# AUTOMATIC TRANSCRIPTION OF ACADEMIC LECTURES FROM DIVERSE DISCIPLINES

Ghada AlHarbi & Thomas Hain

Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK

{g.alharbi,t.hain}@dcs.shef.ac.uk

## ABSTRACT

In a multimedia world it is now common to record professional presentations, on video or with audio only. Such recordings include talks and academic lectures, which are becoming a valuable resource for students and professionals alike. However, organising such material from a diverse set of disciplines seems to be not an easy task. One way to address this problem is to build an Automatic Speech Recognition (ASR) system in order to use its output for analysing such materials. In this work ASR results for lectures from diverse sources are presented. The work is based on a new collection of data, obtained by the Liberated Learning Consortium (LLC). The study's primary goals are two-fold: first to show variability across disciplines from an ASR perspective, and how to choose sources for the construction of language models (LMs); second, to provide an analysis of the lecture transcription for automatic determination of structures in lecture discourse. In particular, we investigate whether there are properties common to lectures from different disciplines. This study focuses on textual features. Lectures are multimodal experiences - it is not clear whether textual features alone are sufficient for the recognition of such common elements, or other features, e.g. acoustic features such as the speaking rate, are needed. The results show that such common properties are retained across disciplines even on ASR output with a Word Error Rate (WER) of 30%.

***Index Terms***— automatic speech recognition, lecture transcription, lecture analysis, text analysis, perplexity.

## 1. INTRODUCTION

Processing of academic lecture recordings has become an area of interest for both research and application communities [1]. The main interest is driven by the use of transcripts in downstream processes such as, for example, information retrieval and summarisation, or for education research [2]. Such technologies can allow applications which simply target accessibility, e.g. through search, or far more complex use. Several internet video-streaming platforms already hold recordings of many lectures and seminars, but such material is not searchable and requires the user to simply watch for the length of the recording.

Giving access by text is desirable for any student, especially if lectures cover a whole course. Text may permit applications which are interactive, and allow a student to compile work packages from spoken text, a new form of learning. Furthermore, written words may also enable students with severe disabilities to participate for the first time, either through automatic note-taking or by displaying the spoken words in real time. But not only students can benefit; lecturers can analyse their presentations [3]. Clearly, an important pre-step for any further applications which can benefit the learning communities is an analysis of the lecture style with respect to the lecture discipline.

As resources are still sparse, research into the acoustic and linguistic components of lectures for ASR purposes is still incomplete. Most resources as yet are either only available to a limited community or do not provide an insight into the wide range of topics across different disciplines inherent to the field.

This work reports for the first time ASR results for lectures from different disciplines using the LLC corpus. In particular, we show different behaviour for different lectures with respect to ASR performance and model accuracy. Furthermore, collections of in-domain and out-of-domain resources are chosen for building language models (LMs) to minimise the Out-of-Vocabulary (OOV) rate as demonstrated in section 2. In this study we further investigate the existence of common properties across different lecture domains. If such features exist one can move towards finding structure within lecture streams regardless of discipline, and as postulated by literature. Structure then allows linking and contextualisation. However, our initial focus is constrained to finding textual features, to show their existence, and investigate whether they are common across disciplines.

Lectures are multimodal experiences. It is not clear whether textual features alone are sufficient for recognition of such common elements or other features, e.g. acoustic features such as speaking rate, are needed. The results presented in this study show that such common features and their generic properties are retained across disciplines even on ASR output with a Word Error Rate (WER) of 30%.

The following sections present an overview of the related literature about ASR for the lecture domain, as discussed in section 1.1. Then, section 1.2 describes general data properties as confirmed in most related literature. After that, there is a discussion in section 1.3 about lecture structure and how different academic lectures could share the same properties. In addition, section 2 provides a detailed description about resources used in this study along with some statistics about LLC corpus as mentioned in section 2.1.

### 1.1. ASR for lectures

One of the largest publicly available data collections in this field is the Corpus of Spontaneous Japanese (CSJ) [4]. However, the lecture style appears to differ substantially from that present in English lectures. Research on English academic talks was first reported in [1] on the publicly available Translanguage English Database (TED) [5]. Several large-scale projects addressed lecture speech, e.g. the European project Computer in the Human Communication Loop (CHIL ) [6] and the LECTRA project for transcribing Portuguese lectures [7]. The iCampus spoken lecture-processing project is an ongoing project at MIT [8]. As in that project, often lecture-processing is seen in a wider context. Lecture slide synchronisation is, for example, part of an application and, if available, such slides allow the improving of recognition performance (e.g [9]).

### 1.2. General data properties

Lecture speech typically is spontaneous in style, and hence contains speech disfluencies such as hesitations, mispronunciations, partial words, filled pauses, and non-grammatical constructs. Such artefacts are known to form a challenge for speech recognisers to perform accurately [7, 8]. Moreover, in a modern academic environment it is quite frequent that speakers may not be presenting in their mother-tongue, which adds another challenge for ASR to deal with. Finally, the recording quality may vary depending on the recording equipment [8] and the acoustic room conditions. By considering all of these challenges, the unadapted ASR error rates have been reported to easily exceed 40% (e.g. [8]).

### 1.3. Structure of lectures

The information presented in lectures follows a logical sequential structure [11]. Lecturers often restate a key idea many times to highlight the importance of some information. Pauses are often used to express transitions between topics [11]. As outlined by Flowerdew in [11], most academic lectures have common features within the lecture structure such as starter, informative description about the topic, and meta-statements, and might contain aside talks and a conclusion or summary. Such acts can be interpreted as a model for a traditional lecture monologue.

However, this structure is difficult to identify due to the complexity of the relationships between different topics and the lack of visual signals such as titles, headings, and subheadings, which are used normally in the written discourse, to detect topic shift [11]. As reported there, one option to recognise the lecture structure is by identifying discourse markers that may signal the point at which there is a change from one topic to another in the given discourse. Identification of such discourse markers, referred to as topic-shift markers, may provide a structural basis for dividing the lecture into smaller units of different topics. Schiffrin has defined discourse markers as sequentially dependent elements which bracket units of talk' [11]. This method of topic identification is called structural analysis of the lecture material [11].

The rest of the paper is organised as follows: In section 3 we described our baseline system and how the acoustic and language model built accordingly. Furthermore, in section 4 we start to investigate whether different academic lectures have the same properties. If such properties are independent of the topics discussed, one can start the modelling of such structures and improve recognition and metadata labelling. Our analysis is based on wider topical groups (academic disciplines or fields of science). We investigate specific vocabulary and language models for these disciplines and measure commonality and difference. Finally, an investigation was made to see whether ASR outputs can hold lectures common properties or not as demonstrated in section 4.3.

## 2. LLC CORPUS

The main lecture resource used in this study was provided by the Liberated Learning Consortium (LLC)[1];. The LLC is an international research network dedicated to improving access to information through speech-recognition-based captioning and transcription systems [12]. In addition to ongoing applied research and development of various systems, a longstanding objective is to improve recognition performance and reduce WER for transcription. To allow improving of recognition performance, the Consortium is developing a lecture corpus comprised of digitalised and transcribed lectures from a wide variety of academic disciplines. As a member of the Consortium, Sheffield University is using this corpus to develop lecture transcription systems for public use on webasr [13].

Aside from these resources, we make use of text from three other collections of lecture speech: the Michigan Corpus of Academic Spoken English (MICASE) consists of 1.8 million words of transcribed speech in a variety of events at the University of Michigan[2]; the British Academic Spoken English (BASE) corpus contains transcriptions of 160 lectures and 39 seminars recorded at the Universities of Warwick and Reading[3]. The aforementioned TED corpus [5] is also used.

| Corpus | #Word | # Uniq. Word |
|---|---|---|
| In-Domain sources | | |
| LLC | 1.4M | 24.5k |
| BASE | 1.5M | 30k |
| MICASE | 600k | 17k |
| TED | 72k | 4k |
| Out-Of-Domain sources | | |
| Wed(CHIL) | 68M | 300k |
| Web(RT07-lect) | 40.5M | 250k |
| HUB4-LM96 | 131M | 200k |
| Web(Fisher-conv) | 500M | 700k |
| Web(ami-rt05) | 78M | 300k |

**Table 1**. Number of words and unique words for different resources in million & thousands.

| Discipline | Dur. | # Word | # Uniq. Word |
|---|---|---|---|
| Arts & humanities | 0.34 | 2854 | 722 |
| Biological Science | 0.34 | 4004 | 666 |
| Business | 0.34 | 3950 | 832 |
| Education | 0.26 | 2677 | 737 |
| Engineering | 0.34 | 3942 | 855 |
| English | 0.64 | 6414 | 1130 |
| Medicine | 0.60 | 6130 | 1281 |
| Physical Science | 0.34 | 3746 | 547 |
| Psychology | 0.34 | 3999 | 994 |
| ALL | 3.59 | 37649 | 4138 |

**Table 2**. Duration and number of words for different disciplines (on the *eval* set).

### 2.1. Corpus statistics

The LLC corpus is a continually growing resource. At this point we have made use of 247 academic lectures covering different topics with a total of 150 hours of speech. Some recorded lectures are in compressed MP3 format. Lecturers are mostly native English speakers from three main accent groups: North American, British, and Australian English. The data was recorded at seven different universities and hence has different recording conditions and qualities. Unfortunately, there is almost no information available about the recording conditions in the reference transcriptions. The topics are wide-ranging and in order to perform analysis we have grouped them by academic discipline: art and humanities, biological science, business, education, engineering, English, medicine, physical science, and psychology. The length of individual lectures ranges from 38 minutes to more than one hour. Interestingly, the LLC corpus includes timings per word. In order to get significant coherent spurts of speech, adjacent words without silence gaps were merged

[1] http://liberatedlearning.com

[2] http://quod.lib.umich.edu/m/micase/

[3] http://www2.warwick.ac.uk/fac/soc/al/research/collect/base

| LM | PPL *dev* | PPL *eval* | %WER |
|---|---|---|---|
| Lect | 128.9 | 129.9 | 37.2 |
| Lect+Wed(chil) | 114.6 | 114.1 | 35.1 |
| Lect+Web(rt07) | 118.0 | 117.5 | 35.5 |
| Lect+HUB4 | 118.3 | 117.8 | 35.5 |
| Lect+SWB(cts) | 127.9 | 128.3 | 36.9 |
| Lect+Web(fisher) | 112.8 | 114.9 | 35.2 |
| Lect+Web(ami) | 116.0 | 110.9 | 34.6 |
| ALL | 111.7 | **110.4** | **30.2** |

**Table 3**. Perplexities on *dev* & *eval*, %WER on *eval*. *Lect* represents in-domain corpora. ALL denotes training on all in-domain and out-of-domain resources.

to form pseudo-sentences for all subsequent processes (acoustic and language modelling; test set scoring). The total number of words is 1.4M, with 24k unique words, as illustrated in Table 1.

The data was split further into a training and test part. The training set covers all disciplines, excluding medicine. The reason for this exception was simply lack of data, as only a few medical lectures are contained in the corpus and it was decided to include those in the test set. The final training set covers 133 hours of speech. The complete test set consists of 17.2 hours of speech and was chosen to hold approximately equal amounts of speech per discipline. As domain adaptation experiments were performed the test material was further split into development (*dev*) and evaluation (*eval*) sets. The *eval* set covers 3.59 hours of speech and consists of the last 10 minutes from all lectures in the complete test set. The remainder of the test set (13.5 hours) formed the *dev* set. Table 2 gives detailed statistics for each of the nine disciplines on the *eval* set. Note that the number of unique words varies between 14 and 27% of set sizes.

## 3. ASR FOR LECTURES

The purpose of this paper is not to show how the best recognition performance can be achieved, but to explore the inherent properties of the data. Thus, both acoustic and language modelling follows standard paradigms in speech recognition. A standard front-end configuration of 12-MF-PLP coefficients and $c_0$ was used [14]. First and second order derivatives were added to form a 39-dimensional feature vector. No further feature normalisation was included. Acoustic models are phonetic decision tree tied state-clustered triphone Hidden Markov Models (HMMs), with Gaussian mixture models representing state output distributions. All acoustic models are trained using HTK[4] under a maximum likelihood criterion and a standard mix-up procedure. The best performance was obtained with models holding 7427 tied states, with a fixed number of 16 mixture components per state. LMs are standard trigram models using Kneser-Ney discounting and standard backoff and are trained with the SRI LM toolkit[5]. Pronunciations are derived from a background dictionary containing pronunciations for over 136k words, which are based on the UNISYN dictionary [15], and manual augmentation[16]. For the training dictionary, further pronunciations for 736 words were manually added using the BOB toolkit [16]. As the strategy for selecting words for the test dictionary (section 3.1) does not require producing additional pronunciations, no additional effort was required for the recognition stage. All recognition experiments are performed with HTK HDecode.

[4] `http://htk.eng.cam.ac.uk`

[5] `http://www.speech.sri.com/projects/srilm`

| Wordlist Source | # Uniq. Word | *dev* | *eval* |
|---|---|---|---|
| Lect | 36116 | 0.7 | 0.8 |
| LLC | 24487 | 0.7 | 0.8 |
| BASE | 30321 | 1.1 | 1.4 |
| MICASE | 17296 | 1.8 | 2.2 |
| TED | 4243 | 4.1 | 6.2 |
| HUB4 | 216418 | 0.3 | 0.4 |

**Table 4**. %OOV of different resources specific vocabularies.

### 3.1. Vocabulary selection

In large vocabulary ASR, typically vocabulary is selected from in-domain sources and potentially augmented with out-of-domain words as described in [17]. Here we follow the same strategy: all words in the training set of the LLC corpus are included, and combined with the words from the BASE, MICASE, and TED corpora. As these sources cover a significant amount of text, no further resources were included. The resulting vocabulary size was 36116, which thus is the size of the recognition dictionary used in all experiments.

### 3.2. Language modelling

While text material for written words is available in abundance, LM training data for conversational speech is typically sparse. Thus, language models have to be constructed from other sources. The sources involved are typically a mix of spoken and written words which represent a fit to the target domain in varying degrees. Hence individual LMs are interpolated (as in e.g. [17]). A set of in-domain and out-of-domain sources was chosen for interpolation as listed in Table 1. The out-of-domain resources are the same as those used for meeting transcription in the AMI systems [17] and are partially based on web data collection [18]. In total, more than 800M words of text are used. Interpolation weights are optimised using maximum-likelihood optimisation on the *dev* set. Table 3 shows perplexity results on the LLC *dev* and *eval* sets. *Lect* denotes the combination of all in-domain resources.

### 3.3. Decoding and performance

All recognition is performed in a single pass without adaptation using HTK HDecode with trigram language models, in first-best mode. Table 3 shows results for a system trained on roughly 132 hours of data for different interpolated language models. Pooling of all lecture-related material is given by *Lect*. One can observe a steady improvement by inclusion of more and more background resources. The inclusion of the Fisher web data (obtained on the basis of telephone conversations) gives good results. The best result both in WER and perplexity is a combination of all sources; the WER improvement from just using lecture data for LM construction is 7.0% absolute and 18% relative.

## 4. LECTURE ANALYSIS

The WERs shown above are surprisingly low, given that the acoustic model structure and the system are very basic. One can expect that state-of-the-art adaptive recognition configurations will lower the errors at least to the region of 20% or less. Nonetheless, for some purposes this may still be too high [13]. In this section, we analyse academic lectures from different perspectives. At first we have looked at the resulting OOVs rates when building the interpolated LMs as well as the interpolated weights for each discipline. Secondly, we

| Category | %WER |
|---|---|
| Arts & humanities | 26.4 |
| Biological Science | 14.4 |
| Business | 26.8 |
| Education | 43.7 |
| Engineering | 24.7 |
| English | 37.0 |
| Medicine | 39.6 |
| Physical Science | 20.4 |
| Psychology | 32.0 |

**Table 5**. % WER for each discipline.

| Discipline | %SR | %H | %FS |
|---|---|---|---|
| Arts & humanities | 2.40 | 1.8 | 0.4 |
| Biological Science | 3.47 | 0.7 | 0.5 |
| Business | 3.37 | 0.6 | 0.5 |
| Education | 3.08 | 0.4 | 0.4 |
| Engineering | 3.29 | 1.9 | 0.3 |
| English | 2.86 | 2.0 | 1.1 |
| Medicine | - | - | - |
| Physical Science | 3.08 | 0.2 | 0.2 |
| Psychology | 3.41 | 1.1 | 0.4 |

**Table 6**. Number of Speaking rate (SR), Hesitations (H), False Start (FS) on LLC *train* set per discipline.

investigate whether different lectures from different disciplines have common properties within its structure. In this work, we defined properties as a coherent behaviour of the given discourse. In other words, it is considered as an indication of the organisation of information in lecture discourse by looking at different parts of lectures such as beginning, middle, and end. Finally, we investigate whether such properties are still sustained even with a total error rate of %30.

### 4.1. Vocabulary and language

The language models constructed associated with the results in Table 3 were based on a wordlist constructed on *Lect* data only, with a vocabulary size of about 36k words. As previous studies have highlighted, vocabulary is an important issue. Table 4 shows OOV rates for wordlists from in- and out-of-domain sources. While some OOV rates are relatively high, by combining the in-domain sources the OOV rate on the *dev* and *eval* sets lowers to only 0.7 - 0.8%. However, the wordlist is significantly different from one obtained from broadcast news (HUB4) sources. Existing research has applied different approaches in order to minimise the perplexity (PPL) by adapting both the topic and style within the lecture transcription, which have a great impact in the resulting PPL of the interpolated LM [10]. The perplexity in their study was 143 in the test and 127 in the development set respectively. In contrast, results on LLC data as shown in Table 3 indicate that the lower perplexities are 111 and 110 on the development set and test sets respectively. Thus, combining different resources for building and interpolating the LM could be sufficient to reduce the perplexity.

### 4.2. Disciplines and structure

The aim in this section is two-fold; at first, we want to analyse academic lectures from different disciplines in order to find out any differences or similarities between them. Second, we want to prove that there are properties common to different lectures, such as the distribution of words or the complexity of the language at the introductory part being different from the middle and end part of the lecture stream.

| dev set | lect | chil | rt07 | hub | ami | fish. |
|---|---|---|---|---|---|---|
| Arts&Hum. | 0.75 | 0.05 | 0.01 | 0.03 | 0.02 | 0.12 |
| Bio._Sci. | 0.75 | 0.07 | 0.04 | 0.02 | 0.05 | 0.07 |
| Business | 0.51 | 0.08 | 0.05 | 0.10 | 0.06 | 0.20 |
| Education | 0.57 | 0.06 | 0.04 | 0.15 | 0.04 | 0.14 |
| Engineering | 0.53 | 0.07 | 0.05 | 0.03 | 0.15 | 0.16 |
| English | 0.59 | 0.04 | 0.01 | 0.02 | 0.03 | 0.30 |
| Medicine | 0.51 | 0.06 | 0.02 | 0.05 | 0.03 | 0.33 |
| Physical_Sci. | 0.85 | 0.09 | 0.02 | 0.00 | 0.01 | 0.03 |
| Psychology | 0.65 | 0.05 | 0.02 | 0.03 | 0.03 | 0.22 |

**Table 7**. Interpolated weights for different disciplines using different LMs.

| dev set | dev | eval | dev$_d$ | eval$_d$ |
|---|---|---|---|---|
| Arts & humanities | 132.7 | 112.4 | 121.9 | 102.1 |
| Biological Science | 104.3 | 90.1 | 98.2 | 82.7 |
| Business | 111.6 | 109.9 | 110.1 | 105.3 |
| Education | 115.7 | 109.2 | 115.4 | 108.1 |
| Engineering | 111.7 | 97.7 | 107.0 | 101.3 |
| English | 135.2 | 111.2 | 125.0 | 112.3 |
| Medicine | 112.0 | 105.3 | - | - |
| Physical Science | 79.1 | 100.6 | 68.1 | 91.0 |
| Psychology | 98.4 | 95.7 | 97.8 | 96.1 |

**Table 8**. Cross-perplexities using Interpolated LM on *eval* & *dev* sets for each discipline data as a dev set. *d* denotes using only data & LM specific to each discipline.

One way to analyse lecture discourse is to investigate the ASR performance, in particular WER for each academic discipline. Another important metric is the interpolation weights when building the interpolated LMs. For this particular aim, we have split the data into nine disciplines and for each discipline we took one academic lecture which varies in length. The selection of the data sources was based on LLC corpus since it covers sources from several academic disciplines, as outlined in Table 2. Thus, for this particular experiment we used our standard baseline system as described in section 3. For constructing the LM we made use of in-domain and out-of-domain resources as illustrated in Table 1.

Thus, Table 5 shows %WER results for each of the disciplines using the interpolated language model of all in-domain and out-of-domain resources described in Table 3. Biological Science with 14 %WER yields the best performance, which indicates that the language model has been optimised perfectly for this discipline. Furthermore, an analysis of some acoustic characteristics for different disciplines were provided as demonstrated in Table 6. Speaking rate for Biological Science is the highest among other disciplines whereas English lecture has the highest hesitations and false start rates in contrast to others. One could argue that this is related to the ASR performance for each discipline. On the LM side, Table 7 provides the interpolated weights of different disciplines using different sources for building the LM. As indicated in this table the LM mostly optimised the in-domain resources. One could argue here that this optimisation effect is due to the selection of the development set. Furthermore, Table 8 indicates that optimising the LM specifically for each discipline will reduce the perplexity further down. The results above indicate inhomogeneity across disciplines. However, in the following we have analysed the lecture according to the variations of PPL.

The second aim in this section is to show that there are common properties among different academic lectures. One needs to address the kind of metrics used to prove this assumption. For this study, we make use of PPL measurement on different parts of lec-

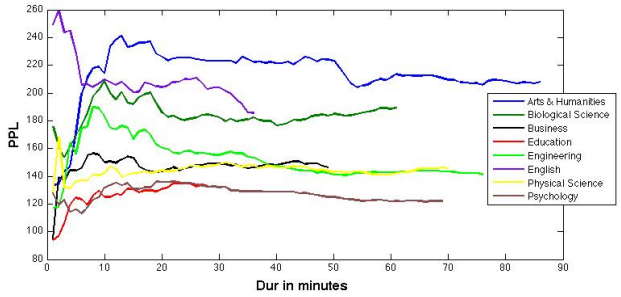


**Fig. 1**. Perplexities change as we go further along the lecture stream, applied on a complete set of eight lectures from different disciplines.

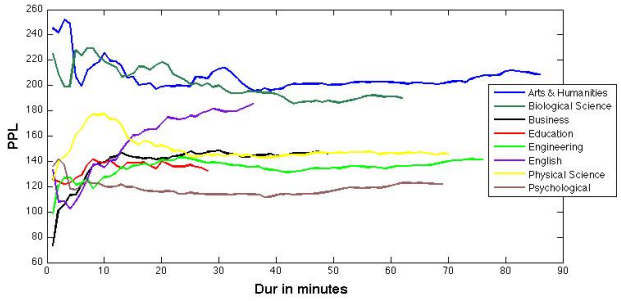


**Fig. 2**. Perplexities measures from the end of the lectures to the beginning (in reverse), applied on a complete set of eight lectures from different disciplines.

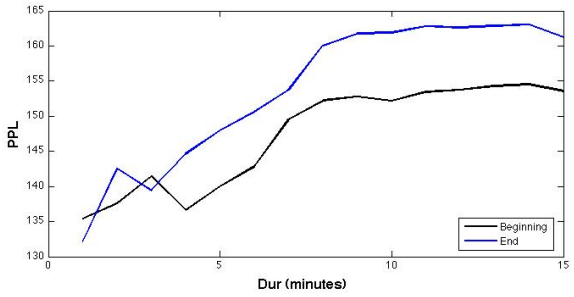


**Fig. 3**. Perplexities measures from both the beginning and end of the lectures until the middle part, applied on the reference transcriptions for a combination of eight different lectures.

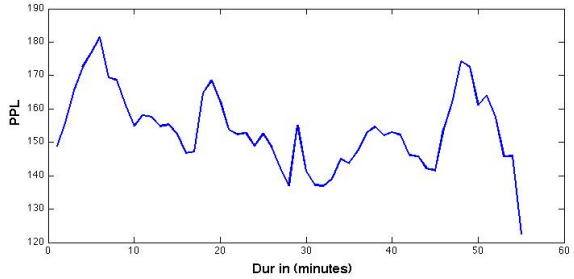


**Fig. 4**. Perplexities measures using a sliding window of five minutes from the beginning to the end, applied on the reference transcriptions for a combination of eight different lectures.

ture transcription. Thus, we rely on PPL since it is used to measure the complexity of speech recognition task, in particular LM performance [19]. The data used here is the reference transcription of a complete set of eight lectures from different disciplines which vary in length. For each lecture, we tried to measure the PPL performance for each segment of the given transcription. The length of the segment here is increased by one minute as we go further along the lecture stream. Another configuration was made by combining all the segments from different lectures together and then measuring the PPL for each of the combined segments. The length of the segment here is again started by one minute and then increased by one minute as well as we go further along the lecture stream. As a last experiment, we used a sliding window of five minutes length and it moved by one minute along the lecture stream. Then we measured the PPL for each window. For this particular experiment, we applied the same principle of combining the segments of different lectures together as explained before. It is important to mention that the LM used in such experiments was optimised on the BASE corpus instead of the LLC development set, since the LLC development set was used here for evaluation purpose only. This decision was made because the LLC development set contains a whole lecture in contrast to the LLC evaluation set as described in section 2.1.

For this particular experiment, our preliminary results indicate homogeneity across disciplines. As stated in section 1.3 the beginnings and endings of presentations differ. Hence we have studied perplexities from those sections only, in contrast to the central discussions. Fig. 1 shows perplexities of the initial part of the complete set of eight lectures, when increasing these on a per-minute-of-speech basis. Fig. 2 shows the perplexities of the end part of the lecture. These two experiments were applied individually on each academic lecture. However, naturally estimates are poor with small amounts of text, but the perplexities tend to increase until reaching the average, given by the middle part. This seems to be generally a common property, across all disciplines.

Fig. 3 and Fig. 4 demonstrate the integration effect when we combine segments from different lecture streams and then measure their PPL. In addition, Fig. 4 applied the concept of the sliding window as explained above. Thus, the results on these figures indicate that there are common properties between different lectures, since PPL tended to increase in the initial part of lectures and decrease at the end parts. However, for the middle parts, it seems stable across disciplines.

### 4.3. ASR transcription

In this section we investigate the usefulness of ASR outputs to find the common properties of lecture disciplines. Surprisingly, it turns out that even with a total WER of 30% the ASR outputs can be useful, even for any text-analysis tasks. Fig. 5 shows such behaviour of PPL on ASR outputs and it mostly resembles Fig. 3 of the reference transcription. In Fig. 6, we applied the concept of windowing as described above but here on ASR output. Again, the PPLs performance tended to be similar as in the case of the reference transcriptions as in Fig. 4 . These experiments prove that ASR output is capable of catching such coherence property of the lecture discourse.

## 5. CONCLUSIONS

In this paper we reported the results of a large-scale ASR system using LLC data with 30.2 % WER. In this context, an attempt was made to analyse lectures from different disciplines in respect of ASR performance. In addition, we provided an analysis of the OOV rate for different disciplines. A second important element addressed in this study is the investigation of the properties common to different lecture disciplines. In addition, the usefulness of the machine-

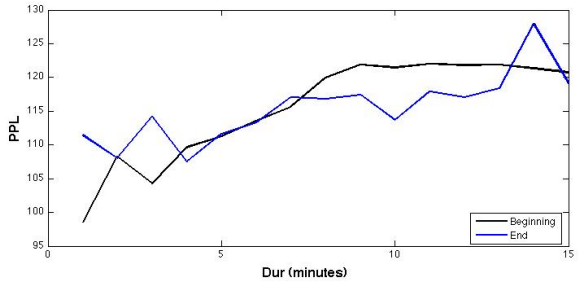


**Fig. 5**. Perplexities measures from both the beginning and end of the lectures until the middle part, applied on the ASR transcriptions for a combination of eight different lectures.

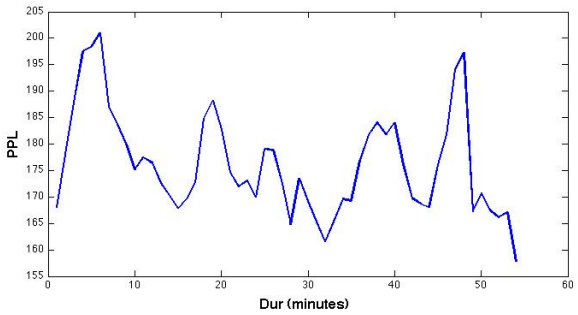


**Fig. 6**. Perplexities measures using a sliding window of five minutes from the beginning to the end, applied on the ASR transcriptions for a combination of eight different lectures.

generated text (ASR transcription) in such a task is seen. The findings highlight the importance of the discourse analysis of lecture speech, which gives a hint about lecture structure and how that could have an impact on any advancing tasks such as text segmentation or summarisation.

## 6. ACKNOWLEDGEMENTS

The authors would like to thank the members at Liberated Learning Consortium (LLC) for providing the lecture data which has been used in this study. In addition, the data used in this study came from the British Academic Spoken English (BASE) corpus project. The corpus was developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson. Corpus development was assisted by funding from BALEAP, EURALEX, the British Academy, and the Arts and Humanities Research Council.

The research leading to these results was in part supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

## 7. REFERENCES

[1] E. Leeuwis, M. Federico and M. Cettolo, "Language modeling and transcription of the TED corpus lectures", Proc. ICASSP'03, Hong Kong,2003.

[2] M. Wolfel and S. Burger, "The ISL baseline lecture transcription system for the TED corpus", Tech. Rep., Karlsruhe University, 2005.

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

[4] S. Furui, "Recent advances in spontaneous speech recognition and understanding", Proc. IEEE Workshop on Spont. Speech Proc. and Rec., Tokyo, Japan, 2003.

[5] L. Lamel, F. Schiel, A. Fourcin, J. Mariani and H.Tillman,"The translanguage English database (TED)", Proc. ICSLP, 1795-1798, Yokohama, Japan, 1994.

[6] L. Lamel, G. Adda, E. Bilinski and J. Gauvain,"Transcribing lectures and seminars", Proc. Interspeech'05, Lisbon, Portugal, 2005.

[7] I. Trancoso, R. Nunes, L. Neves, C. Viana, H. Moniz, D. Caseiro and A. Mata,, "Recognition of Classroom Lectures in European Portuguese", Proc. Interspeech'06, Pittsburgh, USA, 2006.

[8] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh and R. Barzilay, "Recent Progress in the MIT Spoken Lecture Processing Project", Proc. Interspeech'07, Antwerp, Belgium, 2007.

[9] C. Munteanu, G. Penn and R. Baecker, "Web-Based Language Modelling for Automatic Lecture Transcription", Proc. Interspeech'07, Antwerp, Belgium, 2007.

[10] B. J. Hsu and J. Glass, "Style and topic language model adaptation using HMM-LDA", Proc. ACL Conf. on Empirical Methods in NLP EMNLP, 373381, Sydney, Australia, 2006

[11] J. Flowerdew [Ed], "Academic listening: Research perspectives", NY: Cambridge University Press, Cambridge,1994.

[12] K. Bain, S. Basson, A. Faisman and D. Kanevsky, "Accessibility, transcription and access everywhere", IBM Systems Journal, 44(3), pp. 589603, 2005.

[13] T. Hain, A. El Hannani, S. Wrigley and V. Wan, "Automatic speech recognition for scientific purposes - webASR", Proc. Interspeech'08, pp. 504-507, 2008.

[14] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, J. Acoust. Soc. Amer., vol. 87, no. 4, pp. 17381752, Apr. 1990.

[15] S. Fitt, "Documentation and user guide to UNISYN lexicon and post-lexical rules', Tech. Rep., Centre for Speech Technology Research, Edinburgh, 2000.

[16] V. Wan, J. Dines, A. El Hannani and T. Hain, "Bob: A lexicon and pronunciation dictionary generator", Proc. Spoken Language Technology Workshop, Goa, India, 2008.

[17] T. Hain, J. Dines, G. Garau, D. Moore, M. Karafiat, V. Wan, R. Oerdelman and S. Renals, "Transcription of conference room meetings: an investigation", Proc. InterSpeech'05, Lisbon, Portugal, 2005.

[18] I. Bulyko, M. Ostendorf and A. Stolcke, "Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures", Proc. HLT, 2003.

[19] F. Jelinek, R. L. Mercer, L. R. Bahl and J. K. Baker, "Perplexity - A Measure of Difficulty of Speech Recognition Tasks", 94th Meeting of the Acoustic Society of America, Miami Beach, Florida, 1977.