Extension of the LECTRA corpus
Classroom Lecture Transcriptions in European Portuguese

Thomas Pellegrini[1], *Helena Moniz*[1,2], *Ramon Astudillo*[1,3], *Isabel Trancoso*[1,3]

[1]L2F - Spoken Language Lab, INESC-ID Lisboa, Portugal
[2]Faculdade de Letras, Universidade de Lisboa, Portugal
[3]Instituto Superior Técnico, Lisboa, Portugal
{Thomas.Pellegrini}@inesc-id.pt

Index Terms: Speech corpus, European Portuguese, ASR, spoken lectures transcription

1. Summary

The LECTRA corpus was collected in the framework of a national Portuguese project that ran from 2005 to 2007, aiming at doing automatic transcription of university lectures in European Portuguese. This material 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 includes seven 1-semester courses: Linear Algebra (LA), Economic Theory I (ETI), Object Oriented Programming (OOP), Production of Multimedia Contents (PMC), Accounting (A), Graphical Interfaces (GI), Introduction to Informatics and Communication Techniques (IICT). All lectures were taught at IST (Instituto Superior Técnico, Lisbon), recorded in the presence of students, except IICT, recorded in a quiet office environment, targetting an Internet audience. Most classes are 60-90 minutes long. A total of 74h were recorded, of which 10h were multilayer annotated in 2008, including (besides other information) an orthographic tier, a morpho-syntactic tier, and a disfluency tier. For more details, the reader may refer to [1].

In the context of the European project METANET4U that aims at supporting language technology for European languages and multilingualism, six of the courses in the LECTRA corpus will be distributed, under a license that remains to be chosen. Resources to be distributed shall be upgraded and extended. In this paper, a recent extension of the manual transcriptions is described, and automatic speech recognition experiments are reported. Three annotators transcribed 11 additional hours of various lectures.

Due to the idiosyncratic nature of lectures as spontaneous and prepared non-scripted speech, the annotators faced two main difficulties, in punctuating the speech and in classifying the disfluencies. The former is 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 [2]. 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 described in [3]. The latter is related to the different courses and the diffilculty in discriminating the specific types of disfluencies (if it is a substitution, for instance), since the background of the annotators is on linguistics.

In [1], preliminary ASR results were reported, 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. 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, foot steps, blackboard writing noise, etc. may difficults the speech processing, in particular the Speech / Non-speech detection that feeds the recognizer with audio segments labeled as speech. Typical WER reported in the recent literature are between 40-45% [4]. In this study, we report new ASR experiments to be compared to 2008's results, by using more recent acoustic models, new language models, and a larger vocabulary. In mean, our new baseline system achieved a WER of 45.7% on the same test subset as in [1], hence a 25.0% relative reduction. Further improvements were achieved with our best result: 44.0% WER. This performance was obtained by interpolating our generic broadcast news 4-gram LM with an 3-gram LM trained on the training lecture subset. The 100-best hypothesis were rescored with this LM and a recurrent neural network (implementation of the Brno university [5]) trained only on the lecture train subset. An analysis of the ASR errors will be done to determine the main difficulties: in particular Out Of Vocabulary words (OOV) due to jargon use, and frequent disfluencies.

2. References

[1] I. Trancoso, R. Martins, H. Moniz, A. Silva, and M. Ribeiro, "The LECTRA Corpus - Classroom Lecture Transcriptions in European Portuguese," in *Proc. LREC*, Marrakech, 2008.

[2] E. Blaauw, "On the Perceptual Classification of Spontaneous and Read Speech," Ph.D. dissertation, Utrecht, Research Institute for Language and Speech, 1995.

[3] I. Duarte, *Lngua Portuguesa, Instrumentos de Anlise*. Universidade Aberta, 1995.

[4] J. Glass, T. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the MIT Spoken Lecture Processing Project," in *Proc. Interspeech*, Antwerp, 2007, pp. 2553–2556.

[5] T. Mikolov, A. Deoras, S. Kombrink, L. Burget, and J. Cernocky, "Empirical evaluation and combination of advanced language modeling techniques," in *Proc. Interspeech*, Florence, 2011, pp. 605–608.