgraphs vs. 27.1% in reports). This different construction will have an obvious negative impact on the domain adaptation.

3. Baseline Recognizer

Our baseline large vocabulary recognizer was trained for Broadcast News (BN) [7]. It uses hybrid acoustic models that try to combine the temporal modeling capabilities of hidden Markov models with the pattern classification capabilities of MLPs (Multi-Layer Perceptrons). The models have a topology where context-independent phone posterior probabilities are estimated by three MLPs given the acoustic parameters at each frame. The streams of probabilities are then combined using an appropriate algorithm. The MLPs were trained with different feature extraction methods: PLP (Perceptual Linear Prediction), Log-RASTA (log-RelAtive SpecTrAl) and MSG (Modulation SpectroGram). Each MLP classifier incorporates local acoustic context via an input window of 7 frames. The resulting network has a non-linear hidden layer with over 1000 units and 40 softmax output units (38 phones plus silence and breath noises). The vocabulary includes around 57k words. The lexicon includes multiple pronunciations, totaling 66k entries. The language model was created by interpolating a newspaper text language model built from over 400M words with a backoff trigram model using absolute discounting, based on the training set transcriptions of our BN database (45h). The perplexity (PP) is 139.5. For the BN test set corpus, the out-of-vocabulary (OOV) word rate is 1.4%, and the average WER (word error rate) is 31.2% for all conditions, 18.9% for F0 conditions (read speech in studio), and 35.2% for F1 conditions (spontaneous speech).[1]

For the ETI and PMC test sets, this BN recognizer achieved a WER of 56.4% and 63.6%, which was expected, in view of the fact that we are dealing with spontaneous speech recorded in a classroom, with very specific contents, and the above mentioned recording problems. The lack of domain adaptation is specially patent in the high OOV rates and perplexity values obtained for the PMC test set (OOV=3.4%, PP=292.8). For the ETI test set, the values were much lower (OOV=1.6%, PP=175.0).

[1]Computed over the joint evaluation set of the ALERT project, using automatic segmentation into blocks of sentences from the same speaker.

4. Domain Adaptation

The following subsections describe the adaptation stages to the lecture domain of the lexical, language and acoustic models. We tried to make this process as automatic as possible in order to enable the rapid porting to other course domains.

4.1. Lexical Model

The most straightforward way of adapting the lexical model was to select from the text material from the two courses, the vocabulary to be added to the original 57k BN vocabulary. For the ETI course, we selected all the words of the transcribed training material, plus the words of the text book that occurred more than 5 times. That added 325 new words to the original vocabulary. For the PMC course, we selected all the words of the transcribed and text materials, amounting to 3k new words.

Acronyms and foreign words were very frequent among the new entries: 17.5% in the ETI course and 80.4% in the PMC course. Like foreign words, acronyms do not follow the common lexicon rules, even when they are read, and are characterized by a high degree of pronunciation variability among native speakers. Nowadays, we can also find many acronyms in URLs. Hence all these entries were manually transcribed.

Concerning foreign words, we started by automatically detecting them in the new entries, by computing the intersection with an English lexicon of approximately 118k entries.[2] This procedure was followed by phone mapping, excluding xenophones. Multiple pronunciations were often adopted, in order to account for the fact that the phone mapping between the English phone inventory and the Portuguese one may not be unique. For instance, the English phone [θ] can be pronounced either as [s] (closest symbol in terms of pronunciation) or [t] (closest symbol in terms of orthography, since *th* sequences never occur in EP).

The most frequent OOVs of the ETI test set are references to mathematical variables (33.8%) not included in the textbook.

4.2. Language Model

Given the scarcity and inadequacy of written training material for the PMC pilot course, building a language model (LM) on that basis alone would give rise to a very high perplexity (256.1) and OOV rate (8.0%). The best results were hence obtained by interpolating the course LM with the one derived from the BN domain. After several tests, the interpolation coefficient for the new LM was set to around 0.25. The new 3-gram LM was built using the SRILM toolkit [8], with modified Knesser-Ney discounting. Before interpolation, the WER corresponding to this new model was 64.8%. After interpolation, it decreased to 58.7%. The perplexity decreased to 208.6 and the OOV rate to 1.7%.

For the ETI course, the interpolation of the textbook and the transcribed training lecture with the BN model decreased the WER

[2]http://www.speech.cs.cmu.edu/cgi-bin/cmudict

to 54.3%, corresponding to a perplexity of 148.2. The OOV rate was practically the same as with the initial BN model.

4.3. Acoustic Model

Although the transcribed training material was also very scarce, it was worth testing how much one could gain from adapting the acoustic models to the speaker and classroom environment, with just one lecture. We tried adapting the acoustic models without and with language model adaptation. In the first tests, after 3 iterations, the WER was down to 48.0%, for the PMC course and to 45.4% for the ETI course. The WER reduction was hence much more significant than with language model adaptation alone. In the second tests, the WER decreased to 44.8% for the PMC course, and to 44.7% for the ETI course.

5. Vocabulary Reduction

Because of domain mismatch, we expect that a significant number of words in our initial 57k-word BN vocabulary does not contribute to a reduction of the OOV rate, and may even have a negative impact due to acoustic confusability. In an effort to reach a better equilibrium between coverage and acoustic confusability, we investigated the use of smaller vocabularies (30k and 15k), chosen amongst the most frequent entries of the previous 57k vocabulary, excluding foreign celebrity names that were very domain dependent. The OOV rate for the BN test corpus almost duplicates with each reduction (2.4% for 30k, and 5.0% for 15k). OOV values near 5% were considered too high, so further tests with domain adaptation were done using the 30k vocabulary only.

The WER achieved with the 30k vocabulary increased slightly for the BN test set (+1.2%). The interpolation of the corresponding LMs with the ETI and PMC LMs yielded minor changes in WER values (44.8% and 44.5%, respectively). The corresponding perplexity and OOV values were 145.8 and 1.7% for the ETI course, and 197.2 and 2.6% for the PMC one. The best interpolation coefficient for the LMs of the courses increased (0.47 for PCM), reflecting their larger relative weight. When more transcribed material becomes available, it will be worth exploring more elaborate vocabulary selection techniques [4].

In order to compensate the OOV increase caused by the vocabulary reduction, we have also tried to treat clitics as separate unigrams. This was motivated by an analysis of the errors that showed that many correspond to inflected verbal forms (Portuguese verbs typically have above 50 different forms, excluding clitics), and gender and number distinctions in names and adjectives. In the BN domain, 36.6% of the OOV words are verbal forms, and 10.7% are verbal forms with clitics. In the classroom lecture domain, without adaptation, 39.0% of the OOVs are verbal forms, and 6.1% are verbal forms with clitics. Our previous attempts at doing some morphological analysis on verbal forms have not brought any significant improvements [9]. In this work, however, we decided to restrict this analysis to clitics, by treating them as separate unigrams in the lexical and language models.

EP cliticization exhibits an alternation between preverbal (proclitic), intraverbal (mesoclitic), and postverbal (enclitic) realization. Proclisis does not contribute to the vocabulary. Mesoclisis does, but it is restricted to future and conditional forms, it is only represented in 0.06% of the BN vocabulary, and is practically inexistent in the lecture domain. Concerning enclisis, in the past, when defining the vocabulary of the recognizer, our approach has been to consider such clitics together with the verbal form. Hence, for instance, *deu*, *deu-me*, and *deu-lhes* (*gave*, *gave me* and *gave them*, respectively) were 3 different entries. Altogether, each verbal form can have up to 30 different enclitics, although the vocabulary included only the most frequent enclitic verbal forms.

For the BN domain, the treatment of enclitic verbal forms as two separate entries yielded a 2.5% reduction in the original vocabulary size, and a very small OOV reduction (to 1.3%). The recognizer performance did not change, probably because the original percentage of errors in clitics was very small (0.95%). With the smaller vocabularies, the OOV values did not show any significant changes (2.3% for the 30k vocabulary, and 4.9% for the 15k one). This line of research was not pursued any further, and no tests were done in the lecture domain.

6. Filled Pauses

An analysis of the main types of recognition errors in the lecture domain revealed several sources which were common to the BN domain: disfluencies, severe vowel reduction, OOVs, large variability of inflected forms, and inconsistent spelling [10]. Other important sources of errors in the lecture domain are tag questions and negative, affirmative and interrogative grunts (e.g. *humhum*, *hum*, *hã*). Tag questions, in particular, are fairly frequent in both courses, given the need felt by the teachers to make sure that the class was following their presentations. Therefore, the teachers often invited the students to give feedback by using tag questions such as *não é?* (isn't it? - 36 instances in the ETI test set). The variety of possible phonetic realizations for such a word sequence in casual speech (e.g.[nˈɛ], [nɐˈɛ], [nõˈɛ]) coupled with the virtual non-representativeness of such examples in the written corpus makes them very difficult to be recognized. A similar problem was found for discourse markers, the most typical one being *portanto* ('so' - 104 instances in the ETI test set), pronounced mostly in very reduced forms.

Disfluencies, however, assume a particular importance. The disfluency rate (one in every 10 words, for PMC) is higher than the one cited in [1] and is a frequent cause for error bursts (only 19.2% of the errors in PMC occurred in isolation). Since filled pauses (FPs) are by far the most common disfluency type in our corpus, our analysis concentrated on them.

Given that FPs are not random events outside the control of the speaker, but appear to have different forms from language to language and a systematic contextual distribution ([1][12][13]), we aimed at contributing to the establishment of their inventory and determining the locations in which they were more likely to occur for EP. They were transliterated using "%" as the first character, and (i) a sequence of two graphemes representing vowels if they sound as a vowel-like segment only; (ii) 'mm' if they sounded as a nasal murmur only; (iii) a combination of those, if both types of vocalization were present. Only three different vocalizations were found: '%mm' (3.4%), '%aa' (90.4%) and '%aamm' (5.2%), where '%aa' represents a vowel-like segment close to the mid central vowel [ɐ]. This confirms previous observations for other EP corpora, namely those of [11] based on high school classroom presentations (4.4%, 78.5% and 17.1%, respectively). A further comparison with [11] also shows consistent trends in the distribution of these three types of vocalization: '%aamm' generally occurs at major intonational phrase boundaries, forms a prosodic constituent on its own, and is most often associated with large following silent pauses as well as with discourse topic changes. Instead, '%mm' has a clitic behavior, occurring only as a coda of disfluent prolon-

gations. '%aa' - which is by far the most frequent FP - also tends to form an intonational constituent on its own (57.8%) and may be used instead of '%aamm,' (17.4%). It may occur, however, in a much wider variety of contexts: it is the most likely FP form at minor intonational phrase boundaries (95.5% of %aa vs 4.5% of %aamm) and the most common inside complex disfluent sequencies (89.3% of %aa and 10.7% of %mm). Such results partially agree with the ones presented in [12] concerning the English *uh* and *um* respectively, supporting the hypothesis that FPs are conventional forms, under the speaker control, and thus consistently used to signal either minor or major delays in speaking.

This could explain, at least in part, the different frequency distribution of the 3 FP forms found in [11], which show that '%aamm' is more likely to occur in prepared non-scripted presentations (16.4%) than in really spontaneous ones (0.7%). This explanation may also account for the differences observed between the high school and the PMC teachers (13,2% and 6.2%, respectively), as the latter classes include longer stretches of non-prepared speech. Further evidence in favor of this hypothesis may eventually be given by the fact that in the parts of the teacher's prepared presentations which have slide visual support, instances of '%aamm' are not observed and '%aa' forms may be very short (around 40-50 ms). This hypothesis must be further tested by analyzing additional transcribed material.

Unlike the speakers in [11], the PMC teacher makes frequent short pauses at the beginning or middle of prosodic phrases, which do not result in an interruption of F0 general contour, nor trigger a complete repair of that phrase. As those instances of '%aa' are very similar in duration and spectral characteristics to the EP preposition, determiner and pronoun *a* ([ɐ]), serious mismatches may occur, in contexts where '%aa' is not recognized as FP and these homophonous counterparts are not predicted by the LM.

Given the current very high percentage of errors that can be attributed to misrecognized FPs, we hope that the results of this analysis, besides contributing to a better understanding of the role of FPs in European Portuguese may also be important for building more adequate LMs.

7. Conclusions and Future Work

This pilot study with lecture transcription allowed us to learn valuable lessons in terms of recording protocols, and validated the well known importance of large quantities of textual and manually transcribed material for training language and acoustic models. Despite the limited resources, our domain adaptation efforts yielded a significant (although not sufficient) WER reduction, which motivated the test of different strategies for vocabulary selection.

The fact that a significant percentage of the recognition errors occurs for function words (45.1%) lead to us believe that the current performance, although too bad in terms of transcription, may be good enough for indexation purposes. Hence, we plan to include indexation capabilities in our lecture browser application.

As this is our first project on spontaneous speech recognition for EP, the research challenges are enormous, in terms of strategies for dealing with disfluencies [14] [15], for introducing punctuation and capitalization, and for producing a surface rich transcription [16] that would be more intelligible for hearing impaired students. This is in fact our ultimate goal, given the demands that we currently have from students with progressive hearing disabilities.

8. Acknowledgments

The authors would like to thank João Neto, Hugo Meinedo, Ciro Martins, and Joaquim Jorge, for many helpful discussions. This work was partially funded by FCT project POSC/PLP/58697/2004. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

9. References

[1] Shriberg, E., "Spontaneous speech: How people really talk, and why engineers should care", Proc. Interspeech'2005, Lisbon, Portugal, 2005.

[2] Furui, S., Iwano, K., Hori, C., Shinozaki, T., Saito, Y., Tamura, S., "Ubiquitous speech processing", Proc. ICASSP'2001, Salt Lake City, USA, 2001.

[3] Lamel, L., Adda, G., Bilinski, E., Gauvain, J., "Transcribing lectures and seminars", Proc. Interspeech'2005, Lisbon, Portugal, 2005.

[4] Glass, J., Hazen, T., Hetherington, I., Wang, C., "Analysis and processing of lecture audio data: Preliminary investigations", Proc. Human Language Technology NAACL, Speech Indexing Workshop, Boston, 2004.

[5] LDC, "Simple metadata annotation specification version 6.2.", Technical report, Linguistic Data Consortium, 2004.

[6] Lindstrom, A., "English and Other Foreign Linguistic Elements in Spoken Swedish: Studies of Productive Processes and Their Modelling Using Finite-State Tools", PhD thesis, Linkoping University, 2004.

[7] Trancoso, I., Neto, J., Meinedo, H., Amaral, R., "Evaluation of an alert system for selective dissemination of broadcast news", Proc. Eurospeech'2003, Geneva, Switzerland, 2003.

[8] Stolcke, A., "SRLIM - an extensible language modeling toolkit", Proc. ICSLP'2002, Denver, USA, 2002.

[9] Martins, C., Neto, J., Almeida, L.: Using partial morphological analysis in language modeling estimation for large vocabulary portuguese speech recognition. In: Proc. Eurospeech '1999, Budapest, Hungary, 1999.

[10] Trancoso, I., Nunes, R., and Neves, L., "Classroom Lecture Recognition", Proc. 7th Int. Workshop on Computational Processing of the Portuguese Language", Brazil, 2006.

[11] Moniz, H., "Disfluencies in high school oral presentations", M.A. Thesis, Univ. of Lisbon, 2006.

[12] Clark, H., and Fox Tree, J., "Using uh and um in spontaneous speech", Cognition, 84:73-111, 2002.

[13] Swerts, M., "Filled pauses as markers of discourse structure", Journal of Pragmatics, 30(4):485-496, 1998.

[14] Johnson, M., Charniak, E., "A tag-based noisy channel model of speech repairs", Proc. ACL, Barcelona, Spain, 2004.

[15] Honal, M., Schultz, T., "Automatic disfluency removal on recognized spontaneous speech - rapid adaptation to speaker-dependent disfluencies", Proc. ICASSP'2005, Philadelphia, USA, 2005.

[16] Snover, M., Schwartz, R., Dorr, B., Makhoul, J., "RT-S: Surface rich transcription scoring, methodology, and initial results", Proc. of the Rich Transcription 2004 Workshop, Montreal, Canada, 2004.