Automatic Transcription of Courtroom Recordings in the JUMAS project

Daniele Falavigna[1], Diego Giuliani[1], Roberto Gretter[1], Jonas Lööf[2], Christian Gollan[2], Ralf Schlüter[2], and Hermann Ney[2]

[1] FBK-Irst - Via Sommarive 18, 38050 Povo, Trento, Italy
{falavi,giuliani,gretter}@fbk.eu
[2] Lehrstuhl für Informatik 6 - Computer Science Dept.
RWTH Aachen University, Aachen, Germany
{loof,gollan,schluter,ney}@cs.rwth-aachen.de

Abstract. In this paper we present ongoing work on speech recognition for the judicial domain, performed in the European project JUMAS (Judicial management for digital library semantics.) The specific challenges for courtroom speech recognition are discussed, and the development of speech recognition systems for Italian and Polish are described. The results achieved on the target domain are presented and discussed.

1 Introduction

This paper presents work performed in the context of the JUMAS project [1], a European Union project aimed at information extraction and indexing in the judicial domain, specifically for the processing of court recordings. As part of the project Polish, and Italian automatic speech recognition (ASR) system for the domain of court proceedings is being developed.

State of the art ASR systems allow to achieve good performance (approximately word error rates inferior to 10%) when the speech signal is acquired in controlled conditions. However, as demonstrated in recent DARPA evaluations [2] performance significantly decreases if transcription tasks include speech coming from non professional speakers, uttering their sentences in "free" conditions, where the recording environment is not optimal.

Table 1. *State of the art ASR word error rates for different application domains.*

	WER
connected digits	$\leq 0.5\%$
continuous dictation	$\leq 5\%$
studio broadcast news	$\leq 10\%$
telephone news reports	$\leq 20\%$
telephone conversations	$\leq 30\%$
meetings (head mounted microphone)	$\leq 30\%$
meetings (distant microphone)	$\leq 50\%$

Table 1 gives a general idea on the level of performance that can be presently reached with state of the art ASR systems. As can be seen from the Table, Word Error Rate (WER) increases as more degrees of freedom are allowed in the speech signal to recognize.

1.1 Courtroom Speech Recognition – Challenges

The courtroom environment presents most of the phenomena listed in the previous section. In addition, the audio recorded in courtrooms is acquired with several distinct microphones, generally one for each of the actors of the trial, located on fixed positions with respected to the speakers who are free to move inside the room (sometimes we have observed speakers uttering their sentences in the direction opposite to the microphone assigned to them). This type of acoustic environment originates high levels of noise and reverberation in the speech signal, reducing the signal-to-noise ratio and introducing non linear distortions which are difficult to remove and which are known to be detrimental to good ASR performance.

A further problem relies on the language that is used: it comes from sentences spontaneously uttered, syntactically wrong, containing a large number of hesitations, pauses and false starts. Often, speakers are non native or make use of strong dialectal inflections. Foreign words are also frequently present, especially referring foreign people.

2 System Development

2.1 Italian System

For the development of the Italian automatic speech recognition system for the JUMAS Project, no in-domain acoustic data was yet available; the first prototype was trained using acoustic data from the broadcast news domain. But a significant set of text resources, mainly consisting of transcriptions of trials in several Italian Courtrooms are available: this allowed to train different language models using in-domain text corpora.

An initial set of recordings, about 30 hours of radio news programs, has been provided by RAI, the major Italian broadcast company. These recordings were manually segmented, labelled and transcribed and used to train a preliminary version of an automatic broadcast news speech transcription system for Italian [3]. Successively, about 100 hours of TV broadcast news have been collected and automatically transcribed with the. The resulting partially supervised corpus of about 130 hours of audio recordings was used to train the acoustic models employed in the Italian system.

The system consists of two main components: the audio partitioner and the speech recognizer. The aim of the audio partitioner is to divide the continuous audio stream into homogeneous non-overlapping segments and to cluster these segments into homogeneous groups. From this process, each audio file is divided in a set of temporal segments, each with a label that indicates its nature and the cluster to which it belongs (e.g. speaker A, speaker B, etc.) The speech recognizer, which uses continuous density Hidden Markov Models (HMMs), generates a word transcription for each speech segment.

The Italian speech transcription system makes use of two decoding steps: for each clusters of segments, the output of the first step allows to estimate parameters of linear transformations utilized in the second decoding step (for feature normalization and acoustic model adaptation) to maximizing the system performance [4].

Language Resources
Three different 4-gram LMs were used for the experiments in this paper, all estimated using improved Kneser-Ney smoothing [5]. A first LM was trained on a 606M word news text corpus, containing about 1.2M unique words. This LM will be referred as the "Out of Domain" (OD) LM. A second LM was trained on a 25M word corpus of court transcriptions, containing about 150K unique words; this LM will be referred as the "In Domain" (ID) LM. A third LM was estimated by adapting the OD LM with the in-domain data. This latter LM will be referred as the "Adapted" (AD) LM. For each of the three LMs we evaluated the number of 4-grams in the training corpus, the Perplexity (PP) on the test set, and the out of vocabulary (OOV) word rate on the test set. The statistics are presented in Table 2, together with the dictionary sizes of the LMs.

Table 2. *Statistics of the three LMs used in ASR experiments for Italian*

LM	# 4-grams	OOV(%)	PP	dict. size
OD	23.5M	0.49	471	1.2M
ID	2.3M	1.50	250	150k
AD	12.4M	0.43	272	1.2M

The lexicon employed in the Italian transcription system is based on the SAMPA phonetic alphabet, and includes a total of 85 phone-like units. Of these units, 50 are needed for representing the Italian language, while the remaining 35 are needed for representing foreign words. The lexicon was first produced with an automatic transcription tool, and then manually checked to correct possible errors in the transcription of acronyms and foreign words.

2.2 Polish System

In this section, the development of the Polish Automatic Speech Recognition (ASR) system is described. Due to the limited availability of in-domain data at the start of the project, the first efforts on a Polish ASR system was performed on the domain of political (parliamentary) speeches.

Language Resources
In a European Parliament Plenary Session (EPPS) different languages of the EU are spoken, and simultaneously interpreted into every official language of the EU. Starting at the time of the (now finished) TC-STAR project and continuing since then, the recordings of the parliamentary sessions have been collected.

For Polish, several hundreds of hours of parliament recordings are currently available. From this a black-out period was chosen, and a half hour tuning set as well as a three hour development set were extracted and transcribed by native Polish speakers, see Table 3. In addition a development set consisting of audio data from the Court of Wroclaw, was defined at RWTH and transcribed as above. The data of this corpus is also included in Table 3.

Table 3 also describes the EPPS acoustic recordings used for (unsupervised) acoustic training for the current system. This data was taken from outside of the blackout period, and included both original politician speeches as well as interpreter audio. Since this data is completely untranscribed, the word statistics are taken from the automatic transcription output.

For Polish, only the transcriptions of the politician portions of the recordings, and not the interpreted portions, totaling about half a million running words,

Table 3. *Polish acoustic corpora, statistics.*

	EPPS Tune	EPPS Dev	WCC Dev	Train
Net Duration	0.45h	3.03h	2.66h	127.8h
# Segments	195	1326	1904	40995
# Speakers	9	37	49	–
# Running words	2944	21938	21938	788098

are available. Since this is clearly inadequate for language model (LM) training, several additional sources of text data were used. The additional data consisted of official Polish translations of European Union legal documents, as well as news articles collected over the web from two Polish news sources, see Table 4.

Table 4. *Text data used for Polish language modeling.*

Source	Running Words
European Parliament	481 k
EU Legal Documents	29425 k
Kurier Lubelski (News)	15364 k
Nowosci (News)	27720 k

Lexicon and Language Model
Since Polish is a highly inflected language, the out of vocabulary (OOV) rate is typically much higher than that for a language such as English for the same vocabulary size. Since good ASR performance require an OOV rate of about one percent or lower, it is necessary to use an increased vocabulary size when working in Polish.

To achieve this, four different vocabularies were used, approximately of sizes 75, 150, 300 and 600 thousand words, respectively. For each of the vocabulary sizes a three-gram language model using modified Kneser-Ney smoothing was produced. Separate models trained for each of the four portions were combined using interpolation tuned on the perplexity on the tuning corpus. For the pronunciation lexicon the Polish SAMPA phonemes consisting of 37 phonemes were used. The pronunciations for the vocabulary were generated using letter to sound rules described in [6].

Unsupervised Acoustic Training Using Cross-language Bootstrapping
The development of the Polish acoustic model using cross-language bootstrapping and unsupervised training is described in the following section. A more detailed description is available in [7].

Cross-language bootstrapping is the technique of initializing acoustic model training using a acoustic model originally trained on a different language. For the present system a Spanish European Parliament acoustic model, described in [8], was used as a starting point. As described in [9], for cross-language bootstrapping, a mapping from the target language phoneme set (in our case Polish) to the source language phonemes (Spanish) is needed.

Both the source and target model used SAMPA phoneme sets. A manual mapping was constructed by keeping the SAMPA phoneme symbol if present in both phoneme sets, and using the Spanish phoneme with the most similar properties for the remaining 14 Polish phonemes. Once a mapping is available it is possible to use the Spanish acoustic model in combination with the Polish pronunciation lexicon for acoustic model retraining, and even for recognition (with a high error rate).

The thus initialized acoustic model was further improved using unsupervised training. The basic idea of unsupervised training is to improve an acoustic model by iterated recognition and retraining on training data for which no manual transcriptions are available. For effective use of available acoustic data, it is important to use confidence measures to select or weight the contributions of the audio data in such a way that correctly recognized data is more likely to contribute to the modeling. For the present work, the state posterior confidence method, as presented in [10] was used.

3 Experiments

3.1 Italian

The acquisition of the Italian audio baseline for the JUMAS Project is still under way. At the moment we are writing this paper about 4 hours of audio recordings have been collected in the Court of Naples, with the goal being about 30 hours. We plan to use 10 hours for development and test sets, and the remaining 20 hours for acoustic training/adaptation purposes. All the audio recordings are acquired with 4 microphones; one each for the judge, witness, prosecutor and lawyer. The sampling frequency is 16kHz, the precision is 16 bit per sample.

The four above mentioned audio tracks of Naples were automatically transcribed and are going to be manually corrected. From the automatic transcriptions some statistics, reported in Table 5, were estimated, namely:

1. the number of speech segments in each audio track and in total, detected using a start-end-point detection procedure.
2. the average speech segment duration in each audio track and in total;
3. the total speech duration in each audio track and in total;
4. the number of uttered words in each audio track and in total.

Table 5. *Statistics of first* 4 *hours of the Italian audio baseline.*

	# voice segments	avg. segment durat.	tot. durat.	# running words
judge	697	8.6s	1.7h	8501
prosecutor	663	7.9s	1.4h	4466
witness	629	9.1s	1.6h	7301
lawyer	493	7.3s	1.0h	3192
total	2482	8.3s	5.8h	23460

Note in Table 5, that the total duration ($5.8h$) of the automatically detected speech segments is significantly higher than the total duration of the trial recording, which is about $4h$. This is due to speaker overlap between the different microphone channels, and also due to speech being recorded and (automatically) detected as speech in multiple channels simultaneously.

Till now, of the available audio data, only the prosecutor audio track has been completely manually transcribed, together with about $50m$ of the other 3 tracks. On the small audio data set, formed by the manually transcribed parts ($50m$) of the 4 audio tracks, we evaluated the same statistics as for the full set. These are reported in Table 6. We can note a slight higher average duration of the manually detected speech segments compared to the automatic ones (given in Table 5), suggesting to try to reduce the triggering thresholds of the start-end-point detection module.

Table 6. *Statistics of manually transcribed part of Italian audio baseline.*

	# voice segments	avg. segment durat.	tot. durat.	# running words
judge	83	11.5s	16m	1433
prosecutor	124	9.6s	20m	1053
witness	121	11.4s	23m	2260
lawyer	120	9.0s	18m	1924
total	448	10.3s	77m	6670

A pilot ASR experiment has also been carried out on the small data set reported in Table 6, to get some hints on both the possible level of performance achievable for this domain and on what could be critical parameters to tune for the overall automatic transcription system. The obtained results, although lacking of "statistical significance", due to the small size of the test set, are given in Table 7, for each of the 3 LMs decribed in section 2.1.

Table 7. *Performance of Italian system using the three LMs of Table 2.*

	%WER I decoding step	%WER II decoding step
OD	61.5	58.9
ID	56.6	53.9
AD	56.3	53.8

Although the test data set is small, large improvements are necessary if we want to deliver an ASR technology reliable for the judicial domain. In particular, the improvement between the first and second decoding pass is smaller than that obtained on other application domains (typically about 20% relative WER improvement), probably due to the high absolute values of WERs. Furthermore, from the Table are evident the benefits from using in-domain data for LM training/adaptation, giving hope to further improvements if also in-domain data are used for acoustic model training/adaptation.

The final experiment, we report in this paper, was carried out on the four hour audio track of the prosecutor, with transcriptions available. Table 8 reports the statistics of this latter audio track derived from both the automatic and manual segmentations, and the corresponding word error rates, measured in the second decoding steps.

Table 8. *Statistics and WER (%) for prosecutor audio track.*

	# voice segments	avg. segment duration	total duration	# running words	%WER II decoding step
manual seg.	699	8.7s	1.7h	4952	57.1
automatic seg.	663	7.9s	1.4h	4466	59.6

Results of Table 8 still shows a high absolute value of WER, higher than that obtained with state of the art automatic transcription of meetings with distant microphones (see Table 1).

3.2 Polish

The unsupervised retraining of the acoustic model was performed as several recognitions and retrainings on about 130 hours of untranscribed recordings from the European Parliament. On the original Spanish task, the bootstrap

model achieves an error rate of approximately 10%. Using cross-language bootstrapping without retraining, the initial error rate is 60%. Several iterations of retraining are necessary to achieve adequate performance. In Table 9 the recognition performance on the EPPS Tuning set after the different retraining steps, as well as the amount of data selected by confidence thresholding, are summarized.

Table 9. *Results during unsupervised training and adaptation – EPPS Tuning Set.*

Training step	Train. data [h]	Data sel. [h]	WER [%]
Initial Spanish AM	n.a.	n.a.	63.4
First MAP iter.	1.9	1.7	49.6
Second MAP iter.	60.4	29.8	37.1
First training iter.	60.4	59.2	29.9
Second training iter.	60.4	53.9	26.9
First SAT iter.	67.5	66.1	24.1
First full data train.	127.8	103.4	23.2
SAT full data	127.8	106.0	20.7
SAT re-training	127.8	113.7	20.5
SAT re-training	127.8	111.0	20.0

The bootstrap model used vocal tract length normalized (VTLN) mel-frequency cepstral coefficient (MFCC) features, using cepstral mean normalization and linear discriminant analysis, resulting in a 45 dimensional feature vector. The system uses classification and regression tree state tying, with 4500 generalized triphone states. The acoustic models are hidden Markov models with pooled covariance Gaussian mixture model emission probabilities. A fully trained model consist of approximately 900k distributions in total.

Table 10. *Effect of vocabulary size.*

System	EPPS Dev		WCC Dev	
	OOV [%]	WER [%]	OOV [%]	WER [%]
75k+	3.54	21.1	8.16	63.7
150k+	2.09	20.1	5.27	62.5
300k+	1.39	19.8	3.86	62.3
600k+	0.51	19.3	1.54	62.3

For the first two re-estimation iterations maximum unsupervised a posteriori (MAP) adaptation, was used. The final iterations of unsupervised training were made using speaker adaptive training (SAT) with feature space maximum likelihood linear regression (fMLLR), using confidence measures for estimation. The effect of the different vocabulary sizes on out of vocabulary rate, as well as on first pass recognition error rate is shown in Table 10. As a final improvement, maximum likelihood linear regression (MLLR) adaptation was used in recognition. The final results for the two pass system on the WCC and EPPS development sets are presented in Table 11.

Table 11. *Performance of final Polish system – WER [%].*

System	EPPS Dev	WCC Dev
1st Pass	19.3	62.3
+ SAT	16.2	47.2
+ MLLR	15.6	45.8

4 Discussion and Conclusion

In this paper the ASR systems being developed in the context of the JUMAS project have been described, and the current results have been described. Results obtained on the present Italian acoustic data set are very preliminary and need to be confirmed by further experiments to be carried out on the whole set of data forming the planned Italian baseline, under acquisition. We also believe that the usage of unsupervised training methods (similarly to what done for Polish) or of lightly supervised training methods and, in general, of in domain data for training/adapting the acoustic models, can improve the present level of performance. In any case, these latter topics need to be deeply investigated.

The polish system, being until now optimized primarily for the European parliament domain, show good results for this domain. The results for the court recordings should still be considered preliminary, though. Substantial improvements are to be expected by the inclusion of in-domain training data both for the language model and the acoustic model.

5 Acknowledgements

This work was partly funded by the European Union under the FP6 project JUMAS, Contract No. 214306.

References

1. JUMAS: Judicial Management by Digital Libraries Semantics (http://www.jumasproject.eu)
2. Fiscus, J., Ajot, J.: The rich transcription 2007 speech-to-text, stt, and speaker attributed stt, sastt, results. In: Meeting Recognition Workshop. (May 2007)
3. Brugnara, F., Cettolo, M., Federico, M., Giuliani, D.: A Baseline for the Transcription of Italian Broadcast News. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Istanbul, Turkey (June 2000) 1667–1670
4. Giuliani, D., Gerosa, M., Brugnara, F.: Improved automatic speech recognition through speaker normalization. Computer Speech and Language **20**(1) (January 2006) 107–123
5. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. (1995) 181–184
6. Oliver, D.: Polish text to speech synthesis. Master's thesis, Edinburgh University, Edinburgh, UK (1998)
7. Lööf, J., Gollan, C., Rybach, D., Schlüter, R., Ney, H.: The RWTH 2007 TC-STAR evaluation system for European English and Spanish. In: Proc. Int. Conf. on Spoken Language Processing, Antwerp, Belgium (August 2007) 2145 – 2148
8. Lööf, J., Gollan, C., Ney, H.: Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a polish speech recognition system. In: Proc. Int. Conf. on Spoken Language Processing, Brighton, UK (September 2009)
9. Schultz, T., Waibel, A.: Experiments on cross-language acoustic modeling. In: Proc. European Conf. on Speech Communication and Technology, Aalborg, Denmark (September 2001) 2721 – 2724
10. Gollan, C., Hahn, S., Schlüter, R., Ney, H.: An improved method for unsupervised training of LVCSR systems. In: Interspeech, Antwerp, Belgium (August 2007) 2101–2104