SLIGHTLY SUPERVISED ADAPTATION OF ACOUSTIC MODELS ON CAPTIONED BBC WEATHER FORECASTS

Christian Mohr, Christian Saam, Kevin Kilgour, Jonas Gehring, Sebastian Stüker, Alex Waibel

International Center for Advanced Communication Technologies (interACT)
Institute for Anthropomatics
Karlsruhe Institute of Technology, Karlsruhe, Germany
{firstname.lastname}@kit.edu

Abstract

In this paper we investigate the exploitation of loosely transcribed audio data, in the form of captions for weather forecast recordings, in order to adapt acoustic models for automatically transcribing these kinds of forecasts. We focus on dealing with inaccurate time stamps in the captions and the fact that they often deviate from the exact spoken word sequence in the forecasts. Furthermore, different adaptation algorithms are compared when incrementally increasing the amount of adaptation material, for example, by recording new forecasts on a daily basis.

Index Terms: speech recognition, acoustic model adaptation, slightly supervised training, loose transcripts, adaptation methods

1. Introduction

Within the European Union's 7th Framework Programme's project (Bridges Across the Language Divide) (EU-BRIDGE) [1] several tasks on automatic speech recognition are defined over different data sets. The active domains are TED talks[2], a collection of public talks covering a variety of topics, academic lectures and weather bulletins. For the TED task large collections of training data are readily available which are the basis for the IWSLT ASR evaluation track [1]. The mismatch between training and testing data pertains to speaker and domain yet style is relatively consistent. The approximate transcripts of the talks are very close to verbatim. For lectures there is comparatively little training data available. Thus, general models are adapted on small data sets that often do not even have transcripts. Unsupervised adaptation must account for mismatches in speaker, domain and style. The weather bulletin data on the other hand is a new and still very small data set that has weak references in the form of captions. Again, general models must be adapted in a supervised/semi-supervised manner to account for mismatches in style, domain and speakers.

This paper investigates different approaches for acoustic model adaptation on weather forecasts when captions are available. Of special interest is the question of how to deal with imperfect transcripts and unlabeled non-speech audio as investigated by [2]. Similar to [3] we investigate the possible improvements of a system by unsupervised acoustic model training depending on the amount of training data and the reliability of transcripts. Similar to [4, 5], we made use of word level confidence scores. However, we did not exclude data from training based on the word posteriors of the transcription, as we have too little training data available as that we could afford to lose some of it. Our training conditions can be compared to [6] where new data for retraining comes from the same speaker, channel and related conversation topics. Following the implications of [7] we add low confidence score data to the training, but unlike in other work we apply word-based weighting in order to compensate for errors, as it was done by [8] for acoustic model adaptation. The assumption is that erroneous data is helpful to improve system generalization. Unlike other work, e.g. [9], we did not use a lattice-based approach. Furthermore we study the choice of a good adaptation method with increasing adaption set sizes. We assume that sufficient amounts of training data are available in order to transition from transform based techniques, such as maximum likelihood linear regression and its feature space constrained version [10], to maximum likelihood [11] or maximum a posteriori parameter re-estimation [12].

[1] http://www.eu-bridge.eu
[2] http://www.ted.com

2. The BBC Weather Data

The BBC weather data consists of audio recordings of British weather forecasts and manually generated captions. There are two different kinds of forecasts: general bulletins and regional forecasts. The captions for the general bulletins are prerecorded and therefore more accurate than the live captions for the regional weather forecasts.

The data used consists of audio of forecasts recorded between 2008 and 2012 with roughly 50 different speakers. This information is only an estimate since the tagging of speaker names is partly imprecise and inconsistent, and the airing date of the shows is not always given.

Although the speakers are well trained there are some hesitations, grammar errors or lengthy formulations in the recordings which are corrected in the captions (some examples are shown in Table 1). The captions therefore can only be regarded as loose transcripts.

Capt.	*We had some more typical summer weather*
Verb.	*We had some more **of this** typical summer weather*
Capt.	*Downpours across England and Wales*
Verb.	*Downpours **whistling** across England and Wales*

Table 1: *Two examples for differences between captions (Capt.) and the verbatim word sequences (Verb.). Words omitted in the caption are bold-faced.*

Captions are only provided for the forecast itself with time markers relative to the beginning of the forecast but without ab-

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

solute positions in the recording. Also, the recordings often contain untranscribed parts at the beginning—such as trailers and introductions by different speakers—and advertisements at the end. The length of the untranscribed parts in the audio differs, so it is not possible to simply cut it off at a specific time, in order to just obtain the portion of the data that is actually captioned.

For the test corpus described in Section 5 careful transcriptions were available in addition to the captions also covering only the forecast itself, leaving out introduction and trailers. To determine the general degree of faithfulness of the captions as a (training) reference we computed the word error rate (WER) between the verbatim references and the captions. Table 2 shows the result of this on the test data. It can be seen that the captions and the verbatim transcriptions are rather close, indicating that the speakers are indeed well trained.

	Case Sensitive	Case Insensitive
WER	7.4%	5%
# words in reference	12007	12007
Total # errors	890	600
# Substitutions	434	144
# Insertions	21	21
# Deletions	435	435

Table 2: *WER between the captions and the verbatim transcripts of the test set, and statistics on the types of errors.*

The captions' data format contains timestamps that indicate when individual captions are displayed, however these need not exactly correspond to when the respective words were spoken, because captions have to adhere to further constraints in addition to when they were spoken. E.g., they have to adhere to a certain letter rate in order to be readable, have to maintain a certain distance from scene changes and may not span several scenes. The timing information is therefore too inaccurate to be taken as timestamps in the audio.

3. Preprocessing: Finding Suitable Start and End Times

Due to the inaccuracy of the timing information we need to align the captions to the audio to be able to use them as loose transcripts. A naïve Viterbi alignment of the concatenated captions to the corresponding audio file leads to suboptimal results due to the large untranscribed parts in the audio.

To make sure that we use only audio that is properly transcribed we decode the audio data, align the resulting hypotheses to the captions and search for the first and last matching trigram. The start time of the first word of the first trigram is used as the start time of the loose transcript and the end time of last the word of the last trigram as end time. The words preceding the first trigram and following the last trigram are deleted from the transcript. This leads to some data loss but the start and end times can be iteratively refined by repeating the decoding and cutting after the model was adapted on the data obtained in the previous iteration.

Even after one iteration of model re-estimation the amount of data that is lost due to the cut-off is rather small. We tested the approach on a subset of 16 hours of audio data (parts 1-4 of the final database as described in Section 4). The acoustic model as well as the language model of the system used for decoding were trained on British general broadcast data. The baseline system is described in more detail in Section 5. On the test set described in Section 5 this system's WER was 31.9%. The baseline system was adapted on the raw recordings and then achieved a WER of 23.2%. This adapted system in turn was used to refine the start and end times of the audio it was adapted on. After cutting, the amount of audio data was reduced by approximately 37% but the text of the original captions only by around 6%. When applying this method to the final database, the reduction of audio data decreased to 35.1% and the percentage of removed words in the reference to 4.9%. So the cut off audio data consists of a small part of transcribed data plus a very large part of unwanted data.

The different results for the subset and the final database result from the small amount of data in total and from the fact that the length of introductions and trailers differs significantly. Although this heuristic for finding usable start and end times is rather simple, it is convenient for the given task, as only 4.9% of words in the reference were lost.

4. Experimental Set-Up and Data

All experiments were performed with the *Janus Recognition Toolkit* (JRTk) developed at Karlsruhe Institute of Technology and Carnegie Mellon University [13].

The training of our *Hidden Markov Model* (HMM) based acoustic model tries to maximize the likelihood of the model on the training data. In *Viterbi* training only the most probable HMM state sequence is computed and used for re-estimating the HMM's parameters. In *Expectation Maximization* (EM) training all possible alignments are taken into consideration for model estimation. Both training techniques work iteratively and require an initial set of model weights which are improved over several iterations of model re-estimation. Adaptation can be done by performing one iteration of model parameter estimation on new adaptation data using an existing set of models that was trained on different, out-of-domain data. As an alternative *maximum-a-posteriori* (MAP) estimation using the models of an existing speech recognizer as seed models for the ML estimation of the model parameters was investigated—again on the adaptation data. Various weighting factors τ to control the influence of the seed model were evaluated. We denote the MAP weights as *(Weight of the seed model* $\cdot 100$ | *Weight of the adaptation data* $\cdot 100$*)*.

From past experience these approaches are known to outperform *maximum likelihood linear regression* (MLLR) adaptation of acoustic models when the training data exceeds roughly 1.5hrs. The amount of available adaptation data suggested MLLR adaptation to be inferior, thus it was omitted.

Since the captions are not verbatim transcripts we expected the Viterbi as well as the EM training to suffer from transcription errors. The EM algorithm should not be affected as badly as the Viterbi approach, since all possible HMM state sequences are considered and not only the most likely one. To overcome the problem of transcription errors we tried altering the transcripts by introducing successions of filler states between words, that are intended to be aligned to feature vectors from words missing from the transcript. As a final alternative we tested two kinds of unsupervised adaptation on transcripts of the adaptation data that are in fact hypotheses produced by the unadapted speech recognition system. The statistics accumulated in training over these transcripts are either weighted by the confidence value of the respective hypothesis word or the weights are set to 1.0 for all words.

We split up the data and adapted the general system de-

scribed in Section 5 with the different algorithms on different growing subsets of the database. Periodically new packages of data were made available. Our final database consists of the 6 parts described in Table 3.

Part Number	# files	Comment	Duration / hours (net duration)
1	50	bulletins part 1	3.87 (2.43)
2	50	bulletins part 2	4.04 (2.48)
3	50	bulletins part 3	3.89 (2.44)
4	51	bulletins part 4	3.88 (2.49)
5	103	bulletins part 5	7.46 (4.86)
6	54	regional forecasts	1.07 (1.00)
Σ			24.21 (15.7)

Table 3: *Overview of the parts the final database consists of. The size of the general bulletins files varies between 180 and 410 seconds.*

Part 6 (regional forecasts with live captions) contains captions considered to be less verbatim even than the material in the rest of the database. These captions are produced on the fly during live airings and the results depend on the ability of the captioner to keep up with the speaking rate of the presenter.

Since not all parts of the data were available when the experiments began, we tested the general viability of some adaptation approaches only on initially available subsets of the final training data. We tested only the most promising techniques on the larger databases.

5. Results

For all tests a semi-continuous system was used as baseline system to be adapted on the given adaptation data.

As front-end we used *mel-frequency cepstral coefficients* (MFCC) with 13 cepstral coefficients. The mean and variance of the cepstral coefficients were normalized on a per-utterance basis. 15 adjacent frames were combined into one single feature vector. The resulting feature vectors were then reduced to 42 dimensions using *linear discriminant analysis* (LDA).

The acoustic model is a context dependent quinphone system with three states per phoneme, and a left-to-right topology without skip states. It uses 24,000 distributions over 8,000 codebooks. The model was trained using *incremental splitting of Gaussians* (MAS) training, followed by *semi-tied covariance* (STC) [14] training using one global transformation matrix, and one iteration of Viterbi training. The acoustic models have up to 128 mixture components per model and a total of $591k$ Gaussian components. All models use *vocal tract length normalization* (VTLN)[15].

The system was trained on about 200 hours of carefully transcriped British general Broadcast data.

A baseline 4gram case sensitive language model with modified Kneser-Ney smoothing was built for 36 sources with a total word count of 2,935.6 million and a lexicon size of $128k$ words. This was done using the SRI Language Modeling Toolkit [16]. The language models built from the text sources were interpolated using interpolation weights estimated on a tuning set resulting in a language model with $59,293k$ 2grams, $153,979k$ 3grams and $344,073k$ 4grams. For decoding, a pronunciation dictionary was used containing $142k$ entries.

A second, smaller 4-gram language model was trained on the references of the acoustic model training data containing $61,738$ words increasing the lexicon size to $129k$ words. This was interpolated with the baseline language module to produce an adapted language model. Adding pronunciations and variants for the new words in the lexicon to the pronunciation dictionary increased its size to $144k$ entries.

5.1. Test Set

The test set contains 54 minutes of general weather bulletins, the captions for which were manually corrected to be verbatim transcripts. Correct start and end times were also manually determined.

5.2. First Adaptation Tests

First adaptation tests were done on a subset of the final database originally containing 16 hours of audio data and 10.6 hours after recalculation of start and end times as described in Section 3. Of all 6 parts of the final database the first tests were only done on the first 4. Table 4 shows a comparison of the results of the adaptations via one iteration of the Viterbi or the EM algorithm, and the Viterbi-based MAP estimation. Viterbi re-estimation using the original start and end times was used as an additional baseline.

To limit time and memory consumption a segmentation of the audio files using a partial Viterbi-Alignment was performed instead of aligning over whole audio files.

System	WER
Baseline	31.9%
Viterbi 1 iteration	20.9%
Viterbi 2 iterations	26.5%
EM 1 iteration	21.5%
EM 2 iterations	32.1%
MAP 20/80	20.7%
MAP 40/60	20.5%
MAP 60/40	21.0%
MAP 80/20	21.6%

Table 4: *First adaptation results on a subset of the final database.*

It can be seen that the EM re-estimation achieves worse results than the Viterbi re-estimation. These results however are not comparable since our EM training fails for a considerable amount of the training data (approximately 31%). This may be due to the implementation being optimized under the assumption of accurate transcripts and although a pruning technique is applied the EM training exceeds the memory limit for long utterances. Tuning the pruning parameter of the EM algorithm might alleviate this problem.

After two iterations of Viterbi re-estimation the systems performance degrades since the adaptation over-fits to the adaptation data.

5.3. Results on the Iteratively Growing Database

Viterbi estimation and Viterbi MAP estimation were tested in multiple configurations trained on different parts of the database. Results are shown in Table 5 and Figure 1.

It can be seen that Viterbi MAP adaptation outperforms the Viterbi ML re-estimation for all sizes of the database but the difference in performance decreases the larger the amount of training data is. Figure 2 shows the corresponding results of the tests using the adapted language model. Here the difference in

database parts	Viterbi WER	MAP 20/80	MAP 40/60	MAP 60/40	MAP 80/20
1	26.1%	25.1%	23.4%	**22.9**%	24.0%
1+2	23.4%	23.4%	22.6%	**22.0**%	22.8%
1-3	21.9%	21.6%	**21.2**%	21.5%	22.0%
1-4	20.8%	**20.3**%	20.6%	20.7%	21.6%
1-5	20.4%	**20.1**%	20.3%	20.6%	21.3%
1-6	20.1%	**19.8**%	20.0%	20.5%	21.3%
only 6	50.9%	33.7%	31.4%	**30.7**%	31.0%

Table 5: *WERs of adapted systems for different numbers of parts of the final database. The best performance for each size of the database is bold.*

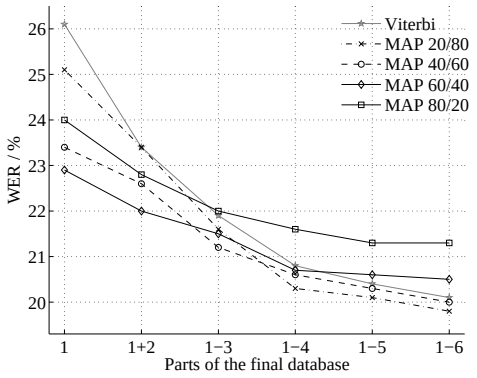

Figure 1: *Word Error rates for different adaptation methods on the test set, plotted over increasing amounts of available adaptation data* .

performance for larger amounts of training data is significantly higher and the performance of the Viterbi ML re-estimation seems to stagnate. Using the adapted language model with the unadapted acoustic model, the resulting WER is 21.5%.

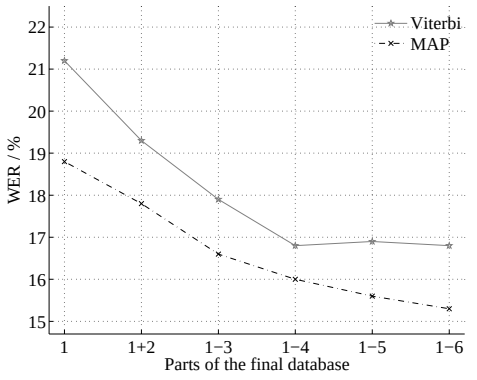

Figure 2: *Word Error rates for different adaptation methods on the test set, plotted over increasing amounts of available adaptation data with adapted language model.*

We took part in an internal EU-BRIDGE evaluation campaign on the Weather Bulletin Task using the presented Viterbi MAP re-estimation method. The initial system training mentioned in Section 5 was redone with the adaptation data also being used during the basic system training. Instead of MFCC features we used *deep bottle neck features* (DBNFs) [17] which have been shown to significantly outperform MFCC features. We also performed fMLLR and MLLR adaptation in a second decoding pass. This resulted in a single 2nd pass system with a WER of 12.4%. Adapting this system, which already saw the Weather Bulletin data during traing, still resulted in a reduced WER of 12.0% for the Viterbi ML re-estimation and 11.9% for the MAP re-estimation.

5.4. Comparison to Unsupervised Training

Table 6 compares the best results from using the captions as training transcriptions with training in an unsupervised manner. One can see that the unsupervised training performs significantly worse. This suggests that the quality of the training references, while not verbatim grade, is still good enough and that they are much more informative than recognition hypotheses. However, when time is not an issue repetitive unsupervised adaptation may yield similar results. The Viterbi ML re-estimation using the original start and end times mentioned in Section 3 was redone using the final database, improving the performance from 23.2% to 21.8%.

system	WER
Unadapted system	31,9%
Viterbi on original start and end times	21.8%
Viterbi on modified start and end times	20.1%
MAP 20/80 on modified start and end times	19.8%
Unsupervised	28.4%
Unsupervised weighted	27.9%

Table 6: *WER of the best adapted system to baseline experiments. Unsupervised adaptations are Viterbi ML re-estimations on the hypotheses from the decoding with the baseline system. In weighted unsupervised training the confidence of a word is used as a weight for the training patterns during the accumulation of the sufficient statistics during training.*

6. Conclusion

We investigated methods for using captions as loose transcripts for adapting acoustic models for automatic speech recognition to weather forecast audio data. Considerable gains can be made by determining the correct start and end times of the captions. This is necessary since the original time segments of the captions only match imprecisely to the corresponding parts in the audio. It turned out that similar to supervised adaptation methods Viterbi ML estimation is outperformed by MAP estimation but for increasing amounts of adaptation material results converge. By using an adapted language model the effect of convergence is decreased.

We showed that the proposed method leads to a WER that is 8.1% abs. lower than when using unsupervised adaptation methods, letting the WER drop from 27.9% to 19.8%. Refining start and end times for incomplete transcriptions by a simple heuristic that searches for matching trigrams of words in the alignment of hypotheses from the decoded audio files to the transcriptions improves the WER by 1.7% abs.

Using the proposed method in combination with language model adaptation and deep BNF features led to a WER of 11.9% in the EU-BRIDGE evaluation campaign on the Weather Bulletin task.

At a level of 5% WER divergence of the available transcripts from verbatim references supervised training is still much more effective than replacing the reference with automatically generated transcripts. A major drawback of the proposed method is the need to decode all of the adaptation material. Depending on the task this might not be feasible due to the time intensity of the approach.

If the divergence is higher, the investigation of the appropriate adaption method would have to be redone and data selection methods might become necessary.

7. Acknowledgements

The work leading to these results has received funding from the European Union under grant agreement no. 287658. '*Research Group 3-01*' received financial support by the '*Concept for the Future*' of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative.

8. References

[1] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the iwslt 2012 evaluation campaign," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2012*, Hong Kong, December 6-7 2012.

[2] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, Jan. 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S088523080190186X

[3] ——, "Unsupervised acoustic model training," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, May 2002, pp. I–877–I–880.

[4] C. Gollan, S. Hahn, R. Schlüter, and H. Ney, "An improved method for unsupervised training of LVCSR systems," *Interspeech, Antwerp, Belgium*, pp. 2101–2104, 2007. [Online]. Available: http://www-i6.informatik.rwth-aachen.de/publications/download/366/An%20Improved%20Method%20for%20Unsupervised%20Training%20of%20LVCSR%20Systems.pdf

[5] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer using TV broadcasts," in *Proc. of ICSLP*, vol. 98, 1998, pp. 2207–2210. [Online]. Available: http://reference.kfupm.edu.sa/content/u/n/unsupervised_training_of_a_speech_recogn_110317.pdf

[6] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve performance," in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 301–305. [Online]. Available: http://reference.kfupm.edu.sa/content/u/t/utilizing_untranscribed_training_data_to_14022.pdf

[7] H. Li, T. Zhang, and L. Ma, "Confirmation based self-learning algorithm in LVCSR's semi-supervised incremental learning," *Procedia Engineering*, vol. 29, no. 0, pp. 754–759, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S187770581200046X

[8] C. Gollan and M. Bacchiani, "Confidence scores for acoustic model adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, Apr. 2008, pp. 4289–4292.

[9] T. Fraga-Silva, J.-L. Gauvain, and L. Lamel, "Lattice-based unsupervised acoustic model training," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 4656–4659.

[10] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, Apr. 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230898900432

[11] S. Stüker, "Acoustic modelling for under-resourced languages," Ph.D. dissertation, PhD thesis, Universität Karlsruhe (TH), Karlsruhe, Germany, 2009. 125, 2009.

[12] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[13] H. Soltau, F. Metze, C. Fuegen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," in *ASRU*, 2001.

[14] M. Gales, "Semi-tied covariance matrices for hidden markov models," Cambridge University, Engineering Department, Tech. Rep., February 1998.

[15] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *ICASSP*, Munich, Germany, April 1997.

[16] A. Stolcke, "Srilm—an extensible language modeling toolkit," in *ICSLP*, 2002.

[17] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stackedauto-encoders," in *Acoustics, Speech and Signal Processing (ICASSP),2013 IEEE International Conference on*. IEEE, 2013.