THE RWTH ENGLISH LECTURE RECOGNITION SYSTEM

Simon Wiesler[1], *Kazuki Irie*[2,*], *Zoltán Tüske*[1], *Ralf Schlüter*[1], *Hermann Ney*[1,2]

[1]Human Language Technology and Pattern Recognition,
Computer Science Department, RWTH Aachen University, Aachen, Germany
[2]École Centrale Paris, Paris, France
[3]LIMSI CNRS, Spoken Language Processing Group, Paris, France

ABSTRACT

In this paper, we describe the RWTH speech recognition system for English lectures developed within the Translectures project.

A difficulty in the development of an English lectures recognition system, is the high ratio of non-native speakers. We address this problem by using very effective deep bottleneck features trained on multilingual data. The acoustic model is trained on large amounts of data from different domains and with different dialects. Large improvements are obtained from unsupervised acoustic adaptation.

Another challenge is the frequent use of technical terms and the wide range of topics. In our recognition system, slides, which are attached to most lectures, are used for improving lexical coverage and language model adaptation.

Index Terms— lecture recognition, speech recognition system, LVCSR

1. INTRODUCTION

Video lectures currently receive a lot of attention. Renowned universities have made lectures available in electronic form, for example on Coursera [1] or edX [2], accompanied by additional material and interaction methods. In addition, many conferences including ICASSP record talks and make them available to a wide audience.

The automatic transcription of these videos is of high interest. Transcriptions allow to perform text search in videos and facilitate access of non-native speakers and people with hearing disabilities. In this work, we describe our English lecture recognition system, which has been developed within the Translectures project [3]. The project aims at transcribing the Videolectures.NET [4] archive, which is a very large online repository for academic videos. Most of the talks are in English language and were given at computer science conferences as ICML, NIPS, and others. This means, the system described in this paper is applied to a large-scale task with real-life data.

The lecture recognition task has several characteristics, which impose challenges for automatic speech recognition (ASR). One problem is the very large vocabulary, in particular the frequent use of rare technical terms. In addition, the video lectures cover a wide range of topics. Another challenge that is specific for English lectures is the large ratio of non-native speakers. Typically, performance of ASR systems already degrades when dealing with different dialects which are not covered by the acoustic training data. The effect of foreign speakers is even more severe. We address this problem by using training data with different dialects and even different languages for the training of bottleneck features. Finally, since the goal is to transcribe the complete Videolectures.NET database, our recognition system must be efficient.

On the other hand, the task offers specific opportunities for improving ASR performance. First, lectures typically only have one speaker and unsupervised speaker adaptation is therefore very effective. Second, many video lectures are attached with slides, which can be used as an additional knowledge source for extending the vocabulary and language model adaptation [5]. Furthermore, many video lectures have been subtitled manually by volunteers. The subtitles are not an exact transcription as known from standard ASR tasks. Still, the data can be used as a basis for acoustic model training [6].

The topic of this paper is therefore a challenging real-life task. The task has several properties different from conventional ASR research tasks, which makes it interesting to study how well existing methods perform on it. Further, it is required to adapt techniques to the task under consideration.

The next section briefly describes the Videolectures.NET database. The remaining sections give an overview of our lecture recognition system, including highly effective deep multilingual bottleneck features and our proposed language model adaptation approach.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287755 (transLectures). H. Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.

*The author performed the work while at RWTH Aachen University.

Table 1. Statistics of the Videolectures.NET database.

	Archive	Development	Test
Videos	9,148	4	4
Time	5,900h	3.2h	3.4h
Sentences	-	1K	1.3K
Words	-	28K	34K
Vocabulary	-	3K	3K

2. THE VIDEOLECTURES.NET DATABASE

Videolectures.NET is a free and open access video lectures repository. Within the Translectures project, all 9,148 video lectures that have been viewed at least fifty times at project start, are automatically transcribed.

A very small part of the database has been carefully transcribed for the development of ASR systems. The statistics of the development and test sets as well as the complete archive are given in Table 1. Until now, additional twenty hours of training data have been transcribed. Due to this small amount, we have not used this data so far.

Most of the lectures are attached with slides. Depending on the format, the slide text can be extracted directly, or has to be generated with optical character recognition (OCR). The text extraction of the slides has already been performed within the Translectures project. We assume the slide texts to be given and do not deal with the difficulties of OCR ourselves. The quality of the extracted slide texts is quite low in general. Even if the text can be extracted without OCR, mathematical formulas and other graphical elements impose difficulties.

3. ACOUSTIC MODEL

Our acoustic model is based on the tandem approach [7]. This allows using well-understood Gaussian mixture model (GMM) speaker adaptation techniques, which is of crucial importance for the task under consideration.

3.1. Resources

The acoustic model has been trained on large amounts of acoustic data from various resources with different domains and dialects, see Table 2. The Quaero English corpus contains podcasts, mostly in British English, which has been provided within the Quaero project [8]. HUB4 and TDT4 are widely used broadcast news corpora in American English. EPPS is a collection of carefully transcribed European Parliament speeches.

The TED data has been collected by ourselves. We downloaded 200 hours of subtitled videos from the TED website [9]. The subtitles are not well aligned to the audio data and contain additional annotation, for example speaker information. Repetitions and disfluencies are also not annotated in subtitles. Therefore, we applied a text postprocessing to the subtitles and removed audio segments which could not be aligned when using a low alignment pruning threshold.

Table 2. Statistics of acoustic training data. BC stands for broadcast conversations, BN for broadcast news, PS for parliament speeches, and L for lectures.

Corpus	Domain	Duration (h)	#Words
Quaero English	BC	268	1.6M
HUB4	BN	206	1.6M
TDT4	BN	186	1.7M
EPPS	PS	102	0.7M
TED	L	200	1.8M
Quaero French	BC	317	3.9M
Quaero German	BC	142	1.4M
Quaero Polish	BC	110	1.0M

In total, 962 hours of English audio data have been used for acoustic model training. Furthermore, we used data from other languages for training multilingual neural network features, see Subsection 3.3. The multilingual bottleneck features have been trained on all Quaero corpora shown in Table 2, in total 837 hours.

3.2. Baseline feature extraction

Sixteen Mel-cepstral coefficients (MFCC) are extracted every 10 ms using a bank of 20 filters. In addition, a voicedness feature is computed. By applying a sliding window of size 9, 154-dimensional features are obtained, which are mapped by a linear discriminant analysis (LDA) to a 45-dimensional subspace.

3.3. Multilingual deep bottleneck features

In addition to the cepstral features, bottleneck (BN) features extracted from a multilayer perceptron (MLP) are used. The neural network has already been trained within the Quaero project [10]. This illustrates a major practical advantage of the tandem approach: The bottleneck features can be shared across different tasks. In our case, the network is trained on multilingual data and the features can even be shared across languages. This approach facilitates system development strongly. In addition, we know from our experience on Quaero data that the multilingual training improves performance by about three percent relative in comparison to using only data from the target language [10]. The observed improvements from multilingual training are in the same range as reported by other groups for data-rich languages [11],[12].

The bottleneck features are extracted from a neural network with a hierarchical structure as described in [13, 14], based on MRASTA filtering [15]. The fast modulation part of the MRASTA filtering is fed to a first MLP. A second MLP is trained on the slow modulation components and the PCA transformed BN output of the first MLP. The modulation fea-

tures fed to the MLPs are always augmented by the critical band energies. Both MLPs have six hidden layers with 2000 hidden nodes and a 60-dimensional BN layer, which is placed before the last hidden layer. The final features are obtained by applying a PCA to the BN activations of the second MLP and reducing the dimensionality to 38.

We applied the multilingual training method proposed by [16]. The MLP training data comprises the English, French, German, and Polish acoustic training data from the Quaero project - in total 837 hours. The feature vectors extracted from the joint corpus of the four languages were randomized and fed to the MLPs. Using language specific softmax outputs, backpropagation has been initiated only from the language specific subset of the output depending on the language-ID of the feature vector. The MLPs were trained according to cross-entropy criterion with 1500 tied-triphone states per language as outputs [17].

3.4. Training

The acoustic model used in this work is a GMM/HMM. The features for the GMM are obtained by concatenating the spectral features with the BN features described in the previous subsection.

The acoustic model has been trained on all English audio data given in Table 2. The parameters have been optimized according to the maximum likelihood (ML) criterion with the expectation maximization algorithm (EM) with Viterbi approximation and a splitting procedure. The GMM has a globally pooled and diagonal covariance matrix and roughly 1.2M densities. It models $4,500$ generalized triphones determined by a decision-tree-based clustering (CART).

Speaker adaptation is of crucial importance for the performance of a lecture recognition system, because there is a lot of data per speaker available. We use several speaker adaptation techniques in our system. First, mean and variance normalization is applied to the spectral features on segment level. Furthermore, vocal tract length normalization (VTLN) is applied to the MFCC features. The VTLN warping factors are obtained from a Gaussian classifier (fast-VTLN) [18]. In addition, we perform speaker adaptation with constrained maximum likelihood linear regression (CMLLR) [19]. The CMLLR transformation has been applied to the training data and a new speaker-adapted GMM has been trained (speaker adaptive training). In recognition, the CMLLR transformations are estimated from a first recognition pass and then, a second recognition pass with the GMM from speaker adaptive training (SAT) is performed.

In addition to these transformations on feature side, Gaussian mixture models can also be adapted directly with maximum likelihood linear regression (MLLR) [19]. In our experience, MLLR usually does not improve performance of tandem systems, see for example [20]. For the lecture recognition system, we found MLLR to be beneficial due to the large amount of adaptation data per speaker.

Table 3. Statistics of text resources. The third column gives the weight of the data source in the interpolated LM.

Corpus	#Words	λ
Acoustic transcriptions	8M	0.27
Slide texts	97M	0.25
IWSLT 2013	3M	0.13
Quaero blog 2011	730M	0.12
WIT TED talks	$3M$	0.11
Gigaword	3B	0.08
WMT 2012 news-crawl	2.8B	0.04

4. LEXICON

Our pronunciation modeling is based on the British English Example Pronunciation (BEEP) dictionary [21]. Missing pronunciations are determined by a grapheme-to-phoneme conversion (g2p) model. We use a combination of a generative g2p model and a conditional random field based approach as described in [22].

The baseline recognition vocabulary has been determined by using the 150k most frequent words from the English Gigaword corpus, see Table 3. Many technical terms which are used in lectures are not covered by the Gigaword corpus. In order to reduce the OOV rate, we added the 50k most frequent unknown words from the slide texts to the recognition lexicon. This adds a lot of noise to the lexicon, but the OOV rate is reduced strongly, because the slides exactly correspond to the data that is recognized.

5. LANGUAGE MODEL

5.1. Resources

The datasets for language model training are summarized in Table 3. Only small amounts of in-domain data are available: the transcriptions of the acoustic training data, the collection of Videolectures.NET slide texts, lecture data provided by the IWSLT 2013 evaluation campaign [23], and a small collection of TED talks. A large archive of blog data has been provided within the Quaero project. Gigaword and news-crawl are very large news text collections. In total, the language model is trained on 6.6 billion running words.

5.2. Training

The texts were normalized and converted to lower case. Punctuations were discarded. For every data source, a single 4-gram LM with modified Kneser-Ney smoothing has been estimated using the SRILM toolkit [24]. These LMs were linearly interpolated by optimizing the perplexity (PPL) on a validation set. Table 3 shows the interpolation factor of each data source. It can be seen that the in-domain datasets have much more weight than the large out-of-domain sources, despite of

Table 4. Word error rates (in %) for the MFCC baseline system and the tandem system on the development and test set.

	Baseline		Tandem	
	Dev.	Test	Dev.	Test
speaker-independent	45.9	36.6	35.4	27.1
+CMLLR	41.0	30.6	31.3	23.6
+MLLR	37.8	28.2	29.4	21.8
+Slides/Viterbi	36.1	27.4	28.5	21.5
+Slides/CN	35.0	26.6	28.1	21.2

their small size. The full language model (11GB) is only applied in lattice rescoring. For recognition, the language model is pruned to 726MB with entropy-pruning.

5.3. Adaptation using slide texts

The language model adaptation is performed in a lattice rescoring framework. For each video lecture, a language model is trained on the corresponding slides. We found a careful text normalization to be very important for the noisy slide texts. In rescoring, the unadapted language model described above is dynamically interpolated with the slide language model. The interpolation weight has been chosen as the average of the weights that minimize the perplexity for each lecture in the development data. We trained N-gram models with order one to four on the corresponding slide texts. Bigrams performed about ten percent relatively better than unigrams in terms of perplexity, and marginally better than trigrams and 4-grams.

6. RECOGNITION SETUP

Our system has a four-pass recognition setup. In an initial unadapted pass, a first transcription is obtained, which is used for the CMLLR-adapted recognition pass. In the subsequent MLLR-adapted recognition pass, word lattices are created. The lattices are rescored using the slide-adapted language model. Finally, a confusion network (CN) decoding is performed on the lattice [25].

In contrast to other systems developed in our group, for example the Quaero evaluation systems [20], we do not use system combination or cross-adaptation, because our aim is to apply our system to a large-scale dataset.

7. EXPERIMENTAL RESULTS

Table 4 shows detailed word error rate (WER) results for an MFCC baseline system and the tandem system described above.

The results highlight the importance of adaptation on this task. The relative improvement of all adaptation techniques on the tandem system is 20.7% on the test data. On the MFCC baseline system, the improvement is even 25.1% relative. Slight additional gains are obtained by using CN decoding instead of Viterbi decoding on the final lattices.

Table 5. Effect of using slide vocabulary on the OOV and WER rate. Perplexity (PPL) improvements by language model (LM) adaptation using slide texts.

		Dev.	Test
OOV [%]	baseline lexicon	2.5	1.0
	baseline + slides lexicon	1.1	0.7
WER [%]	baseline lexicon	28.8	21.1
	baseline + slides lexicon	28.1	21.2
PPL	baseline LM	174	143
	slide-adapted LM	146	140

The multilingual bottleneck features strongly improve system performance from 26.6% WER to 21.2% WER on the final system. The relative improvement is even higher if less adaptation techniques are applied. Overall, the bottleneck features are clearly highly valuable.

Using the subtitled videos as acoustic training data only gave a moderate improvement of 1.8% relative in WER.

The system benefits from the availability of slide texts in several ways. First, the OOV rate is reduced strongly, in particular on the development data, see Table 5. It can be assumed that most of the relevant technical terms also appear on the slides attached to the videos. The reduction of the OOV rate also reduces the word error rate on the development data. Second, the slides already improve the unadapted language model. They are used as one of the LM text sources and have a high interpolation weight, see Table 3. Finally, the lecture-specific slides are used for LM adaptation. The slide adaptation gives improvements in perplexity (see Table 5) and in the recognition result, but varies with the quality of the slides.

8. CONCLUSION

In this work, we described the RWTH speech recognition system for English video lectures in detail. The system described in this paper is applied to a large-scale dataset with real-life data. The lecture recognition task is challenging due to the large vocabulary and the high ratio of non-native speakers. Adaptation of the acoustic model and the language model plays an important role on this task. The tandem acoustic model has been adapted within a multipass CMLLR and MLLR framework. We obtained additional gains by using a language model adaptation method similar to [5] based on the slides which are attached to most videos.

We only obtained small improvements by using subtitled videos as acoustic training data. In future work, we plan to use this data more efficiently by making use of unsupervised training techniques as described in [6].

9. REFERENCES

[1] "Coursera," www.coursera.org.

[2] "edx," www.edx.org.

[3] "Translectures," www.translectures.eu.

[4] "Videolectures.NET," www.videolectures.net.

[5] A. Martinez-Villaronga, M. A. del Agua, J. Andrés-Ferrer, and A. Juan, "Language model adaptation for video lectures transcription," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, Canada, May 2013, pp. 8450–8454.

[6] A. Rousseau, F. Bougares, P. Deléglise, H. Schwenk, and Y. Estève, "LIUM's systems for the IWSLT 2011 speech translation tasks," in *Proc. of IWSLT*, San Francisco, USA, Dec. 2011.

[7] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Dallas, USA, Mar. 2000, pp. 1635–1638.

[8] "Quaero," http://www.quaero.org.

[9] "Ted," http://www.ted.org.

[10] Z. Tüske, R. Schlüter, and H. Ney, "Multilingual hierarchical MRASTA features for ASR," in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 2222–2226.

[11] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, Canada, May 2013, pp. 8619–8623.

[12] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, Canada, May 2013, pp. 7304–7308.

[13] F. Valente and H. Hermansky, "Hierarchical and parallel processing of modulation spectrum for ASR applications," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Las Vegas, USA, Mar. 2008, pp. 4165–4168.

[14] C. Plahl, R. Schlüter, and H. Ney, "Hierarchical Bottle Neck Features for LVCSR," in *Proc. Interspeech*, Makuhari, Japan, Sep. 2010, pp. 1197–1200.

[15] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 361–364.

[16] S. Scanzio, P. Laface, L. Fissore, R. Gemello, and F. Mana, "On the Use of a Multilingual Neural Network Front-End," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, pp. 2711–2714.

[17] Z. Tüske, R. Schlüter, and H. Ney, "Deep hierarchical bottleneck MRASTA features for LVCSR," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, Canada, May 2013, pp. 6970–6974.

[18] L. Welling, S. Kanthak, and H. Ney, "Improved methods for vocal tract normalization," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Phoenix, USA, Mar. 1999, pp. 761–764.

[19] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, no. 2, p. 171, 1995.

[20] M. Sundermeyer, M. Nußbaum-Thom, S. Wiesler, C. Plahl, A. El-Desoky Mousa, S. Hahn, D. Nolden, R. Schlüter, and H. Ney, "The RWTH 2010 Quaero ASR Evaluation System for English, French, and German," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011, pp. 2212–2215.

[21] T. Robinson, "BEEP - The British English Example Pronunciation Dictionary," 1995, ftp://svr-ftp.eng.cam.ac.uk/comp.speech/dictionaries/.

[22] S. Hahn, P. Lehnen, S. Wiesler, R. Schlüter, and H. Ney, "Improving LVCSR with Hidden Conditional Random Fields for Grapheme-to-Phoneme Conversion," in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 495–499.

[23] "IWSLT 2013," http://www.iwslt2013.org.

[24] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. Int. Conf. on Spoken Language Processing*, Denver, USA, Sep. 2002, pp. 901 – 904.

[25] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. NIST Speech Transcription Workshop*. Baltimore, USA, 2000.