ADVANCES IN THE AUTOMATIC TRANSCRIPTION OF LECTURES

Mauro Cettolo, Fabio Brugnara, and Marcello Federico

ITC-irst
Centro per la Ricerca Scientifica e Tecnologica
I-38010 Povo di Trento - Italy
{cettolo,brugnara,federico}@itc.it

ABSTRACT

Transcribing lectures is a challenging task, both in acoustic and in language modeling. In this work, we present recent results on the automatic transcription of lectures from the Translanguage English Database, which contains the recordings of talks given in English at Eurospeech '93, by mostly non-native speakers.

Concerning acoustic modeling, the acoustic model trained for a broadcast news transcription task was adapted on the lectures training data through Maximum Likelihood Linear Regression adaptation, including models of spontaneous speech phenomena. Moreover, a normalization procedure was embodied in the training stage, consisting in a cluster-based mean and variance normalization of the static features.

Language modeling was based on adaptation of a background language model estimated on broadcast news transcripts, conference proceedings, lecture transcripts, and conversational speech transcripts. Among the examined adaptation techniques, the most effective one was obtained by exploiting the paper presented in each lecture to be processed.

The best transcription performance on a 2 hours test set was 32.4% word error rate.

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 the Translanguage English Database (TED) [4], which was issued in 2002 by ELRA and LDC. Briefly, TED 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 recent improvements of the TED transcription system at ITC-irst, whose baseline was described in [5]. After the overview of the methods experimented for language model (LM) adaptation, the paper presents subsequent activities carried out to improve the acoustic model (AM) for lecture transcription.

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

TED consists of 48h audio recordings of 188 lectures given in English by mostly non-native 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.

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

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

3. BASELINE SYSTEM

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

The system has been applied to several large vocabulary tasks: Italian broadcast news [6], American English broadcast news (HUB4) and Wall Street Journal newspaper dictation (WSJ).

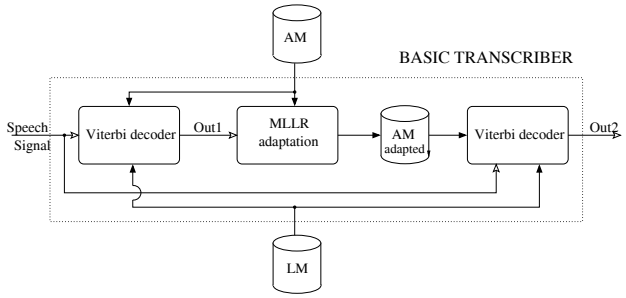

Fig. 1. Architecture of the ITC-irst transcription system.

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 [7]. 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 [7]. 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 [7].

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 [8], an empirically exponent (adaptation rate), estimated on a development data set, 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 [8]. 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 [8]. In this way, statistically sound constraints about a trigram LM can be derived from very little data.

4.1. Experiments

4.1.1. Baseline Development

The baseline system for transcribing the TED lectures is that of Fig. 1. The 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% word error rate (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.

Interpolated LMs estimated on corpora `Lect`, `Proc` and `Conv`, described at the beginning of this section, have been mixed in different combinations in order to explore the relationship between their characteristics and transcription performance.

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	-	-	279	1.51	46.3
TED	Proc	Conv	-	230	0.62	45.1
TED	Proc	Lect	-	215	0.55	45.2
TED	Lect	Proc	-	200	0.55	**43.9**
TED	Lect	Proc	Conv	**194**	0.53	44.0

Table 3. Baseline recognizer performance by using various LMs.

In Table 3, results in terms of perplexity (PP), out of vocabulary rate (OOV) and 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 this task. The baseline LM has a dictionary of 36Kw.

4.1.2. 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 [9]. Actually, also AM adaptation was performed again, which leads to the adaptation scheme depicted in Fig. 2.

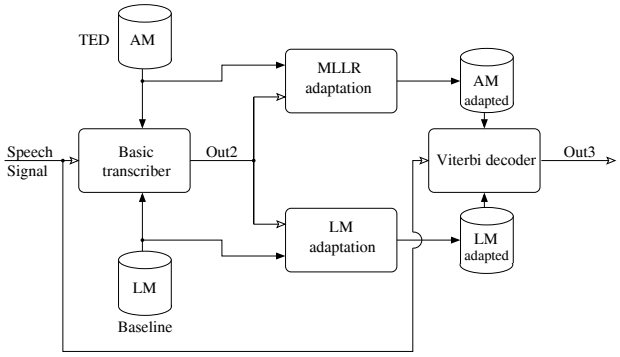

Fig. 2. Unsupervised LM adaptation experiments scheme.

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 [7]. The best performance was achieved with an adaptation rate of 0.7.
PLSA adaptation was based on a set of 100 kernel distributions 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.

	Base	MIX	MDI	PLSA
PP	194	**168**	177	187
WER	44.0	44.3	43.9	**43.8**

Table 4. Unsupervised LM adaptation per speaker.

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. 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.

4.1.3. 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	194	**120**	157	185	187	190
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.

5. AM ESTIMATION AND ADAPTATION

The baseline AM has been improved by an explicit modeling of spontaneous speech phenomena, and by using a normalization technique on a rich training corpus.

5.1. Investigated methods

5.1.1. Spontaneous speech phenomena

In the recognizer used for assessing experimental results on language modeling, spontaneous speech phenomena were mapped into a single filler model. In order to model them in a more detailed way, seven different non-linguistic phenomena have been trained: one for silence, that means absence of any linguistic sound, and six for filled pauses, that are vowels and voiced consonants (like long "m", "e", "o" etc.) which are typically uttered when one thinks out what to say. In Table 7, this updated AM is referred with the label `WSJ+ssp`. The same extra-linguistic phenomena have been modeled for the HUB4 AMs described in the following.

5.1.2. Improvements in acoustic modeling

The AM used for experiments of Section 4.1 was trained on the WSJ corpus. Meanwhile, we acquired the HUB4 corpus, which contains almost 200 hours of transcribed audio data, coming from several American news programs. These data were used to build a much richer model set, with an augmented acoustic resolution, and

trained on varying acoustic conditions. The addition of training data required to enlarge the lexicon, which was built by merging different sources: the LIMSI '93 lexicon, the CMU lexicon and the PRONLEX lexicon. A first $HUB4_1$ model set was built by using the same procedure used for the WSJ models, and using the union of the HUB4 data with the WSJ data for training.

	#models	#Gauss	#Tied states	training amount	WER %
$HUB4_1$	32.7K	171.1K	10.9K	$\approx$ 204h	22.9
$HUB4_2$	28.0K	145.7K	9.1K	$\approx$ 150h	21.2

Table 6. Features of the model sets built on HUB4 data, and results on HUB4 Eval'98.

Thanks to the increase in training data, these models already produced a significant improvement in performance, but it was felt that both the lexicon and the training procedure could be improved to better exploit the large size of the corpus. Therefore, new rules were designed for combining the source lexicons into the global lexicon. These provided a smaller but more consistent representation of pronunciation alternatives. Moreover, a normalization procedure was embodied in the processing stages, consisting in a cluster-based mean and variance normalization of the static features. Each cluster in training and test had zero mean and unity variance. Segmentation and clustering in training were completely automatic and data-driven, while in testing each lecture was considered a cluster. With normalized data, new models were built from scratch, including the redesigning of the phonetic decision tree-based state tying. The WSJ data were not included anymore in training, as some experiments showed that they did not give any benefit. This new AM will be referred in the following as $HUB4_2$.

Table 6 details the characteristics of $HUB4_1$ and $HUB4_2$ model sets, along with the performance obtained on the HUB4 '98 evaluation set with a single step of decoding (without adaptation).

AM	LM				PP	WER (%)
	FG	BG_1	BG_2	BG_3		
WSJ	Lect	Proc	Conv	-	120	39.2
WSJ+ssp	Lect	Proc	Conv	-	120	38.4
WSJ+ssp	Lect	Proc	Conv	HUB4	113	37.6
$HUB4_1$	Lect	Proc	Conv	HUB4	113	35.5
$HUB4_2$	Lect	Proc	Conv	HUB4	**113**	**32.4**

Table 7. WER with different configuration of the recognizer.

5.2. Experimental results

Table 7 summarizes performance obtained with different configurations of the recognizer. For the sake of comparison, the first row of the table gives the best WER reported in the previous section (see Table 5). The second row allows to quantify the improvement obtained by modeling spontaneous speech phenomena, while the WER of the third row has been obtained by introducing in the LM mixture a new component estimated on the HUB4 transcripts (130 Mw). The updated LM improves the previous one in terms of PP by the following values: from 194 to 180 for the baseline, and from 120 to 113 for the LMs adapted to the papers of each speaker through the MIX method.

The last two rows show the WERs of the recognizer employing the updated LM and the $HUB4_1$ and $HUB4_2$ AMs, respectively.

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 paper, we have described our efforts on both language and acoustic 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. The use of the ITC-irst WSJ AM adapted on 8h of TED training data, yielded a WER of 44.0%. Unsupervised LM adaptation did not show worth mentioning improvements in WER, but the decreases in perplexity indicate that future research could prove beneficial. Significant improvements (39.2% WER) were obtained by adapting the baseline LM on the papers of the speakers.

Afterwards, efforts have been devoted in order to improve acoustic modeling. First of all, the AM derived from a dictation task (WSJ) has been replaced by models built for a broadcast news task (HUB4). The rationale behind this choice is that the speaking style of broadcast news speakers, although "controlled", is closer to that of lecturers than the read speech of dictation tasks. In fact, the new AM allowed to reach 32.4% WER, which represents an almost 14% relative WER reduction with respect to the best result obtained with the WSJ AM (37.6%).

7. REFERENCES

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

[2] H. Nanjo and T. Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," *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] `www.ldc.upenn.edu/Catalog/LDC2002S04.html`

[5] E. Leeuwis, M. Federico, and M. Cettolo, "Language modeling and transcription of the TED corpus lectures," *ICASSP*, Hong Kong, 2003.

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

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

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

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