LANGUAGE MODELING AND TRANSCRIPTION OF THE TED CORPUS LECTURES

Erwin Leeuwis[1], *Marcello Federico*[2] *and Mauro Cettolo*[2]

[1]University of Twente
Department of Computer Science
P.O. Box 217, 7500 AE Enschede, The Netherlands

[2]ITC-irst
Centro per la Ricerca Scientifica e Tecnologica
I-38010 Povo di Trento - Italy

ABSTRACT

Transcribing lectures is a challenging task, both in acoustic and in language modeling. In this work, we present our first results on the automatic transcription of lectures from the TED corpus, recently released by ELRA and LDC. In particular, we concentrated our effort on language modeling. Baseline acoustic and language models were developed using respectively 8 hours of TED transcripts and various types of texts: conference proceedings, lecture transcripts, and conversational speech transcripts. Then, adaptation of the language model to single speakers was investigated by exploiting different kinds of information: automatic transcripts of the talk, the title of the talk, the abstract and, finally, the paper. In the last case, a 39.2% WER was achieved.

1. INTRODUCTION

Automatic lecture transcription is arising as an important task both for research and applications [1, 2]. It is a challenge for speech recognition as, in contrast to broadcast news, lectures typically present a higher variability in terms of speaking style, linguistic domain, and speech fluency. From the application point of view, spoken document retrieval based on automatic transcripts has shown to be a promising mean for accessing content in audiovisual digital libraries [3]. Hence, envisaging digital repositories of recorded speeches and lectures, which can be searched and browsed through the net, is quite natural now.

A useful and publicly available resource for investigating automatic lecture transcription is given by the TED corpus, which was issued in 2002 by ELRA and LDC. Briefly, the Translanguage English Database contains 188 recordings of talks in English at Eurospeech '93, a part of which has been manually transcribed.

The lectures in TED present several kinds of problems to cope with. Speakers are often non-native, have a strong accent, and, sometimes, are not even fluent. Despite the speaking style being in general planned, spontaneous speech phenomena occur quite frequently. Recordings were made with a lapel microphone, hence the signal often contains some noise from the auditorium and from the speaker as well. Finally, relatively little supervised data is available for acoustic and language model training. For the sake of language modeling, the lack of transcripts is compensated by the availability of electronic texts of that conference.

This work describes the development of a TED baseline system at ITC-irst. Acoustic models were estimated starting from an existing baseline for American English newspaper dictation and by exploiting a small amount of TED training data. Most of the effort was put in estimating the language model (LM). Several LM adaptation configurations are reported which make use of different sources of data and adaptation techniques, both supervised and unsupervised. In particular, increasing amounts of conference papers were used together with conversational speech corpora, and transcripts provided with TED. Explored LM adaptation techniques are mixture models, minimum discrimination information, and probabilistic latent semantic analysis.

The purpose of this work is twofold. Besides reporting preliminary results on the TED database, we would like to elicit interest in this task by proposing an experimental set-up which other labs can follow. In doing so, we hope to start some benchmarking activity around the TED task.

This work was partially financed by the European Commission under the project FAME (IST-2000-29323, http://isl.ira.uka.de/fame/index.html).

2. TED CORPUS

The Translanguage English Database corpus consists of 48h audio recordings of 188 lectures given by, often non-native, English speakers at Eurospeech '93. Of the total lectures, 39 are provided with manual transcripts. Also included are information about the recorded speakers, and electronic versions of over 400 papers presented at the conference.

# speakers	eng.	ger.	lat.	other	n.a.	fem.	mal.
transcribed	5	12	12	6	4	7	32
in test set	1	3	3	1	0	2	6

Table 1. Test set composition in terms of native language groups (English, Germanic, neo-latin, others, not available) and gender.

The 39 manually transcribed lectures were divided in a test set of 8 speakers (2 hours of speech) and a training set of 31 speakers (8 hours of speech). Test speakers were selected by taking into account the proportion of each native language group and gender (Table 1). The test set speakers are listed in Table 2.

3. BASELINE SYSTEM

The ITC-irst transcription system (Fig. 1) features a Viterbi decoder, context-dependent cross-word continuous-density HMMs, MLLR adaptation, and a trigram LM.

The system has been applied to several large vocabulary tasks: Italian broadcast news [4], American English broadcast news (HUB4)

speaker	native language	gender
cj29s3	english	male
dc57s2	italian	male
fd29s5	french	male
hb64s4	french	female
ld29s2	danish	female
ph50s2	german	male
ro31s4	dutch	male
yi59s5	japanese	male

Table 2. Test set speaker identifier, mother tongue, and gender.

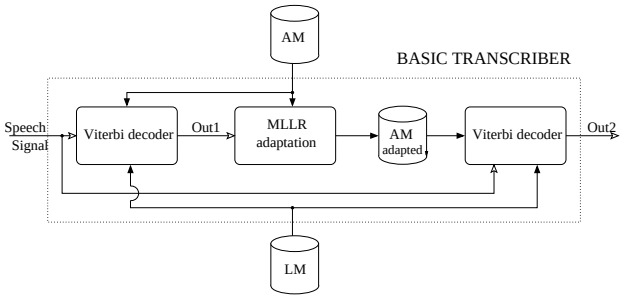

Fig. 1. Architecture of the ITC-irst transcription system.

and newspaper dictation (Wall Street Journal, WSJ).

3.1. Acoustic Model

The acoustic model (AM) for TED was developed starting from a WSJ baseline, featuring 27K triphone units and 71k Gaussians trained on 66.5h of speech. By using the standard 20k-word trigram LM, the WSJ baseline scores a 12.9% WER on the 1993 DARPA evaluation test set. The WSJ AM was adapted on the TED training data (8 hours) through MLLR adaptation. In this step, spontaneous speech phenomena were mapped into a single filler model.

4. LM ESTIMATION AND ADAPTATION

For LM estimation, three different types of data were used:

`Lect` 55Kw of lecture transcripts from the TED training data;

`Proc` 15Mw of scientific papers from speech conferences and workshops (Eurospeech, ICASSP, ICSLP, etc.);

`Conv` 300Kw of transcripts of conversational speech (Verbmobil, HUB5).

The `Lect` corpus has the most suitable data, but unfortunately is rather small. Therefore bigger corpora are also used that are less suitable, but have useful qualities: `Proc` does not have the required style, but has suitable content (speech research); `Conv` on the contrary, does not have suitable content, but has the required style (conversational).
LMs estimated for the TED task make use of trigram statistics and are based on a recursive interpolation scheme and non-linear smoothing [5]. For the sake of LM estimation, three different LM adaptation methods have been investigated.

Mixture Model (MIX). Given two or more interpolated language models, a mixture model can be derived which applies a convex combination at the level of discounted relative frequencies [5]. The mixture model can be used to combine one or more general background (BG) LMs with a foreground (FG) LM representing new features of the language we want to include. In this case, the mixture weights can be estimated on the foreground data by applying a cross-validation scheme that simulates the occurrence of new n-grams [5].

Minimum Discrimination Information (MDI). Assuming a small adaptation text sample, one may reasonably assume that only unigram statistics can be reliably estimated. These statistics can be used as constraints when estimating the adapted LM as the one minimizing the Kullback-Leibler distance from a background trigram model. Practically speaking, the adapted n-gram conditional probability is obtained by scaling and normalizing the background LM distribution. As shown in [6], an empirically estimated exponent (adaptation rate) can be applied to the scaling factor to improve the effect of adaptation. This adaptation rate has a value between 0 and 1, with 0 corresponding to no adaptation and 1 to full adaptation.

Probabilistic Latent Semantic Analysis (PLSA). PLSA can be interpreted as the problem of estimating a kernel of r unigram distributions which better fits the word distribution of each document, in a collection $\mathcal{D}$, through a suitable convex combination [6]. Assuming that $\mathcal{D}$ contains documents talking about different topics, the compression effect induced by the model should force semantically related words, e.g. words associated with a specific topic, to have meaningful probabilities concentrated in one or few basis distributions. An appealing feature of PLSA is that a document/topic word distribution can be estimated from a small amount of adaptation data relatively easily. Combination of MDI with PLSA naturally follows given that the PLSA distribution estimated from the adaptation data can be used to constrain a higher-order background LM [6]. In this way, statistically sound constraints about a trigram LM can be derived from very little data.

5. EXPERIMENTS

5.1. Baseline Development

The baseline system for transcribing the TED lectures is that of Fig. 1, with the AM developed as explained in Section 3. Interpolated LMs estimated on corpora `Lect`, `Proc` and `Conv`, described in Section 4, have been mixed in different combinations in order to explore the relationship between their characteristics and transcription performance.
In Table 3, results in terms of perplexity (PP), out of vocabulary rate (OOV) and word error rate (WER) are reported for different mixture models. In particular, for each mixture model, the foreground and background models are indicated. For the sake of comparison, the first two rows show the performance of the recognizer developed for the WSJ task, and of the recognizer using the TED AM and the WSJ LM.
Since in terms of PP and OOV rate its results are the best, and its recognition accuracy is not worse than the best one in a statistically significant way, the LM of the last row was selected as baseline LM. Intuitively, we assume that it adapts the style of `Conv` and the content of `Proc` to suit `Lect`, which is the most proper data for

AM	LM			PP	OOV	WER
	FG	BG_1	BG_2		(%)	(%)
WSJ	WSJ	-	-	1240	5.33	93.2
TED	WSJ	-	-	1240	5.33	59.7
TED	Lect	-	-	634	8.07	56.3
TED	Proc	-	-	288	1.51	46.3
TED	Proc	Conv	-	239	0.93	45.1
TED	Proc	Lect	-	218	0.55	45.2
TED	Lect	Proc	-	202	0.55	43.9
TED	Lect	Proc	Conv	197	0.53	44.0

Table 3. Baseline recognizer performance by using various LMs.

this task. The baseline LM has a dictionary of 36Kw; the 44.0% WER was achieved using the basic transcriber with a real time ratio of 65 on a Pentium III 933 MHz processor.

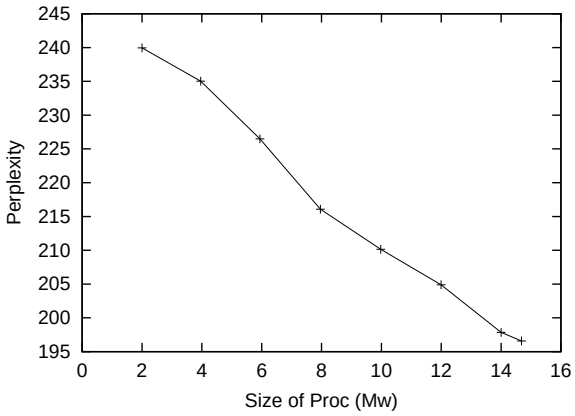

Fig. 2. PP as function of the LM estimation corpus size.

In Fig. 2, the relationship of the PP of the baseline LM with the size of the `Proc` corpus is plotted. It shows that increasing the amount of proceedings used, decreases perplexity significantly. Thus, we expect PP to go further down when more proceedings will be used.

5.2. Unsupervised LM adaptation

A first set of experiments aimed at improving the baseline performance by adapting the LM on each single test lecture. In particular, unsupervised LM adaptation was carried out on the automatic transcripts output by the baseline [7]. Actually, also AM adaptation was performed again, which leads to the adaptation scheme depicted in Fig. 3.

MIX adaptation was applied by extending the baseline mixture with a new component estimated on the automatic transcript. For estimating the mixture weights, the new component was taken as foreground model. MDI adaptation was performed in the same way by only extracting unigram statistics from the transcript. In order to smooth the effect of recognition errors, words in the transcripts with frequency below 2 were mapped into the out-of-vocabulary word class [5]. The best performance was achieved with an adaptation rate of 0.7.

PLSA adaptation was based on a set of 100 kernel distributions

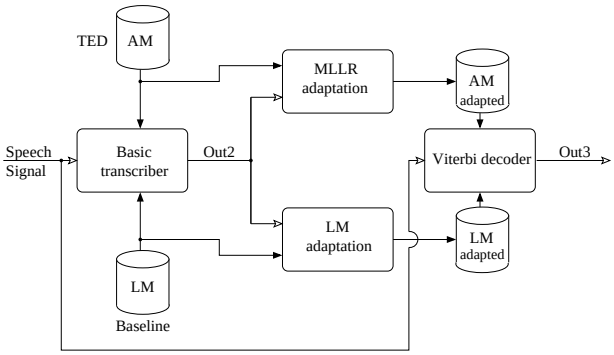

Fig. 3. Unsupervised LM adaptation experiments scheme.

estimated on the `Proc` corpus, which includes over 6,000 documents. As adaptation data the 10 most frequent non-stop words in the transcript were used. The unigram mixture estimated from the kernel distributions and the adaptation data was then used for MDI adaptation. This time the optimal adaptation rate was 0.2.

In order to reduce the bias of perplexity measures after unsupervised adaptation, perplexity computation of MIX and MDI was not performed on the whole transcript, but using a leaving-one-out scheme. The transcript was split at sentence level; iteratively, a sentence was left out of the adaptation data and that sentence was used to compute perplexity on. Finally, the resulting perplexities were combined. Results of the experiments are reported in Table 4.

	Base	MIX	MDI	PLSA
PP	197	157	170	190
WER	44.0	44.3	43.9	43.8

Table 4. Unsupervised LM adaptation per speaker.

Even though the leaving-one-out strategy should reduce the bias, there is a decrease in PP for MIX and MDI that is not reflected in the WER. Perhaps the PPs are still biased on sentence level, but probably the discrepancy is due to the significantly higher probability assigned to recognized n-grams. From the WER point of view, performance does not change substantially, as the LM is suggesting the same n-grams the recognizer produced in the previous step.

Hence, a reduction of the bias could be achieved by filtering out less frequent words from the transcript or by using only unigram statistics, as is done by the MDI and PLSA adaptation methods. In general, we expect that the availability of more transcribed material or, alternatively, of multiple quite independently produced transcripts of the same data should help to reduce the bias.

5.3. Supervised LM adaptation

Supervised LM adaptation was performed using instead the presented paper or parts of it to adapt the baseline LM. In order to assume an increasing amount of supervision, adaptation was performed just on the title (PLSA), on the abstract (PLSA), or on the full paper (PLSA, MDI, MIX). PLSA adaptation was applied by using the same kernel distributions estimated for the unsupervised adaptation experiments. MIX adaptation extended the baseline

components with an additional LM estimated on the adaptation data and used as foreground model.
Results for each approach are given in Table 5. As expected, performance became better when the amount of supervision increased.
Very marginal improvement is achieved with PLSA adaptation, probably due to the fact that papers in the collection are not easily decomposed into very distinct topics.

				PLSA		
	Base	Mix	MDI	Paper	Abstract	Title
PP	197	133	166	188	190	193
WER	44.0	**39.2**	42.3	43.8	43.9	44.2

Table 5. Supervised LM adaptation.

The other two methods instead gave reasonable improvements in terms of PP and WER. Fig. 4 and 5 show the PP and WER respectively for each speaker using the baseline LM and using both MIX and MDI supervised adaptation. For each speaker and each method both PP and WER decrease significantly. There is a strong correlation between the difference in PP and in WER. Speakers CJ and YI show bigger improvements with mixture adaptation than the other speakers, since they held lectures in a style similar to their papers.

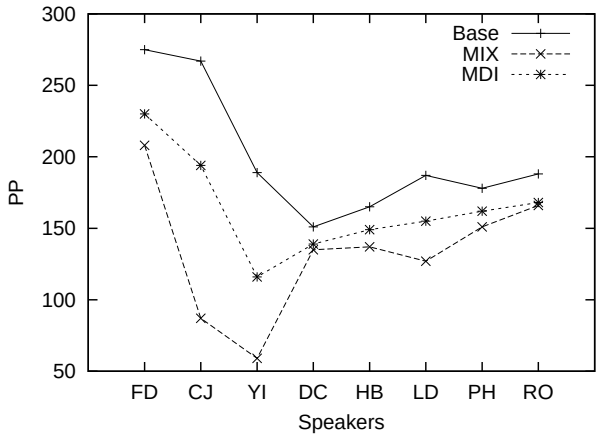

Fig. 4. PP after supervised adaptation per speaker.

6. CONCLUSION

Lecture transcription is a difficult task, both from an acoustic and a linguistic point of view. Non-native speech, background noise, different and varying speaking rates and many spontaneous speech phenomena, are all characteristics of lecture speech that make acoustic modeling difficult. Language modeling is hampered due to the sparseness of suitable data and the mixed style of lecture spoken language, combining colloquial expressions with formal jargon.
In this work, we concentrated our effort on language modeling. A baseline LM was estimated using various types of data, which were all flawed, but used in such a way that their qualities were highlighted and not their deficiencies. Using the ITC-irst WSJ AM

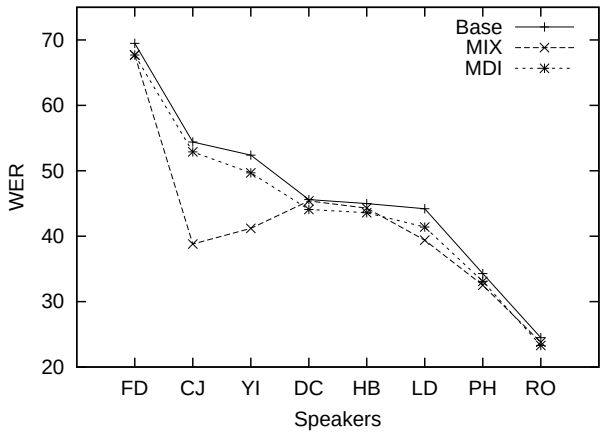

Fig. 5. WER after supervised adaptation per speaker.

adapted on 8h of TED training data, it resulted in a WER of 44.0%. Unsupervised LM adaptation did not show mentionable improvements in WER, but the decreases in perplexity indicate that future research could prove beneficial. Significant improvements were obtained by adapting the baseline LM on the papers of the speakers: 39.2% WER. That represents a good starting point for further research developments. Future work will be devoted to investigate acoustic and lexical modeling for non-native speech, and unsupervised adaptation/training methods for acoustic and language modeling, for which there are 38 hours of untranscribed speech available in the TED corpus.

7. REFERENCES

[1] M. Novak and R. Mammone, "Use of non-negative matrix factorization for language model adaptation in a lecture transcription task," in *Proc. ICASSP*, Salt Lake City, UT, USA, 2001.

[2] H. Nanjo and T. Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," in *Proc. ICASSP*, Orlando, FL, USA, 2002.

[3] F. Kubala, S. Colbath, D. Liu, A. Srivastava, and J. Makhoul, "Integrated technologies for indexing spoken language," *Communications of the ACM*, vol. 43, no. 2, pp. 48–56, 2000.

[4] N. Bertoldi, F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani, "From broadcast news to spontaneous dialogue transcription: Portability issues," in *Proc. ICASSP*, Salt Lake City, UT, 2001.

[5] M. Federico and N. Bertoldi, "Broadcast news LM adaptation using contemporary texts," in *Proc. Eurospeech*, Aalborg, Denmark, 2001.

[6] M. Federico, "Language model adaptation through topic decomposition and MDI estimation," in *Proc. ICASSP*, Orlando, FL, USA, 2002.

[7] D. Giuliani and M. Federico, "Unsupervised language and acoustic model adaptation for cross domain portability," in *Proc. ISCA Workshop on Adaptation Methods for Speech Recognition*, Sophia-Antipolis, France, 2001.