LMELECTURES: A MULTIMEDIA CORPUS OF ACADEMIC SPOKEN ENGLISH

K. Riedhammer, M. Gropp, T. Bocklet, F. Hönig, E. Nöth, S. Steidl

Pattern Recognition Lab, University of Erlangen-Nuremberg, GERMANY

noeth@cs.fau.de

Abstract

This paper describes the acquisition, transcription and annotation of a multi-media corpus of academic spoken English, the *LMELectures*. It consists of two lecture series that were read in the summer term 2009 at the computer science department of the University of Erlangen-Nuremberg, covering topics in pattern analysis, machine learning and interventional medical image processing. In total, about 40 hours of high-definition audio and video of a single speaker was acquired in a constant recording environment. In addition to the recordings, the presentation slides are available in machine readable (PDF) format. The manual annotations include a suggested segmentation into speech turns and a complete manual transcription that was done using BLITZSCRIBE2, a new tool for the rapid transcription. For one lecture series, the lecturer assigned key words to each recordings; one recording of that series was further annotated with a list of ranked key phrases by five human annotators each. The corpus is available for non-commercial purpose upon request.

Index Terms: corpus description, academic spoken English, e-learning

1. Introduction

The *LMELectures* corpus of academic spoken English consists of high-definition audio and video recordings of two graduate level lecture series read in the summer term 2009 at the computer science department of the University of Erlangen-Nuremberg. The *pattern analysis (PA)* series consists of 18 recordings covering topics in pattern analysis, pattern recognition and machine learning. The *interventional medical image processing (IMIP)* series consists of 18 recordings covering topics in medical image reconstruction, registration and analysis. The lectures are read by a single, non-native but proficient speaker, and acquired in the *E-Studio*[1] which ensures a constant recording environment in the same room using a clip-on cordless close-talking microphone. The recordings were professionally edited to achieve a constant high audio and video quality. Note that not all lectures are consecutive; some recordings had to be dropped from the corpus because of a different speaker, sole use of German language, or technical issues such as a misplaced or defect close-talking microphone.

This paper documents the acquisition of the audio and video data (Sec. 2), the semi-automatic segmentation (Sec. 3), the subsequent manual transcription (Sec. 4), and the additional annotations (Sec. 5). Sec. 6 lists possible uses of the *LMELectures* and places the corpus in context with other corpora of academic spoken English. Sec. 7 suggests a partitioning of the data that is recommended for research on automatic speech recognition and key phrase extraction.

[1] RRZE MultiMediaZentrum, http://www.rrze.uni-erlangen.de/dienste/arbeiten-rechnen/multimedia/

2. Audio and Video Data

The audio data was acquired at a sampling rate of 48 kHz and 16 bit quantization, and stored in the *Audio Interchange File Format (AIFF)*. A 16 kHz version for the use with speech recognition systems was produced using down-sampling. The cordless close-talking microphone was able to reduce most of the room acoustics and background noises.

The video was acquired using an HD camera with manually controlled viewpoint and zoom setting to track the lecturer. Furthermore, the currently displayed presentation slide and, if applicable, on-screen writings is captured seperately. The video data is available in two formats:

- Presenter only, 640 x 360 pixel resolution, H.264 encoded (see Fig. 1, inset on the top left).
- Presenter, currently displayed slide and on-screen writings and lecture title, 1280 x 600 pixel resolution, H.264 encoded (see Fig. 1).

In total, 39.5 hours of audio and video data was acquired from 36 lecture recordings. The video recordings feature an AAC encoded audio stream based on the original 48 kHz data.

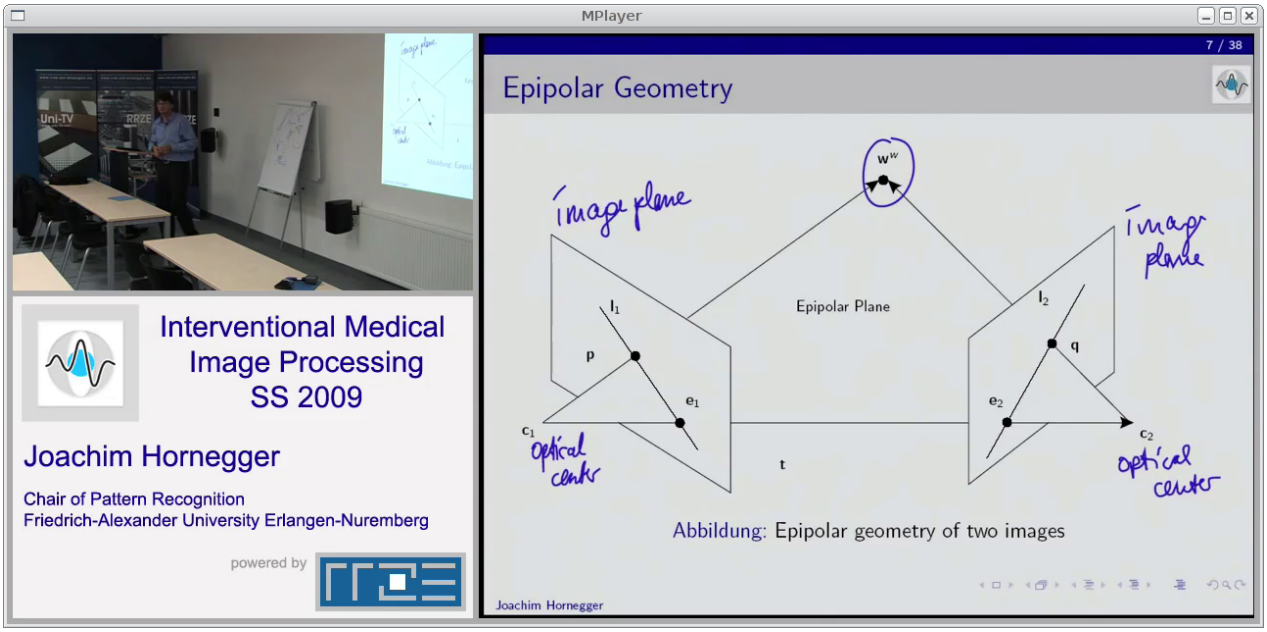

Figure 1: Example image from the video of lecture *IMIP01*. The left side shows the lecturer (top) and the lecture title (bottom), the right side shows the current slide and on-screen writings.

3. Semi-Automatic Segmentation

For the manual transcription, as well as for most speech recognition and understanding tasks, long recordings are typically split into short segments of speech. Another benefit is that longer periods of silence are removed from the data. The segmentation of the *LMELectures* is based on the time alignments of a Hungarian phoneme recognizer [1] that has been successfully used for speech/non-speech detection in various speaker and language identification tasks. The rich phonetic alphabet of the Hungarian language was found to be advantageous in the presence of various languages (here German and English) or wrong pronunciations. The set of phoneme strings was reduced by mapping the 61 original symbols to two groups: the pause (*pau*), noise (*int*, *e.g.*, a door slam) and speaker noise (*spk*, only if following *pau*, *e.g.*, cough) symbols were mapped to *silence* and the remaining symbols to *speech*. Merging adjacent segments of *silence* and *speech* results in an initial speech/non-speech segmentation (*cf.* Fig. 2).

Due to the design of the phoneme recognizer, the resulting segmentation has very sharp cut-offs and does not necessarily reflect the actual utterance or sentence structure, as even a very short pause may terminate a speech segment. With the aim of producing speech segments of an average length of four to five seconds[2], consecutive speech segments are merged based on certain criteria regarding segment lengths and intermediate silence (*cf.* Tab. 1).

Algorithm 1: Merge of consecutive segments based on their duration and interleaving silence.

```
for all segments i do
    if Pau(i, i+1) < min. pau or Dur(i) <
    min. dur then
        required ← true
        while required or Dur(i) < max. dur do
            if ! required then
                if Dur(i) > med. dur or
                Dur(Merge(i, i+1)) >
                max. dur or Pau(i, i+1) >
                max. pau then break
            end
            i ← Merge(i, i+1)
            required ← (Pau(i, i+1) <
            min. pau)
        end
    end
end
```

Algorithm 1 outlines the greedy merging procedure. 150 ms were added to the end of each segment to ease the sharp cut-offs. Given the desired target length, the major control variables are the pauses. Allowing too long pauses within a segment (`max.pau`) may lead to segments that contain the end and beginning of two separate

[2] as suggested by previous experiences of the group with manual transcription and speech recognition system training and evaluation

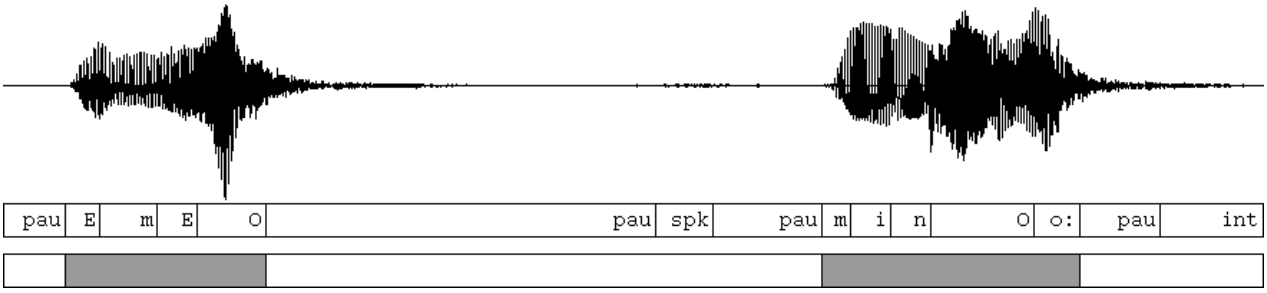

Figure 2: ≫*And then (breath) we know*≪. Adjacent segments of *silence* or *speech* phonemes are merged to an initial speech (gray) and non-speech (white) segmentation.

quantity	*description*	*value*
`min.dur`	if segment is shorter than *min. dur*, merge with following	2 s
`med.dur`	stop if merged segment is longer than *med. dur*	4 s
`max.dur`	only merge if resulting segment is shorter than *max. dur*	6 s
`max.pau`	maximum duration of pause within a segment	1 s
`min.pau`	minimum duration of pause between two segments	0.5 s

Table 1: Final merging criteria for consecutive speech segments.

utterances. Requiring long silences between segments (`min.pau`) leads to unnaturally long segments.

The segmentation closest to the desired characteristics comprises 23 857 speech turns with an average duration of 4.4 seconds, and a total of about 29 hours of speech. Note that these segments are for the purpose of recognition, and do not necessarily resemble dialog acts or "actual" speech turns. The right column of Tab. 1 shows the respective merging criteria. The typically 0.5 s to 3 s of silence between speech segments accumulate to about 10 hours.

4. Manual Transcription

The manual transcription of speech typically requires about ten to 50 times the duration of speech using professional tools like TRANSCRIBER [2, 3]. TRANSCRIBER, similar to other tools, allows to work on long recordings by identifying segments of speech, noise and other acoustic events. Furthermore, higher level information like speaker, speech or language attributes can be annotated. However, this higher level information regarding the data at hand is usually known in advance, and lectures are typically very dense in terms of speech, thus reducing the main task to the (desirably) fast transcription of the speech segments.

The segments were manually transcribed using BLITZSCRIBE2,[3] a platform independent graphical user interface specifically designed for the rapid transcription of large amounts of speech data. It is inspired by research of Roy *et al.* [3] and is publicly available as part of the Java Speech Toolkit (JSTK) [4].[4] Fig. 3 shows the interface that displays the waveform of the currently selected speech segment, a progress bar indicating the current playback position, an input text field to type the transcription, and a list of turns, optionally with prior transcription.

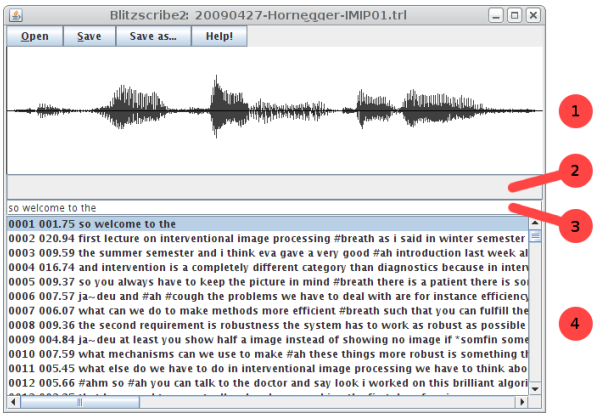

Figure 3: Screenshot of the BLITZSCRIBE2 transcription tool; (1) waveform of the currently selected speech segment, (2) progress bar indicating the current playback position, (3) text field for the transcription, (4) list of segments with transcription (if available).

The key idea to speed up the transcription is to simplify the way the user interacts with the program: although the mouse may be used to select certain turns for transcription or replay the audio at a desired time, the most frequent commands are accessed via keyboard shortcuts listed in Tab. 2.

For a typical segment, the transcriber types the transcription as he listens to the audio, pauses the playback if necessary (CTRL+SPACE), and hits ENTER to save the transcription, which loads the next segment and starts the playback. This process is very ergonomic as the hands

[3] `http://www5.informatik.uni-erlangen.de/en/research/software/blitzscribe2/`

[4] `http://code.google.com/p/jstk`

key combination		*command*
ENTER		save transcript, load **and play** next segment
SHIFT	+BACKSPACE	save transcript, load previous segment
SHIFT	+ENTER	save transcript, load next segment
CTRL	+SPACE	start/pause/resume/restart playback
CTRL	+BACKSPACE	rewind audio and restart playback
ALT	+S	save transcription file

Table 2: Keyboard shortcuts for fast user interactions in BLITZSCRIBE2.

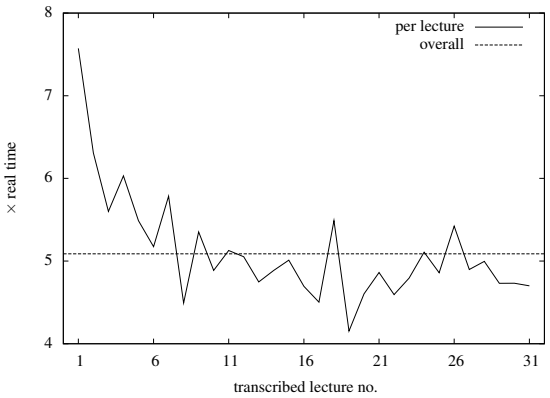

Figure 4: Change of the median transcription real time factor required by transcriber 1 throughout the transcription process.

remain on the keyboard during all times.

The lectures were transcribed by two transcribers. The work was shared among the transcribers and no lecture was transcribed twice. As the language is very technical, a list of common abbreviations and technical terms was provided along with the annotation guidelines. The overall median time required to transcribe a segment was about five times real time, which is a significant improvement over traditional transcription tools. Fig. 4 shows the decreasing transcription real time factor of one transcriber while adapting to the BLITZSCRIBE2 tool.

In total, about 300 500 words were transcribed with an average of 14 words per speech segment. Intermittent German words were transcribed and marked; those typically include greetings or short back-channel. Other foreign, mispronounced or fragmented words were transcribed as closely as possible, and marked for later special treatment. The resulting vocabulary size is 5 383 including multiple forms of words (*e.g.*, plural, composita), but excluding words in foreign languages and mispronounced or word fragments.

5. Further Manual Annotations

The presentation slides are available in machine readable (PDF) format, however, only the video provides accurate information about the display times. The lecturer added key words to each of the lecture recordings in series *PA*.

Lecturer's Phrases	*Annotator 1*	*Annotator 2*	*Annotator 3*	*Annotator 4*	*Annotator 5*
linear regression	●	●	●	●	●
norms	●		●	●	○
dep. linear regression			○		○
ridge regression	●		●		●
discriminant analysis	○		○	○	
motivation					
AP(5)	0.90				
NDCG(5)	0.73				

Table 3: Master key phrases of lecture *PA06* assigned by the lecturer, coverage indicators (●) for the human annotators, and phrase rank of the automatic rankings, if applicable. The empty bullets (○) indicate a partial match, *e.g.*, "linear discriminant analysis" satisfies "discriminant analysis."

The individual lecture *PA06* was further annotated with a ranked list of key phrases by five human subjects that have either attended the lecture or a similar lecture in a different term. The annotators furthermore graded the phrases present in their ranking in terms of quality from 1 – "sehr relevant" *(very relevant)* to 6 – "nutzlos" *(useless)*. This additional annotation can be used to assess the quality of automatic rankings using measures such as average precision (AP) [5] or normalized distributed cumulative gain (NDCG) [6, 7], two measures popular in the search engine and information retrieval community.

Tab. 3 shows, for *PA06*, the lecturer's phrases, whether the raters also extracted them, and the average AP and NDCG when comparing each rater to the remaining ones when considering the top five ranked terms.

6. Intended Use and Distinction from Other Corpora of Academic Spoken English

The corpus, with its annotations, is an excellent resource for various mono- and multi-modal research. The roughly 30 hours of speech of a single speaker provide a great base to work on acoustic and language modeling, speaker adaptation, prosodic analysis and key phrase extraction. The spoken language is somewhere in between read text and spontaneous speech, with passages of well-structured and articulated speech followed by a mumbled utterance with disfluencies and hesitations. At a higher level, the video can be used to determine slide timings, on-screen writing and other interactions of the lecturer. The two series of consecutive lectures provide a good scenario to work on automatic vocabulary extension and language model adaptation as required for a production system.

name	*duration*	*# turns*	*# words*	*% OOV*
train	24h 31m 55s	20 214	250 536	—
dev	2h 07m 28s	1 802	21 909	0.87 %
test	2h 12m 30s	1 750	23 497	0.99 %

Table 4: Data partitioning for the *LMELectures* corpus; the number of words excludes word fragments and foreign words. The percentage of OOV words is given with respect to the words present in the *train* partition.

The two main corpora of academic spoken English are the BASE corpus,[5] and the Michigan Corpus of Academic Spoken English (MICASE) [8]. Although both corpora cover more than 150 hours of speech, their setting is different from the *LMELectures*. The BASE corpus covers 160 lectures and 40 seminars from four broad disciplinary groups (Arts and Humanities, Life and Medical Sciences, Physical Sciences, Social Sciences). Audio, video and transcription material are available for licensing. The MICASE corpus features a wide variety of recordings of academic events including lectures, colloquia, meetings, dissertation defenses, *etc.*. Again, audio and transcripts are subject to licensing, but video data is unavailable.

The main distinction of the *LMELectures* is however the technical homogeneity in terms of recording environment, speaker, and topic of the two lecture series.

7. Suggested Data Partitioning

For experiments on speech recognition and key phrase extraction, the authors suggest to partition the data in three parts. The development set, *devel*, consists of the four lecture sessions IMIP13, IMIP17, PA15 and PA17, and has a total duration of about two hours. The test set, *test*, consists of the four lecture sessions IMIP05, IMIP09, PA06 and PA08, and has also a total duration of about two hours. The remaining 28 lecture sessions form the training set, *train*, with a total of about 24 hours. Tab. 4 summarizes the partitioning and lists details on the duration, number of segments and words, and out-of-vocabulary (OOV) rate with respect to a lexicon based on the training set. A baseline speech recognition experiments using the KALDI toolkit resulted in a word error rate of about 11 % on the test set [9]. For any other partitioning, the authors suggest to include PA06 in the test set as it was annotated with key phrases.

8. Summary

This paper describes the collection and annotation of a new corpus of academic spoken English that consists of audio/video recordings of two series of computer science lectures at the graduate level. The data was acquired in high definition, and was edited to achieve a constant quality; there are two versions of the video available: one that shows only the presenter (including accidental parts of the blackboard and projector canvas), and a combined view that shows both the presenter and the currently displayed slide including on-screen writing. The PDF slides are available, although there exists no exact lecture to slide set alignment: some slide sets overlap multiple sessions, some sessions focus on classic blackboard oriented teaching.

In addition to the plain data, several manual annotations are available:

- The newly developed BLITZSCRIBE2 was used to transcribe the roughly 30 hours of speech in about five times real time instead of ten to 50 times real time as reported for other transcription tools. BLITZSCRIBE2 is freely available as part of the JSTK.
- The lecturer assigned a rough set of key phrases to each lecture, which can be considered a ground truth from a teaching perspective.
- For an individual lecture *PA06*, five human annotators that either observed that very lecture or a similar one in previous years extracted and ranked a set of key phrases.

The collected corpus forms a good base for future research on ASR for lecture-style, non-native speech (a significant percentage throughout the world), supervised and unsupervised key phrase extraction, topic segmentation, slide to speech alignment, and other e-learning related issues. The corpus is available for non-commercial use upon request, please contact the authors for details. Further details of the transcription and annotation process can be found in [10].

9. Acknowledgments

The authors would like to thank Prof. Dr.-Ing. Joachim Hornegger for authorizing the release of the lecture recordings and related PDF slide material. The recording, editing, media encoding and data export was done by the Regionales Rechenzentrum Erlangen (RRZE). The authors would furthermore like to thank Dr. Anton Batliner for his very valuable advice on how to structure, organize and execute a large scale data set acquisition.

[5] The British Academic Spoken English (BASE) corpus project. Developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson.

10. References

[1] P. Matejka, P. Schwarz, J. Cernocky, and P. Chytil, "Phonotactic Language Identification using High Quality Phoneme Recognition," in *Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH)*, 2005, pp. 2237–2240.

[2] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: Development and use of a tool for assisting speech corpora production," *Speech Communication*, vol. 33, no. 1-2, pp. 5–22, 2001.

[3] B. Roy and D. Roy, "Fast transcription of unstructured audio recordings," in *Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH)*, 2009, pp. 1647–1650.

[4] S. Steidl, K. Riedhammer, T. Bocklet, F. Hönig, and E. Nöth, "Java Visual Speech Components for Rapid Application Development of GUI based Speech Processing Applications," in *Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH)*, 2011, pp. 3257–3260.

[5] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[6] K. Järvelin and J. Kekäläinen, "IR Evaluation Methods for Retrieving Highly Relevant Documents," 2000, pp. 41–48.

[7] K. Järvelin and J. Kekäläinen, "Cumulated Gain-Based Evaluation of IR Techniques," vol. 20, no. 4, pp. 422–446, 2002.

[8] R. C. Simpson, S. L. Briggs, J. Ovens, and J. M. Swales, "The michigan corpus of academic spoken english," Tech. Rep., University of Ann Arbor, MI, USA, 2002.

[9] K. Riedhammer, M. Gropp, and E. Nöth, "The FAU Video Lecture Browser system," in *Proc. IEEE Workshop on Spoken Language Technologies (SLT)*, 2012, pp. 392–397.

[10] K. Riedhammer, *Interactive Approaches to Video Lecture Assessment*, Logos Verlag Berlin, 2012.