Extension of the LECTRA corpus: classroom LECture TRAnscriptions in European Portuguese

Thomas Pellegrini[(1)], Helena Moniz[(1,2)], Fernando Batista[(1,3)], Isabel Trancoso[(1,4)], Ramon Astudillo[(1)]

[(1)] Spoken Language Systems Lab, INESC-ID, Lisbon, Portugal
[(2)] Faculdade de Letras, Universidade de Lisboa, Lisbon, Portugal
[(3)] ISCTE-IUL - Instituto Universitário de Lisboa
[(4)] Instituto Superior Técnico, Lisbon, Portugal
E-mail: thomas.pellegrini@inesc-id.pt, helena.moniz@inesc-id.pt

Abstract

This paper presents the recent extension of the LECTRA corpus, a speech corpus of university lectures in European Portuguese that will be partially available. Eleven additional hours of various lectures were transcribed, following the previous multilayer annotations, and now comprising about 32 hours. This material can be used not only for the production of multimedia lecture contents for e-learning applications, enabling hearing impaired students to have access to recorded lectures, but also for linguistic and speech processing studies. Lectures present challenges for automatic speech recognition (ASR) engines due to their idiosyncratic nature as spontaneous speech and their specific jargon. The paper presents recent ASR experiments that have clearly shown performance improvements on this domain. Together with the manual transcripts, a set of upgraded and enriched force-aligned transcripts was also produced. Such transcripts constitute an important advantage for corpora analysis, and for studying several speech tasks.

Keywords: Lecture domain speech corpus, ASR, speech transcripts, speech alignment, structural metadata, European Portuguese.

1. Introduction

This paper aims at a description of the corpus collected within the national project LECTRA and its recent extension. The LECTRA project aimed at transcribing lectures, which can be used not only for the production of multimedia lecture contents for e-learning applications, but also for enabling hearing-impaired students to have access to recorded lectures. The corpus has been already described in (Trancoso *et al*., 2008). We describe the recent extension of the manual annotations and the subsequent automatic speech recognition and alignment experiments to illustrate the performance improvements compared to the results reported in 2008. The extension was done in the framework of the METANET4U European project that aims at supporting language technology for European languages and multilingualism. One of the main goals of the project is that languages resources are made available online. Thus, the LECTRA corpus will be available through the central META-SHARE platform and through our local node: *http://metanet4u.l2f.inesc-id.pt/*.

Lecture transcription can be very challenging, mainly due to the fact that we are dealing with a very specific domain and with spontaneous speech. This topic has been the target of much bigger research projects such as the Japanese project described in Furui *et al.* (2001), the European project CHIL (Lamel *et al.*, 2005), and the American iCampus Spoken Lecture Processing project (Glass, 2007). It is also the goal of the Liberated Learning Consortium[1], which fosters the application of speech recognition technology for enhancing accessibility for students with disabilities in the university classroom. In some of these projects, the concept of lecture is different. Many of our classroom lectures are 60-minute long, and quite informal, contrasting with the 20-minute seminars used in (Lamel *et al.*, 2005), where a more formal speech can often be found.

After a short description of the corpus itself and the annotation schema in Sections 2 and 3 respectively, ASR experiments are reported in Section 4. Section 5 describes the creation of a dataset that merges manual and automatic annotations and that provides prosodic information. Section 6 presents the conclusions and the future work.

2. Corpus description

The corpus includes seven 1-semester courses: Production of Multimedia Contents (PMC), Economic Theory I (ETI), Linear Algebra (LA), Introduction to Informatics and Communication Techniques (IICT), Object Oriented Programming (OOP), Accounting (CONT), Graphical Interfaces (GI). All lectures were taught at Technical University of Lisbon (IST), recorded in the presence of students, except IICT, recorded in another university and in a quiet office environment, targeting an Internet audience. A lapel microphone was used almost everywhere, since it has obvious advantages in terms of non-intrusiveness, but the high frequency of head turning causes audible intensity fluctuations. The use of the head-mounted microphone in the last 11 PMC lectures clearly improved this problem. However, this microphone was used with an automatic gain control, causing saturation in some of the recordings, due to the increase of the recording sound level during the students' questions, in the segments after them. Most classes are 60-90 minutes long (with the exception of IICT courses which are given in 30 minutes). A total of 74h were recorded, of which 10h were multilayer annotated in 2008 (Trancoso *et al.*, 2008). Recently additional 11 hours were orthographically

[1] www.liberatedlearning.com

transcribed. Table 1 shows the number of lectures per course and the audio duration that was annotated, where V1 corresponds to the 2008 version of the corpus, *Added* is the quantity of added data, and V2 corresponds to the extended actual version.

	# Lectures			Duration		
	V1	Added	V2	V1	Added	V2
LA	5	+3	8	2h25	2h30	4h55
GI	3	+1	4	2h50	0h51	3h41
CONT	6	+1	7	4h40	1h02	5h42
ETI	3		3	3h11		3h11
IICT	4		4	1h37		1h37
OOP	5	+1	6	4h00	2h22	6h22
PMC	2	+5	7	2h00	4h09	6h09
Total	28	+11	39	20h43	+10h54	31h37

Table 1: Number of lectures and durations per course.

For future experiments, the corpus was divided into 3 different sets: Train (78%), Development (11%), and Test (11%). Each one of the sets includes a portion of each one of the courses. The corpus separation follows a temporal criterion, where the first classes of each course were included in the training data, and the final classes were included in the development and test sets. Figure 1 shows the portion of each course included in each one of the sets.

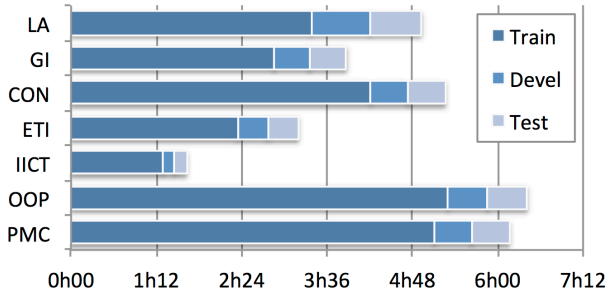

Figure 1 - Corpus distribution

3. Corpus annotation

The orthographic manual transcriptions were done using Transcriber[2] and Wavesurfer[3] tools. Automatic transcripts are used as a basis that the transcribers corrected. At this stage, speech is segmented into chunks delimited by silent pauses, already containing audio segmentation related to speaker and gender identification and background conditions. Previously, the annotation schema comprised multilayers of orthographic, morpho-syntactic, structural metadata (Liu *et al.*, 2006; Ostendorf *et al.*, 2008), *i.e.*, disfluencies and punctuation marks, and paralinguistic information as well (laughs, coughs, etc.). The multilayer annotation aimed at providing a suitable sample for further linguistic and speech processing analysis in the lectures domain. The extension reported in this work respects the previous schema, however does not comprise the morpho-syntactic information tier, since automatic classifications of part-of-speech (POS) tags and of syntactic parsing is automatically performed, initially by *Marv* (Ribeiro *et al.*, 2003) and more recently by *Falaposta* (Batista *et al.*, 2012). Thus, the extension of the annotation comprises the full orthographic transcription, enriched with punctuation and disfluency marks and a set of diacritics fully reported in Trancoso *et al.* (2008). Segmentation marks were also inserted for regions in the audio file that were not further analyzed (background noise, signal saturation).

Three annotators (with the same linguistics background) transcribed the extended data. However, two courses could not benefit from the extension for different reasons: the IICT, since no more lectures were recorded, and the ETI due to the fact that the teacher did not accept to make his recordings publicly available.

Due to the idiosyncratic nature of lectures as spontaneous and prepared non-scripted speech, the annotators reported in the five sessions of the guidelines instructions two main difficulties: in punctuating the speech and in classifying the disfluencies. The punctuation complexities are mainly associated with the fact that speech units do not always correspond to sentences, as established in the written sense. They may be quite flexible, elliptic, restructured, and even incomplete (Blaauw, 1995). Therefore, to punctuate speech units is not always an easy task. For a more complete view on this, we used the summary of grammatical and ungrammatical locations of punctuation marks for European Portuguese described in Duarte (2000). The latter is related to the different courses and the difficulty in discriminating the specific types of disfluencies (if it is a substitution, for instance), since the background of the annotators is on linguistics. To sum up, the guidelines given to our annotators were: the schema described in Trancoso *et al.* (2008) and the punctuation summary described in Duarte (2000).

The general difficulty of measuring the inter-transcriber agreement is due to the fact that two annotators can produce token sequences of different lengths. This is equivalent to measuring the speech recognition performance, where the length of the recognized word sequence is usually different from the reference. For that reason, the inter-transcriber agreement was calculated for pairs of annotators, considering the most experienced[4] as reference. The standard F1-measure and Slot Error Rate (SER) (Makhoul *et al.*, 1999) metrics were used, where each slot corresponds to a word, a punctuation mark or a diacritic:

$$F1 - measure = \frac{2 \text{ x Precision x Recall}}{Precision + Recall} \quad , \quad SER = \frac{\text{errors}}{\text{ref\_tokens}}$$

where *ref_tokens* is the number of words, punctuation marks and diacritics used in the reference orthographic

[2] http://trans.sourceforge.net/
[3] http://www.speech.kth.se/wavesurfer/

[4] The annotator in question had already transcribed other corpora with the same guidelines.

tier, and *errors* comprise the number of inserted, deleted or substituted tokens.

The inter-transcriber agreement of the three annotators is based on a selected sample of 10 minutes of speech from one speaker involving more than 2000 tokens. The selection of the sample has to do with the reported difficulties of the annotators, in annotating disfluencies (*e.g.*, complex sequences of disfluencies) and also punctuation marks. Table 2 reports the inter-transcriber agreement results for each pair of annotators. The table shows the number of (*Cor)rect slots*, (*Ins)ertions*, *(Del)etions*, (*Sub)stitutions*, (*F1)-measure*, and *slot accuracy* (SAcc), which corresponds to *1-SER*. There is an *almost perfect agreement* between A1 and the remaining annotators, and a *substantial agreement* between the pair A2-A3. These results may well be the outcome of a thorough process of annotation in several different steps and with intermediate evaluations during the 5 guidelines instruction sections. Moreover, several other annotators for other corpora already tested the guidelines here in use.

Annotator	Cor	Ins	Del	Sub	F1	SER	SAcc
A1-A2	1714	67	79	224	0.852	0.184	**0.816**
A1-A3	1632	38	34	351	0.808	0.210	**0.790**
A2-A3	1480	81	97	444	0.735	0.308	**0.692**

Table 2: Evaluation of the inter-transcriber agreement.

4. ASR experiments

Transcribing lectures is particularly difficult since lectures are very domain-specific and speech is spontaneous. Except the IICT lectures where no students were present, students demonstrate a relatively high interactivity in the other lectures. Nevertheless, since only a lapel microphone was used to record the close-talk speech of the lecturers, the audio gain of the student interventions is very low. The presence of background noise, such as babble noise, footsteps, blackboard writing noise, etc. may difficult the speech processing, in particular the Speech / Non-speech detection that feeds the recognizer with audio segments labelled as speech. Typical WER reported in the recent literature is between 40-45% (Glass *et al.*, 2007).

4.1 Overview of our ASR system

Our automatic speech recognition engine named Audimus (Neto *et al.*, 2008; Meinedo *et al.*, 2008) is a hybrid automatic speech recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of Multi-Layer Perceptrons (MLPs). The MLPs perform a phoneme classification by estimating the posterior probabilities of the different phonemes for a given input speech frame (and its context). These posterior probabilities are associated to the single state of context independent phoneme HMMs.

The most recent ASR system used in this work is exactly the ASR system for EP described in (Meinedo *et al.*, 2010). The acoustic models were initially trained with 46 hours of manually annotated broadcast news (BN) data collected from the public Portuguese TV, and in a second time with 1000 hours of data from news shows of several EP TV channels automatically transcribed and selected according to a confidence measure threshold (non-supervised training). The EP MLPs are formed by 2 hidden layers with 2000 units each and have 500 softmax output units that correspond to 38 three state monophones of the EP language plus a single-state non-speech model (silence) and 385 phone transition units which were chosen to cover a very significant part of all the transition units present in the training data. Details on phone transition modeling with hybrid ANN/HMM can be found in (Abad & Neto, 2008).

The Language Model (LM) is a statistical 4-gram model that was estimated from the interpolation of several specific LMs: in particular a backoff 4-gram LM, trained on a 700M word corpus of newspaper texts, collected from the Web from 1991 to 2005, and a backoff 3-gram LM estimated on a 531k word corpus of broadcast news transcripts. The final language model is a 4-gram LM, with Kneser-Ney modified smoothing, 100k words (or 1-gram), 7.5M 2-gram, 14M 3-gram and 7.9M 4-gram. The multiple-pronunciation EP lexicon includes about 114k entries.

These models, both AMs and the LM, were specifically trained to transcribe BN data. The Word Error Rate (WER) of our current ASR system is under 20% for BN speech in average: 18.4% for instance, obtained in one of our BN evaluation test sets (RTP07), composed by six one hour long news shows from 2007 (Meinedo *et al.*, 2010).

4.2 ASR results

A test subset was selected from the corpus in 2008, by choosing one single lecture per course. In (Trancoso *et al.*, 2008), preliminary ASR results were reported on this test set, showing the difficulty to transcribe lectures. Very high word error rates (WER), 61.0% in mean, were achieved for a subset of various lectures chosen as a test set. It has a vocabulary of around 57k words. Applying this recognize without any type of domain adaptation, obviously yielded very bad results

Table 3 illustrates the performance of the old and the recent systems without and with adaptation of the LM for the recent system. Our recent system, which was described in the previous section, achieved a WER of 45.7% on the same test subset, hence showed a 25.0% relative reduction. The lexicon was almost twice the size of the one of the previous system. Further improvements were achieved with a 44.0% WER. This performance was obtained by interpolating our generic broadcast news 4-gram LM with a 3-gram LM trained on the training lecture subset. 100-best hypotheses were generated per each sentence and rescored with this LM and a RNN (implementation of the Brno University (Mikolov *et al.*, 2011)). This RNN was trained only on the lecture train subset.

An analysis of the ASR errors showed that most of the

misrecognitions concerned small function words, such as definite articles and prepositions, the backchannel word "OK" also appeared to be very often misrecognized. Then, words specific to each jargon of the courses also were error-prone. For instance variable names in the Linear Algebra lecture, such as "alfa", "beta", "vector" were often substituted. In the PMC lecture, words such as "MPEG", "codecs", "metadados" (metadata), "URL" were subject to frequent errors.

ASR system	LM adapt?	OOV (%)	WER (%)
2008	no	_	61.0
2011	no	2.8	45.7
2011	yes	1.7	44.0

Table 3: Comparison of the ASR results reported in 2008 and obtained with our most recent system. OOV stands for out-of-vocabulary words.

5. Enriched annotations

The ASR system is able not only to produce automatic transcripts from the speech signal, but also to produce automatic force-aligned transcripts, adjusting the manual transcripts to the speech signal. Apart from the existing manual annotations of the corpus, automatic force-aligned transcripts have been produced for the extended version of the corpus, and will be available in our META-SHARE node. These force-aligned transcripts were updated with relevant information coming from the manual annotations, and finally enriched with additional prosodic information (Batista *et al.*, 2012). The remainder of this Section provides more details about this process.

5.1 Automatic alignment

Force-aligned transcripts depend on a manual annotation and therefore do not contain recognition errors. A number of speech tasks, such as the punctuation recovery, may use information, such as pause durations, which most of the times is not available in the manual transcripts. On the other hand, manual transcripts provide reduced or error-free transcripts of the signal. For that reason, force-aligned transcripts, which combine the ASR information with manual transcripts, provide unique information, suitable for a vast number of tasks.

An important advantage of using force-aligned transcripts is that they can be treated in the exact same way as the automatic transcripts, but without recognition errors, requiring the same exact procedures and tools. However, the alignment process is not always performed correctly due to a number of reasons, in particular when the signal contains low energy levels. For that reason, the ASR parameters can be adjusted to accommodate the manual transcript into the signal. Our current force-alignment achieves 3.8% alignment word errors in the training, 3.1% in the development, and 4.5% in the evaluation sets.

5.2 Merging manual and automatic annotations

Starting with the previously described force-aligned transcripts, we have produced a self-contained dataset that provides not only the information given by the ASR system, but also important parts of the manual transcripts. For example, the manual orthographic transcripts include punctuation marks and capitalization information, but that is not the case of force-aligned transcripts, which only includes information, such as: word time intervals, and confidence scores. The required manual annotations are transferred by means of alignments between the manual and automatic transcripts.

Apart from transferring information from the manual transcripts, the data was also automatically annotated with part-of-speech information. The part-of-speech tagger input corresponds to the text extracted from the ASR transcript, after being improved with the reference capitalization. Currently, the Portuguese data is being annotated using *Falaposta*, a CRF-based tagger robust to certain recognition errors, given that a recognition error may not affect all its input features. It accounts for 29 part-of-speech (POS) tags and achieves 95.6% accuracy.

The resulting file, structured using the XML format, corresponds to the ASR output, extended with: time intervals to be ignored in scoring, focus conditions, speaker information for each region, punctuation marks, capitalisation, disfluency marks, and POS information.

5.3 Adding prosodic data

The previously described extended XML file is further improved with phone and syllable information, and other relevant information that can be computed from the speech signal *(e.g.*, pitch and energy). The data provided by the ASR system allows us to calculate the phone information. Marking the syllable boundaries as well as the syllable stress are achieved by means of a lexicon containing all the pronunciations of each word together with syllable information, since these tasks are currently absent in the recognizer. A set of syllabification rules was designed and applied to the lexicon, which account fairly well for the canonical pronunciation of native words, but they still need improvement for words of foreign origin.

Pitch (f_0) and energy (E) are two important sources of prosodic information, currently not available in the ASR output, and directly extracted from the speech signal. Algorithms for automatic extraction of the pitch track have, however, some problems, *e.g.*, octave jumps; irregular values for regions with low pitch values; disturbances in areas with micro-prosodic effects; influences from background noisy conditions; *inter alia*. We have removed all the pitch values calculated for unvoiced regions in order to avoid constant micro-prosodic effects. This is performed in a phone-based analysis by detecting all the unvoiced phones. We also had a calculation cost to eliminate octave-jumps. As to the influences from noisy conditions, the recognizer has an Audio Pre-processing or Audio Segmentation module, which classifies the input speech according to different focus conditions (*e.g.*, noisy, clean), making it possible to isolate those speech segments with unreliable pitch values.

After extracting and calculating the above information, all data was merged into a single data source. The existing XML data has been upgraded in order to accommodate the additional prosodic information.

6. Conclusions

This paper described our lecture corpus in European Portuguese, and its recent extension. The problems it raises for automatic speech recognition systems were illustrated. The fact that a significant percentage of the recognition errors occurs for function words led to us believe that the current performance, although far from ideal, may be good enough for information retrieval purposes, enabling keyword search and question answering in the lecture browser application. ASR performance is still poor but as stated in Glass *et al.* (2007), "accurate precision and recall of audio segments containing important keywords or phrases can be achieved even for highly-errorful audio transcriptions (*i.e.*, word error rates of 30% to 50%)''. Together with the manual transcripts, a set of upgraded and enriched force-aligned transcripts were produced and made available. Such transcripts constitute an important advantage for corpora analysis, and for studying a number of speech tasks. Currently, the LECTRA corpus is being used to study and perform punctuation and capitalization tasks, and spontaneous speech phenomena. We believe that producing a surface rich transcription is essential to make the recognition output intelligible for hearing impaired students. Six courses of the corpus will be soon available to the research community via the META-SHARE platform.

7. Acknowledgements

This work was partially funded by the European project METANET4U number 270893, by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2011, and by DCTI - ISCTE-IUL.

8. References

Abad, A. and Neto, J. (2008). Incorporating Acoustical Modelling of Phone Transitions in a Hybrid ANN/HMM Speech Recognizer. In *Proc. Interspeech*, pp. 2394-2397, Brisbane.

Batista, F., Moniz, H., Trancoso, I., Mamede, N. and Mata, A. I. (2012), Unified Data Representation for Prosodically-based Speech Processing. *JOSS – Journal of Speech Sciences* (submitted).

Batista, F., Moniz, H., Trancoso, I., and Mamede, N. J. (2012). Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. *IEEE Transactions on Audio, Speech and Language Processing, Special Issue on New Frontiers in Rich Transcription*, 20(2):474 – 485.

Blaauw, E. (1995). On the Perceptual Classification of Spontaneous and Read Speech. PhD Diss, Research Institute for Language and Speech, Utrecht.

Duarte, I. (1995). *Língua Portuguesa, Instrumentos de Análise*. Universidade Aberta.

Furui, S., Iwano, K., Hori, C., Shinozaki, T., Saito, Y. and Tamura, S. (2001). Ubiquitous speech processing. In *Proc. ICASSP*, Salt Lake City.

Glass, J., Hazen, T., Cyphers, S., Malioutov, I., Huynh, D. and Barzilay, R. (2007). Recent Progress in the MIT Spoken Lecture Processing Project. In *Proc. Interspeech*, Antwerp.

Lamel, L., Adda, G., Bilinski, E. and Gauvain, J. (2005). Transcribing lectures and seminars. In *Proc. Interspeech*, Lisbon.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., and Harper, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. In *Transactions on Audio, Speech and Language Processing*, 14(5):1526–1540.

Makhoul, J., Kubala, F., Schwartz, R. and Weischedel, R. (1999). Performance measures for information extraction. In *Proc. of the DARPA BN Workshop*.

Meinedo, H., Viveiros, M. and Neto, J. (2008). Evaluation of a live broadcast news subtitling system for Portuguese. In *Proc. of Interspeech*, Brisbane, Australia.

Meinedo, H., Abad, A., Pellegrini, T., Neto, J. and Trancoso, I. (2010). The L2F Broadcast News Speech Recognition System. In *Proc. Fala*, pp. 93—96, Vigo.

Mikolov, T., Deoras, A., Kombrink, S., Burget, L. and Cernocky, J.H. (2011). Empirical evaluation and combination of advanced language modeling techniques. In *Proc. Interspeech*, pp. 605-608, Florence.

Moniz, H., Batista, F., Trancoso, I. and Mata, A. I. (2012), Prosodic context-based analysis of disfluencies. In *Proc. Interspeech*, Portland, Oregon.

Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C. and Caseiro, D. (2008). Broadcast News Subtitling System in Portuguese. In *Proc. ICASSP*, Las Vegas.

Ostendorf, M., *et al*. (2008). Speech segmentation and spoken document processing. *IEEE Signal Processing Magazine*, 25(3):59–69.

Ribeiro, R., Oliveira, L. C. and Trancoso, I. (2003). Morphossyntactic information for TTS systems: comparing strategies for European Portuguese. In *Proc. Propor*, Springer-Verlag, LNAI, pp. 143-150, Faro.

Trancoso, I., Neto, J., Meinedo, H. and Amaral, R. (2003). Evaluation of an alert system for selective dissemination of broadcast news. In *Proc. Eurospeech*, Geneva.

Trancoso, I., Nunes, R., Neves, L., Viana, M. C., Moniz, H., Caseiro, D. and Mata, A. I. (2006). Recognition of Classroom Lectures in European Portuguese. In *Proc. Interspeech*, Pittsburgh.

Trancoso, I., Martins, R., Moniz, H., Mata, A. I. and Viana, M. C. (2008). The LECTRA Corpus - Classroom Lecture Transcriptions in European Portuguese. In *Proc. LREC*, Marrakech.