EVALUATION OF AUTOMATIC TRANSCRIPTION SYSTEMS FOR THE JUDICIAL DOMAIN

J. Lööf[(1)], D. Falavigna[(2)], R. Schlüter[(1)], D. Giuliani[(2)], R. Gretter[(2)], H. Ney[(1)]

(1) Computer Science Department, RWTH Aachen University, Germany
(2) HLT research unit, Fondazione Bruno Kessler, Trento, Italy

ABSTRACT

This paper describes two different automatic transcription systems developed for judicial application domains for the Polish and Italian languages. The judicial domain requires to cope with several factors which are known to be critical for automatic speech recognition, such as: background noise, reverberation, spontaneous and accented speech, overlapped speech, cross channel effects, etc.

The two automatic speech recognition (ASR) systems have been developed independently starting from out-of-domain data and, then, they have been adapted using a certain amount of in-domain audio and text data.

The ASR performance have been measured on audio data acquired in the courtrooms of Naples and Wroclaw. The resulting word error rates are around 40%, for Italian, and around between 30% and 50% for Polish. This performance, similar to that reported for other comparable ASR tasks (e.g. meeting transcriptions with distant microphone), suggests that possible applications can address tasks such as indexing and/or information retrieval in multimedia documents recorded during judicial debates.

Index Terms— Automatic transcription, judicial domain, domain adaptation, cross-channel effects

1. INTRODUCTION

A challenging transcription scenario has been investigated within the European Project JUMAS (Judicial Management by Digital Libraries Semantics, EU contract FP7-214306, see `http://www.jumasproject.eu/` for more information). The most important goal of JUMAS project is to collect, enrich and share multimedia (audio/video) documents, annotated with embedded semantic, minimizing manual transcription activities. Actually, since the automatic audio transcription system can produce the uttered word sequence with the related time instants, it is possible to construct a flexible and effective indexing and retrieval system, even despite the presence of recognition errors, for the multimedia documents recorded during judicial debates.

Hence, JUMAS system will be used for managing the work flow and supporting information sharing and retrieving in all the different phases of the investigation and judicial decision process.

Within the JUMAS project, a set of audio recordings were carried out in the courtrooms of Naples and Wroclaw, during several trial sessions, and made available for ASR experiments. The recordings have been done using different fixed microphones located on desktops inside the room: each actor of the trial (i.e. judge, prosecutor, witness and lawyer) is assigned a particular microphone. Given this acquisition setup, the overall signal intensity, as well as the reverberation level in the audio signal, vary according to the source-microphone distances, that in turn can largely change due to the movements of the bodies of the speakers. This type of audio recording environment, as well as the particular linguistic application domain, presents many critical factors for ASR. More specifically, after a preliminary analysis we identified the following major issues to address:

1. evaluate and try to reduce the effects of both background noise and reverberation;

2. evaluate and try to reduce the effects of cross microphone interference;

3. cope with spontaneous speech and non-native speech;

4. adapt, possibly dynamically, the Language Model (LM) using domain data and information specific of a trial (e.g. proper names, dates, etc).

Above all, it is important to note that the level of reverberation, in relation to the actual signal level is often high. This latter condition is known to be severely detrimental to good ASR performance.

In this paper we will describe the automatic transcription systems developed for both Polish and Italian languages and the related recognition results obtained after having adapted them to the judicial domain.

The performance of the Italian transcription system has been evaluated on audio data, formed by about 7 hours of multiple tracks recordings, acquired in two different dates in the Court of Naples. In particular, each single audio track contains the voice of one of the actors of the trial (i.e. judge, prosecutor, lawyer and witness), where each actor can be represented by more than one speaker. This acquisition setup can induce cross-channel effects, e.g. the speech uttered by a speaker is captured, with a significant level of intensity, by more than one microphone. In this latter case the transcription system should be able to detect and discard the interfering speech segments recorded by secondary microphones.

The Polish transcription system was evaluated on about 6.5 hours of courtroom recordings. The Polish trials were also recorded using four microphones, one for each actor, but in contrast to the Italian data, the recordings were only available as mix-downs of all four microphones. This generates a significant amount of overlapped speech which poses severe problems for correct recognition.

Another issue with this setup is that since the signal from all the microphones are used at the same time, the already prominent reverberation is made more severe, since the room signal is recorded on all microphones and added in the mix-down. This is detrimental to the error rate of the system, since reverberation is known to pose problems for automatic speech recognition.

This work has been partially funded by by the European project JUMAS, under the contract FP7-214306

Table 1. *Italian IT-Apr09 development set*

	# utterances	net duration	# running words
judge	1330	54.4min	3900
prosecutor	666	40.3min	8880
witness	915	52.0min	5814
lawyer	459	24.6min	8175
total	3370	171.4min	26769

Table 2. *Italian IT-Nov09 test set*

	# utterances	net duration	# running words
judge	554	40.4min	1109
prosecutor	554	37.4min	5454
witness	966	66.4min	5739
lawyer	104	7.9min	8640
total	2178	152.1min	20942

2. ACQUISITION OF THE AUDIO DATA

The acquisition of the Italian audio data has been carried out in the following two different recording sessions, in the courtroom of Naples:

- **IT-Apr09**: consists of about 4 hours of multi-track audio recordings acquired in April 2009;
- **IT-Nov09**: consists of about 3 hours of multi-track audio recordings acquired in November 2009.

Each audio track corresponds to one of the actors of the judicial trial, i.e.: judge, prosecutor, witness and lawyer.

Tables 1 and 2 report some statistics of each audio track and in total.

The reference transcriptions have been produced in FBK using a multi-track displaying system that helps to exclude from the reference itself the cross-channel echoes (see section 3.1.1).

The available Polish in-domain audio data was divided into about 6.5 hours of recognition corpora, and about 24 hours of acoustic training data. In Table 3 the different corpora used for system evaluation and model estimation are described. Corporas **PL-Dev08** and **PL-Eval08** were recordings provided by the Court of Wroclaw, and transcribed by native Polish speakers at RWTH. The training corpus was also provided by Court of Wroclaw; it was also transcribed at the Court of Wroclaw. The corpus **PL-Dev09** was recorded as part of the Jumas project; in addition to the audio recording videos were also recorded, for use by other partners of the project. This corpus was also transcribed at RWTH.

Table 3. *Polish in-domain acoustic data (WCC)*

	PL-Dev08	PL-Eval08	PL-Dev09	Train
net duration	2.66h	3.13h	0.85h	23.8h
# segments	1904	2720	528	22866
# speakers	49	44	22	176
# running words	22318	29395	6709	184502

3. ITALIAN AUTOMATIC TRANSCRIPTION SYSTEM

For the development of the automatic transcription system no in-domain acoustic data are still available. Hence, as will be seen below, we have trained the acoustic models on data recorded in other different application domains. Instead, a significant set of text resources, mainly consisting of official transcripts of judicial trials performed in several Italian courtrooms are available: this allowed to train different language models using in-domain text corpora.

The automatic transcription system used in this work is the one developed in FBK during the past years. It works according to the following processing steps.

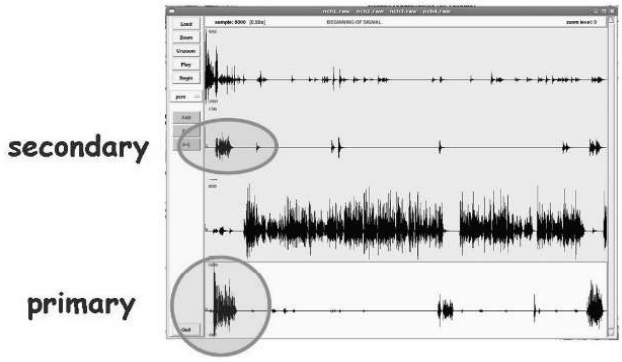

Fig. 1. Example of cross-talk effect: the same speech is present both in the primary and in the secondary channel

3.1. Audio Segmentation and Classification

In the current version of the automatic transcription system the segmenter just identifies the regions of the audio stream with high energies through the application of the multiple channel start-end point detector module described below. The detected speech segments are then classified in terms of broad acoustic classes (e.g. telephone bandwidth speech, wide band speech, music, male, female, etc). To this purpose a set of Gaussian Mixture Models (GMMs), one for each broad class, is used.

3.1.1. Cross-channel Effects

In quite all of the acquired courtroom recordings we have observed cross-channel effects: two actors can speak at the same time, each one in a different microphone, but the voices appear with a considerable volume in different channels. Sometimes the volume of the cross channel signal is so high that, when listening to a single audio track, it is hard to understand if the voice is a cross-channel effect or not. Figure 1 exemplifies this phenomenon.

Since we want to transcribe only the speech segments uttered by the primary speakers, i.e. the ones uttered by the main actor at a given time, we need to develop a system that can detect "echoes" along the primary audio tracks and remove them before sending the tracks themselves to the ASR engine for decoding.

Using a standard energy based start-end-point detector (i.e. a module that identifies inside a single audio track the time intervals where speech is present), not only the primary voice but also the cross talks can be detected.

To reduce the effects of these latter ones, since we have available multiple parallel channels, an algorithm has been implemented which loads all of the channels, computes the energy contour for each of them and, frame by frame, detects the channel having the highest energy. Basically, a local decision identifies the primary channel, which ideally corresponds to the sole channel to transcribe.

Unfortunately, there is often overlapped speech, and in these cases more than one channel should be taken into account. Furthermore, sometimes nobody is really speaking, so none of the channels should be transcribed. To cope with these phenomena, a number of parameters were introduced into the algorithm, which allow a secondary channel to be considered for transcription, or a primary channel to be discarded. These parameters are listed below:

1. number of frames per second
2. minimum accepted energy
3. accept based on min/max energy ratio
4. discard based on min/max energy ratio
5. accept based on energy percent wrt the max
6. discard based on energy percent wrt the max
7. discard based on cross-correlation.

Table 4. *Performance of the multiple channel automatic voice detection algorithm*

cost function	precision	recall	f-measure
f-measure	84.9	84.7	84.8
f-measure + recall	76.8	90.5	83.1

To estimate parameters above we adopted a greedy procedure that minimizes a given cost function (e.g. f-measure) on a manually labeled (in terms of speech/non speech) set of data. The performance reported in Table 4, in terms of frame-by-frame precision, recall, and f-measure has been computed with a resolution of 0.1 seconds and using two different cost functions in the optimization step.

3.2. Segment clustering

Acoustically homogeneous segments, previously classified, are clustered employing a method [1] based on the Bayesian Information Criterion (BIC). At the end of this process, each audio file to transcribe has assigned a set of temporal segments, each having associated a label that indicates the cluster to which it belongs (e.g. "female_1", "male_1", etc). Note that since different speakers can play the role of a given actor, e.g. different witnesses are recorded by the same microphone, speaker clustering has to be performed on each audio track.

3.3. Decoding Process

For each cluster of speech segments, the system, which makes use of continuous density Hidden Markov Models (HMMs), generates a word transcription by performing two decoding passes interleaved by acoustic feature normalization and acoustic model adaptation. Best word hypotheses, generated by the first decoding pass, are exploited for performing cluster-wise acoustic feature normalization, based on Constrained Maximum Likelihood Linear Regression (CMLLR) [2] and acoustic model adaptation. For this latter purpose, just Gaussian means of triphone HMMs are adapted through the application of up to 4 full matrix transforms estimated in the MLLR framework [3].

3.4. Acoustic Models

In both decoding passes AMs are state-tied, cross-word, speaker-independent triphone HMMs. Each HMM is characterized by a 3 state left-to-right topology, with the exception of the model associated to the background noise, which has a single state. In addition to triphones, several spontaneous speech phenomena are also modeled. Output probability distributions are modeled by mixtures of Gaussian probability density functions (PDF) having diagonal covariance matrices. A phonetic decision tree is used for tying states and for defining the context-dependent allophones.

The set of modeled speech units is based on the Italian SAMPA phonetic alphabet, and includes a total of 48 phone-like units.

Each speech frame is parametrized into a 52-dimensional observation vector composed of 13 mel frequency cepstral coefficients (MFCCs) plus their first, second and third order time derivatives. Cepstral mean subtraction and variance normalization is performed on static features on a cluster-by-cluster basis. A projection of acoustic feature space, based on heteroscedastic linear discriminant analysis (HLDA), is embedded in the feature extraction process as follows [4]. A GMM with 1024 Gaussian components is first trained on the original 52-dimensional observation vectors. Acoustic observations in each, automatically determined, cluster of speech segments, are then normalized by applying an affine transformation estimated w.r.t. the GMM through CMLLR [2]. After normalization of training data, an HLDA transformation is estimated w.r.t. a set of state-tied, cross-word, gender-independent triphone Hidden Markov Models (HMMs) with a single Gaussian per state, trained on the normalized 52-dimensional observation vectors. The HLDA transformation is then applied to project the set of 52 normalized features into a 39-dimensional feature space. Recognition models used in the first and second decoding pass are trained on these normalized, HLDA projected, acoustic features. HMMs for the first decoding pass are trained through a conventional maximum likelihood procedure. HMMs used in the second decoding pass are trained through a speaker adaptive procedure [5]: for each cluster of speech segments an affine transformation is estimated through CMLLR exploiting as target-models triphone HMMs with a single Gaussian density per state trained on the HLDA projected acoustic features. The estimated affine transformation is then applied on the cluster data [5]. Acoustic models are finally trained from scratch on the obtained normalized acoustic data.

For the Jumas project we decided to use and compare HMMs trained on different sets of out-of-domain audio data, namely:

1. an Italian Broadcast News ("IBN") corpus, about 130 hours of audio recordings mostly acquired from TV and radio broadcast news programs. The resulting triphone HMM set used in the second decoding pass has about 5,500 tied-states, for a total of about 169,300 Gaussian densities (note that similar number of parameters were allocated for the corresponding HMM set used in the first decoding pass);

2. a corpus consisting of audio recordings of Italian political speeches acquired in the Italian Parliament, about 93 hours of recordings. In the following we will refer to this corpus as: Italian Parliament Political Speeches ("IPPS"). Both HMM sets, used in the first and second decoding pass, have about 3,700 tied-states, for a total of about 120,000 Gaussian densities;

3. a corpus formed by the union of IBN and IPPS data, i.e. about $223 = 130 + 93$ hours of speech. Both HMMs sets, used in the first and second decoding pass, have about 6,900 tied-states, for a total of about 212,000 Gaussian densities. We will call this corpus "IBN+IPPS".

Table 5 shows some statistics of IBN and IPPS audio databases.

3.5. Language Models

As previously seen, for LM, differently from AM training, we have at our disposal a certain amount of in-domain text data coming from the proceedings of trials performed in some Italian courtrooms,

Table 5. *Italian audio training databases*

	IBN	IPPS
# hours	129h:30m	92h:48m
# utterances	115,024	27,041
# words	1.3M	711k
# speakers	9,542	1,744

Table 6. *Statistics of the Italian LMs*

LM	# running words	# 4-grams	OOV(%)	PP	# unique words
OD	606M	23.5M	0.49	471	1.2M
ID	25M	2.3M	1.50	250	150k
AD	-	12.4M	0.43	272	1.2M

namely: Florence, Naples, Nola and Lecce. In addition, we have a large set of out-of-domain texts, collected in the past years from several sources such as: newswire and newspaper articles, broadcast news, web news and web in general, ecc. With all of this available text data we trained three different "4-gram" based LMs.

The first one was trained on an Out-of-Domain (OD) news corpus. The corpus is mainly formed by newswire and newspaper articles.

A second LM was trained on an In-Domain (ID) corpus mainly formed by judicial proceedings.

A third LM was estimated by adapting the OD LM with the in-domain judicial data. The adopted adaptation method is the one that weighs and mix [6] the 4-gram counts of both OD and ID sources.

Table 6 reports some statistics, including perplexity (PP) and Out-Of-Vocabulary (%OOV) rate measured on "IT-Apr09" development set, for out-of-domain, in-domain and adapted (AD) LMs.

Preliminary experiments have shown that both ID and AD LMs give similar results and outperform OD LM (as one can expect), hence, all of the results given in the next section refer to the usage of AD LM.

The lexicon of the judicial domain was first generated with an automatic phonetic transcription tool, that produces the phone sequence (based on the previously mentioned 48 units) of each of the words in the recognition dictionary, then it was manually checked to correct possible errors in the transcription of acronyms and foreign words. Finally, the LM, together with the lexicon, has been used to compile a Finite State Network (FSN) that defines the search space of the decoder.

4. POLISH AUTOMATIC TRANSCRIPTION SYSTEM

The Polish transcription system is a two pass recognition system, based on HMM acoustic models, with GMM emission probabilities. The feature extraction front-end is based on vocal tract length normalized (VTLN) 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 HMMs use pooled covariances. A fully trained model consist of approximately 900k Gaussian densities in total.

The processing steps of the Polish system are similar to those of the Italian system:

1. Segmentation and Clustering
2. First Pass Recognition
3. CMLLR Estimation
4. MLLR Estimation
5. Adapted Recognition

In the following, the details of the development of the Polish system is given.

4.1. Audio Segmentation and Clustering

The audio segmentation of the Polish system is based on the output of a run of a speech recognition decoder on the complete audio file to be transcribed. The length of each speech pause, recognized as sequences of non speech events, is determined. Segment boundaries are then defined at each pause, starting with the longest pause, and continuing until no segment is longer than 30 seconds.

The clustering was performed in the same way as for the Italian system.

4.2. Lexicon and Language Model

Since Polish is a highly inflected language, the OOV rate is typically much higher than that for a language such as English for the same vocabulary size. Since good ASR performance requires an OOV rate of about one percent or better, it is thus necessary to use an increased vocabulary size when working with Polish.

To achieve this, four different vocabularies were used, with approximate sizes of 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 were trained for each of the text data sources given in Table 7, and were combined using interpolation. Interpolation weights were tuned by optimizing the perplexity on the text of the development set corpus.

As seen in Table 7, the amount of in-domain text data (rows 1 and 2) is very limited. Nevertheless, the inclusion of even such a small amount of data is beneficial, although it is expected that the availability of larger amounts of in-domain text data would improve performance noticeably.

Table 7. *Text data used for Polish language modeling.*

Source	# running words
Wroclaw Audio Transcripts	185 k
Wroclaw Reports	170 k
European Parliament	481 k
EU Legal Documents	29,425 k
Kurier Lubelski (News)	15,364 k
Nowosci (News)	27,720 k

For the pronunciation lexicon the Polish SAMPA phoneme set, consisting of 37 phonemes were used. The pronunciations for the vocabulary were generated using letter to sound rules.

4.3. Acoustic Model

The basis of the acoustic model was the cross-language unsupervised trained acoustic model described in [7]. This model was originally trained on 128 hours of untranscribed recordings of Polish, from the European Parliament.

For the development of the Polish transcription system, in contrast to the Italian system, a small amount of in-domain acoustic data was available. Due to the limited amount of data available

Table 8. *%WER achieved on both "IT-Apr09" dev set and "IT-Nov09" test set for each audio track (and in total). The HMMs used are those trained on Broadcast News (IBN)*

	IT-Apr09 pass 1	IT-Apr09 pass 2	IT-Nov09 pass 1	IT-Nov09 pass 2
lawyer	44.1	44.0	41.8	40.6
judge	37.7	37.5	33.4	31.5
prosecutor	47.5	47.4	40.4	39.7
witness	39.1	37.4	41.3	41.3
all	41.1	40.6	39.0	38.3

Table 9. *%WER achieved on both "IT-Apr09" dev set and "IT-Nov09" test set for each audio track (and in total). The HMMs used are those trained on Italian Parliament Political Speeches (IPPS)*

	IT-Apr09 pass 1	IT-Apr09 pass 2	IT-Nov09 pass 1	IT-Nov09 pass 2
lawyer	49.1	47.8	42.8	40.0
judge	40.7	39.8	42.1	37.8
prosecutor	56.2	53.8	44.4	41.9
witness	41.3	39.0	50.4	46.8
all	45.4	43.7	46.1	42.8

(c.f. Table 3) it was decided to use maximum a-posteriori adaptation (MAP) [8] to adapt the model to the judicial domain.

As for the Italian system, the second pass acoustic model of the Polish system was trained using speaker adaptive training (SAT). The system uses three acoustic models:

- First-pass model: Speaker independent model used in first recognition pass.
- Target model: Single Gaussian model, used to estimate CMLLR matrices.
- SAT model: Used for second pass adapted recognition.

Both of the recognition models need to be adapted. Since it is not a problem if the target model is not a close match to the data at hand, it was decided to keep the target model from the baseline system. Thus, the following process was used to arrive at the MAP models for the system:

1. MAP adapt the first-pass model to the in-domain acoustic training data.
2. Estimate SAT CMLLR matrices for each speaker on the new acoustic training data using the original target model.
3. Using the estimated CMLLR matrices, adapt the SAT model to the new training data.

In this way, the complete SAT based recognition system has been adapted to the new acoustic condition.

5. EXPERIMENTS AND RESULTS

5.1. Italian test data

As mentioned above, the performance of the Italian automatic transcription system has been measured on both "IT-Apr09" development set and "IT-Nov09" test set.

Table 8 gives the WERs obtained using the IBN acoustic models while results obtained with IPPS acoustic models, i.e. the ones trained on political speeches, are given in Table 9. Comparing the results reported in the two tables, it can be observed that the use of IBN models results in better performance than using the IPPS models. Overall, with the IBN models, 40.6% and 38.3% WER are achieved on the development and and evaluation sets, respectively. For both model sets, the two pass decoding system ensures a reduced margin of improvements over the one pass decoding system, especially in case of the IBN models. A possible explanation is that the acoustic models used in the first decoding pass are speaker adaptively trained, through cluster-wise acoustic feature normalization based on CMLLR, which leads to pretty good recognition models for the first decoding pass leaving a reduced margin of improvement to the second pass [9]. Furthermore, in all cases significant performance differences can be observed among the different actors.

Table 10. *%WERs obtained on the whole test set (IT-Apr09 + IT-Nov09) with: IBN acoustic models, IPPS acoustic model, with their ROVER combination and with IBN+IPPS acoustic models*

	IBN	IPPS	ROVER (IBN,IPPS)	IBN+IPPS
lawyer	43.4	46.3	43.4	43.4
judge	35.2	39.0	35.1	35.1
prosecutor	43.5	47.8	43.4	45.1
witness	39.4	39.0	38.7	40.6
all	39.5	43.3	39.4	40.3

Since we have at disposal two different sets of acoustic models (IBN and IPPS) we can also perform combination of the corresponding ASR outputs. It is known from the literature [10, 11, 12] that combining the outputs of individual systems can significantly improve the overall WER, assuming that the outputs themselves have comparable WERs and provide, at the same time, complementary word hypotheses (i.e. they contain different errors). To this purpose we decided to use ROVER [13], taking into account both mutual agreement between word hypotheses and related confidence scores. Confidence scores (i.e. word posterior probabilities) were estimated, using a method similar to the one proposed in [14], from word graphs generated in the second decoding pass.

Output combination using ROVER was then compared with the performance provided by the system that uses HMMs trained on the combined audio training database "IBN+IPPS" (see section 3.4).

Table 10 gives the results achieved on the whole test set (IT-Apr09+IT-Nov09) with: IBN models (same of Table 8), IPPS models (same of Table 9), ROVER combination between IBN and IPPS and IBN+IPPS models.

As can be seen, the effects of combination with ROVER is negligible while the union of both IPPS and IBN training sets does not result into any performance improvements with respect to use the IBN training set alone. This trend of results is not observed on other application domains where, in general, ROVER combination and the addition of training data is somehow beneficial. Although a deeper analysis of the ASR errors is necessary, a possible explanation of the inefficiency of ROVER is that the separate systems entering the combination provide large numbers of similar errors and, hence, little complementary information is provided.

5.2. Polish Test Data

The evaluation of the recognition performance of the Polish system was performed on the data sets described in Table 3. Table 11 shows the performance of the system before the inclusion of in-domain acoustic data. The Table shows the WER of the first pass recognition, as well as of the second pass CMLLR SAT recognition, with

and without MLLR adaptation.

Table 11. *Results with out-of-domain AM – WER [%].*

System	PL-Dev08	PL-Eval08	PL-Dev09
1st Pass	46.3	68.1	82.1
+ SAT	36.5	58.3	61.7
+ MLLR	35.2	56.2	59.8

In Table 12, the effect of the vocabulary size on the (first pass) recognition performance is presented. The Table also includes the OOV rate for each vocabulary size. We see that although the OOV rate continues to improve for each vocabulary increase, the error rate does not significantly improve over a vocabulary size of 300k words.

Table 12. *Effect of vocabulary size [%]*

System	PL-Dev08		PL-Eval08		PL-Dev09	
	OOV	WER	OOV	WER	OOV	WER
75k	8.16	49.2	10.1	69.4	7.70	83.4
150k	5.27	46.9	7.54	68.4	3.14	81.8
300k	3.86	46.4	6.05	68.1	1.99	82.0
600k	1.54	46.3	2.55	68.1	0.59	82.1

Table 13 shows the performance of the final system. Here the acoustic model MAP adapted to the judicial domain, as described in Section 4.2, is used. We see that the inclusion of the in-domain acoustic data give a sizable improvement over the out-of-domain case. The final error rates of the system are in the range of 30% to 50%.

Table 13. *Final system, in-domain adapted AM – WER [%].*

System	PL-Dev08	PL-Eval08	PL-Dev09
1st Pass	37.7	57.0	79.4
+ SAT	33.4	48.9	56.4
+ MLLR	32.7	47.4	52.3

6. CONCLUSIONS AND FUTURE WORK

In this paper, automatic speech transcription in the judicial domain has been investigated by targeting two languages, Italian and Polish, for which acoustic data have been acquired in the courtrooms of Naples and Wroclaw. Several factors make this transcription tasks difficult, such as: distant talk microphone, cross channel effects, overlapped speech, spontaneous and accented speech, speech under stress and background noise. In addition, the scarcity of in-domain data, both textual and acoustic, makes the adaptation of existing transcription system very challenging.

The initial results achieved by exploiting a small amount of in-domain data for system adaptation, that is around 40% WER for the Italian language and varying between 30% and 50% WER for Polish, show that the transcription systems need to be customized for the specific task in order to ensures adequate performance.

For Italian, current work is mostly devoted to exploit in-domain lightly supervised audio data with the aim of improving acoustic models.

To cope with the high OOV rates for Polish, future activities on the Polish system will include morphological decomposition and hybrid language modeling [15].

7. REFERENCES

[1] M. Cettolo, "Porting an Audio Partitioner Across Domains," in *Proc. of ICASSP*, Orlando, May 2002, pp. I–301–304.

[2] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[4] G. Stemmer and F. Brugnara, "Integration of Heteroscedastic Linear Discriminant Analysis (HLDA) into Adaptive Training," in *Proc. of ICASSP*, Toulouse, France, May 2006, pp. I–1185–1188.

[5] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization.," *Computer Speech and Language*, vol. 20, no. 1, pp. 107–123, Jan. 2006.

[6] M. Bacchiani and B. Roark, "Unsupervised Language Model Adaptation," in *Proc. of ICASSP*, Hong-Kong, 2003, pp. 224–227.

[7] Jonas Lööf, Christian Gollan, and Hermann Ney, "Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a polish speech recognition system," in *Interspeech*, Brighton, U.K., Sept. 2009, pp. 88–91.

[8] J. L. Gauvain and C.-H. Lee, "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models," in *Proc. of the DARPA Speech and Natural Language Workshop*, Palo Alto, Feb. 1991, pp. 272–277.

[9] G. Stemmer, F. Brugnara, and D. Giuliani, "Using Simple Target Models for Adaptive Training," in *Proc. of ICASSP*, Philadelphia, PA, March 2005, vol. 1, pp. 997–1000.

[10] C. Breslin and M.J.F. Gales, "Directed decision trees for generating complementary systems speech communication," *Speech Communication*, vol. 51, no. 3, pp. 284–295, 2009.

[11] B. Hoffmeister, D. Hillard, S. Hahn, R. Schlüter, M. Ostendorf, and H. Ney, "Cross-site and Intra-Site ASR System Combination: Comparisons on Lattice and 1-Best Methods," in *Proc. of ICASSP*, Honolulu, Hawai, April 2007, pp. IV–1145–1148.

[12] D. Falavigna, M. Gerosa, R. Gretter, and D. Giuliani, "Phone-To-Word Decoding Through Statistical Machine Translation and Complementary System Combination," in *Proc. of the ASRU Workshop*, Merano, Italy, Dec. 2009, pp. 519–524.

[13] J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)," in *Proc. of ASRU*, Santa Barbara, CA, 1997, pp. 347–352.

[14] G. Evermann and P. .C. Woodland, "Large Vocabulary Decoding and Confidence Estimation Using Word Posterior Probabilities," in *Proc. of ICASSP*, Istanbul, Turkey, June 2000, pp. 2366–2369.

[15] A. El-Desoky Mousa, R. Schlüter, and H. Ney, "A Hybrid Morphologically Decomposed Factored Language Models for Arabic LVCSR," in *Proc. of the NAACL/HLT*, Los Angeles, California, USA, June 2010, pp. 701–704.