The KIT-NAIST (Contrastive) English ASR System for IWSLT 2012

Michael Heck[1], *Keigo Kubo*[2], *Matthias Sperber*[1], *Sakriani Sakti*[2], *Sebastian Stüker*[1], *Christian Saam*[1]
Kevin Kilgour[1], *Christian Mohr*[1], *Graham Neubig*[2], *Tomoki Toda*[2], *Satoshi Nakamura*[2], *Alex Waibel*[1]

[1]Institute for Anthropomatics, Karlsruhe Institute of Technology, Germany
[2]Augmented Human Communication Laboratory, Nara Institute of Science and Technology, Japan

{michael.heck,matthias.sperber}@student.kit.edu
{sebastian.stueker,christian.saam,kevin.kilgour,christian.mohr,alex.waibel}@kit.edu
{keigo-k,ssakti,neubig,tomoki,s-nakamura}@is.naist.jp

Abstract

This paper describes the KIT-NAIST (Contrastive) English speech recognition system for the IWSLT 2012 Evaluation Campaign. In particular, we participated in the ASR track of the IWSLT TED task. The system was developed by Karlsruhe Institute of Technology (KIT) and Nara Institute of Science and Technology (NAIST) teams in collaboration within the interACT project. We employ single system decoding with fully continuous and semi-continuous models, as well as a three-stage, multipass system combination framework built with the Janus Recognition Toolkit. On the IWSLT 2010 test set our single system introduced in this work achieves a WER of 17.6%, and our final combination achieves a WER of 14.4%.

1. Introduction

Similar to the IWSLT 2011 Evaluation Campaign [1], IWSLT 2012 featured an Automatic Speech Recognition (ASR) track whose task it was to recognize the recordings made available by TED on their website[1][2]. The TED talks collection is a web repository of recordings of public speeches/talks of about 5-25 minutes by people from various fields of expertise covering repetitive topics related to technology, entertainment and design (TED). This paper describes the ASR (contrastive) system developed for this campaign by the KIT-NAIST team in collaboration under the interACT project. Detail descriptions of the KIT-NAIST primary submission which was a system combination between the KIT primary submission and this contrastive submission can be found in [3].

The main challenge of this ASR track is to develop a system that is capable of recognizing spontaneous and open-domain speeches. Here, we employ: (1) acoustic models trained on European Parliament Plenary Sessions (EPPS) recordings [4] and additional publicly available transcribed TED audio data crawled from the web; (2) 4-gram language models that were trained by interpolating TED data with other provided corpora, as well as a topic adaptated LM using latent Dirichlet allocation (LDA); (3) a pronunciation dictionary in which the pronunciations of unknown words were constructed using several grapheme-to-phoneme methods; (4) single system decoding with fully continuous and semi-continuous models, as well as a three-stage, multipass system combination framework.

The rest of this paper is structured as follows. Section 2 summarizes data resources used for the experiments, and Section 3 provides a description of acoustic front-ends used in our system. An overview of the techniques and data used to build our acoustic models is given in Section 4. We describe the language model used for this evaluation in Section 5 and pronunciation lexicon in Section 6. Our decoding strategy and experimental results are explained in Section 7. Finally, the conclusion is drawn in Section 8.

2. Data Resources

2.1. Training Corpora

For acoustic model training, the following speech corpora were used:

- 80 hours of manually transcribed English European Parliament Plenary Session (EPPS) speeches, provided by RWTH Aachen within the TC-STAR project [4].
- 157 hours of TED talks released before the cut-off date of 31 December 2010, downloaded from the TED websites with the corresponding subtitles.

For language model training, the following text corpora provided by the IWSLT organizer were used:

- 2M words of TED transcripts.
- The English portion of the English-French training data from the Sixth Workshop on Statistical Machine Translation (WMT 2011), including News Commentary (NC), EuroParl (EPPS), NEWS, and GIGA data.

2.2. Test Corpora

Table 1 describes both test sets (“tst2011” and “tst2012”) used for this year’s evaluation campaign, as well as our de-

[1] http://www.ted.com/talks

velopment set for system development and parameter optimization ("tst2010"). "tst2010" is a data set which was also used as development set for last year's ASR task. "tst2011" comprises of TED talks newer than December 2010, is the test set for the IWSLT 2011 ASR task and serves as progress test set to measure the improvement in systems from 2011 to 2012. "tst2012" is a collection of some of the most recent recordings made available by TED. All sets were used with the original pre-segmentation provided by the IWSLT organizers.

Set	#talks	#utt	dur	dur/utt
tst2010	11	1664	2.5h	5.4s
tst2011	8	818	1.1h	4.9s
tst2012	11	1124	1.7h	5.6s

Table 1: Statistics of the development set ("tst2010") and the test sets ("tst2011" and "tst2012"), including the total number of talks (*#talks*), the total number of utterances (*#utt*), the overall speech duration (*dur*), and average speech duration per utterance (*dur/utt*).

3. Front-end

We trained the system with a front-end based on the widely used mel-frequency cepstral coefficients (MFCC). The front-end provides features every 10ms. During decoding this was changed to 8ms after the first stage, so that in ROVER hypotheses from first and second pass can be combined. This is done because it may be beneficial for various sounds to have a higher frame rate, while for some other that may not be the case. Therefore a hypotheses combination from different frame rates may lead to better results. During training and decoding, the features were obtained by a discrete Fourier transform followed by a Mel-filterbank. Vocal tract length normalization (VTLN) is done in the linear domain [5]. The MFCC front-end uses 13 cepstral coefficients. Mean and variance are normalized on a per-utterance basis. Finally, to incorporate the temporal structures and dependencies, 15 adjacent (center, 7 left, and 7 right) frames are stacked into one single feature vector leading to 195 dimensional super vector (15x13 dimensions). It then reduced to an optimum 42 dimensions by applying a linear discriminant analysis.

4. Acoustic Modeling

4.1. Data Preprocessing

Segmenting the TED data into sentence-like chunks used for building a training set was performed with the help of a decoding pass on the input data in order to discriminate speech and non-speech regions and doing a forced alignment given the subtitles. Beforehand, the relevant speech part of each downloaded video soundtrack was cut with the time stamps given by the subtitle files. The segmentation was done by splitting at non-speech regions of notable length. In order to compensate for occasional inaccuracies of the computed time stamps, we merged successive segments by the simple heuristic, "As long as the transcription of the subsequent segment does not start with an uppercase letter, add it to the current segment." This resulted in a sentence-like segmentation of the TED data. While the manually transcribed EPPS data has predefined speaker labels and therefore does not need to be clustered, we made the simple assumption for the TED data, that each talk is spoken by exactly one speaker. Table 2 lists the details of the resulting utterances.

Data	#talks	#utt	dur	dur/utt
EPPS	1,894	52,464	80h	5.5s
TED	711	105,692	157h	5.3s

Table 2: Statistics of speech data for acoustic model training, including the total number of talks (*#talks*), the total number of utterances (*#utt*), the overall speech duration (*dur*), and average speech duration per utterance (*dur/utt*).

4.2. AM Training

All models are context-dependent quinphones with a standard three-state left-to-right HMM topology without skip states. The models use 24,000 distributions over 8,000 codebooks. First, a fully continuous system using 2,000 distributions and codebooks was trained by using incremental splitting of Gaussians training (MAS) [6], followed by optimal feature space training (OFS) which is a variant of semi-tied covariance (STC) [7] training using one global transformation matrix. After generating new labels for the training data, a system using 8,000 distributions and codebooks was trained in the same way, and further refined by 2 iterations of Viterbi training. The semi-continuous system was trained after clustering the models resulting in 24,000 distributions over 8,000 codebooks with 2 iterations of Viterbi training.

5. Vocabulary and Language Model

5.1. Data Preprocessing

We normalized the training data sources of TED, NEWS, NC, EPPS, and GIGA, in a case-insensitive fashion. Noisy parts were omitted from the GIGA corpus, using rules to detect, e.g., HTML tags and very short sentences. Table 3 shows the resulting text corpora along with their total size (word count) and vocabulary size.

Data	Size	Vocabulary
TED	2.4m	43k
EPPS	52m	79k
NC	4.5m	50k
NEWS	2,300m	986k
GIGA	576m	501k

Table 3: Total size (word count) and vocabulary size of the individual text corpora.

5.2. Vocabulary

For the vocabulary selection, we followed an approach proposed by Venkataraman et al. [8]. We built unigram language models using Witten-Bell smoothing [9] from all text sources except GIGA, and determined unigram probabilities that maximized the likelihood of a held-out TED data set. We then defined the 150k most probable words as the vocabulary.

5.3. LM Training

Using the SRILM toolkit [10], we built 4-gram language models with modified Kneser-Ney smoothing [11] from each of the text corpora. These were then combined using linear interpolation as follows:

$$P(w|h) = \lambda_1 P_1(w|h) + \lambda_2 P_2(w|h) + \cdots + \lambda_k P_k(w|h). \quad (1)$$

The interpolation weights $\lambda_1, \ldots, \lambda_k$ were chosen to maximize the likelihood of a held-out TED data set. The resulting language model contains 43 million bigrams, 190 million trigrams, and 382 million 4-grams. The effect of the different training corpora on the language model perplexity is summarized in Table 4.

Data	Perplexity
TED only	184.03
+ EPPS, NC	167.84
+ NEWS	133.51
+ GIGA	133.16

Table 4: Language model perplexities on tst2010 for different amounts of training data.

5.4. Topic Adaptation

During development, we further applied topic adaptation using LDA (see [12]). Using the given document structure of the TED corpus, we inferred 50 topics, using a vocabulary of 10k words. We estimated a separate 4-gram language model for each topic by using all sentences in the TED training data that had at least one word assigned to this topic. This strategy allows assigning a sentence to several topics, as opposed to much of the previous work that enforces a hard assignment decision for each training unit (e.g. see [13]). For the actual decoding of a specific talk, all words from the first-pass hypothesis that have a confidence value higher than a certain threshold are used to estimate the current topic distribution. The top 10 topics (a limitation imposed by SRILM) are linearly interpolated with weights according to that distribution. Finally, this adapted language model is interpolated with the background language model described above. The confidence threshold and the weight for the interpolation of adapted and background language models were chosen to optimize perplexity on a development data set. Topic model adaptation reduced the perplexity on the talks in the development set ("tst2010") by 0.9% on average. The effect in overall system performance is discussed in Section 7.1.

6. Pronunciation Lexicon

6.1. Phoneme Set

We employ the same phoneme set used by KIT with 45 phonemes, and utilize the existing pronunciation dictionary: (1) the CMU Pronouncing Dictionary [14]; a machine-readable pronunciation dictionary for North American English that contains over 125,000 words and their transcriptions based on 39 phonemes; (2) the EPPS dictionary with KIT phoneme set. Since both pronunciation dictionaries use different phoneme sets, our first step is to convert the 39-phonemes of the CMU dictionary into the KIT phoneme set. This is done using the Sequitur grapheme-to-phoneme (G2P) tool based on joint n-gram models [15]. All words that were covered by both the CMU dictionary and the EPPS dictionary were used as phoneme-to-phoneme training data. Then, by utilizing the trained phoneme-to-phoneme model, the pronunciation of words included in CMU dictionary but not included in EPPS dictionary were converted into new pronunciations based on the KIT phoneme set. Finally, we obtained 135k words of the CMU dictionary with the KIT phoneme set (45 phonemes) as baseline dictionary.

6.2. G2P Conversion

Next, we explored various G2P conversion techniques for handling pronunciations of words that have not been covered by the baseline CMU dictionary (135k words, 45 phonemes). These include: (1) Sequitur G2P based on joint n-gram models (denoted as *Sequitur*); (2) DirecTL+ based on online discriminative training [16, 17] (denoted as *DirecTL+*); and (3) merging 1-best of *Sequitur* and *DirecTL+* results (denoted as *Merge*(1)+(2)).

To find the optimum G2P technique, we employed the baseline CMU dictionary (135k words, 45 phonemes) with a 10% test set, a 5% development set, and the remaining data as training set. Table 5 summarizes the results in terms of Recall, Precision, F-value.

	Recall	Precision	F-measure
(1) *Sequitur*	55.19	55.16	55.17
(2) *DirecTL+*	55.61	55.61	55.61
Merge(1)+(2)	63.23	49.80	55.71

Table 5: Recall, Precision and F-measure for various G2P conversion techniques on the baseline CMU dictionary (135k words, 45 phonemes).

Note that, the *Merge*(1)+(2) G2P may result in one or two pronunciations per word, while other techniques only result in one pronunciation per word. In our experiments the DirecTL+ obtains 55.61% in terms of F-value and Sequitur is 55.17%. These results are lower than those of previous research [15, 16, 17] because we employ a more complex phoneme set than the CMU phoneme set and did not delete heteronyms, which are words that share the same written form but have different pronunciations and meanings. Finally, the optimum *DirecTL+* G2P conversion is selected for dictionary construction.

6.3. Dictionary Construction

Last, we constructed a dictionary that would be used for open domain TED talks. Here, we retrain the selected *DirecTL+* G2P conversion using the baseline CMU dictionary (135k words, 45 phonemes) with a 5% development set, and the remaining data as training set. Then, for all words that are included in the LM, but have not been covered by the baseline CMU dictionary (except the capitalized words), the pronunciations were constructed based on *DirecTL+* G2P conversion. For capitalized words, the pronunciations were converted based on rule in which each alphabet included in the word is converted to the alphabetical sound. The number of the converted words was 65k words in the defined 150k vocabulary (see Section 5.2).

7. Decoding Strategy and Results

During development, we evaluated our system using the IWSLT 2010 test set for the lecture task, which was explicitly declared held out data during model training due to the fact that both the IWSLT 2010 development set and test set were initially included in the downloaded raw TED talks intended for training. For comparison we also evaluated the performance on the test2011 set released by the IWSLT organizers.

All speech recognition experiments, i.e. the decoding—as well as acoustic model training—were performed with the Janus Recognition Toolkit (JRTk) that includes the IBIS single pass decoder, developed at Karlsruhe Institute of Technology and Carnegie Mellon University [18]. During development, we evaluated our system mainly using the IWSLT 2010 test set for the lecture task, which was explicitly declared held out data during model training due to the fact that both the IWSLT 2010 development set and test set were initially included in the downloaded raw TED talks intended for training. We observed the recognition accuracy in terms of word error rate (WER) after first pass decoding.

7.1. Single System

Table 6 shows the results given various configurations of the fully continuous system after MAS, OFS and Viterbi training, and the performance of the semi-continuously trained system after two iterations of Viterbi training. For comparison we also evaluated the performance on the test set ("tst2011").

The "tst2010" set was further used for tuning the system and determining the best language model size and dictionary size for decoding data that is very close to the target domain. The IBIS decoder used by JRTk scores the hypothesis related to an input utterance [18] as follows:

$$score(W|X) = logP(X|W) + logP(W) \cdot lz + lp \cdot |W| \quad (2)$$

The lz parameter defines the language model weight, i.e. determines the impact of the language model on the decoding process relative to the acoustic model. The parameter lp is a word transition penalty, helping to normalize the sequence lengths of words W. Note that applying topic model adaptation LM on our development systems improved the WER by up to 2.2% relative. However, results using the final system were mixed, and the adaptation scheme was not included in the final submission.

Data	System		tst2010	tst2011
EPPS	FCHMMs	MAS	36.5%	31.6%
+TED	FCHMMs	MAS	18.8%	16.5%
		OFS	18.8%	16.0%
		VIT1	18.1%	15.9%
		VIT2	18.2%	16.1%
	SCHMMs	VIT1	17.7%	15.6%
		VIT2	17.6%	15.5%

Table 6: Performance of the single system on the development set ("tst2010") and test set ("tst2011") in WER. The fully continuous system uses 8000 codebooks and distributions, the semi-continuous system 24000 distributions.

7.2. System Combination

The decoding strategy for the final submission is based on the principle of system combination and cross-system adaptation. The underlying assumption of system combination is that different systems commit different errors which may cancel each other out. Cross-system-adaptation profits from the fact that the unsupervised acoustic model adaptation methods work better when applied on hypotheses generated by multiple systems that perform about equally well [19].

Our framework for system combination consists of three stages. In the first stage multiple systems, including our system described in this paper, are run. The additional systems differ in the applied front-ends and acoustic models (see [3]) in a way that achieves a high system diversity among the full set of applied systems. The same combination of dictionary and language model is used for all decoding runs. The system outputs of the first stage are combined via confusion network combination (CNC) [20]. The acoustic models of all systems for the second pass are then adapted on this output using VTLN, maximum likelihood linear regression (MLLR) [21] and feature space constrained MLLR (fMLLR) [22]. After the first stage, the frame shift was changed to 8 ms. In the second stage a second CNC is performed. The third and final stage of our system combination framework is a ROVER combination of seven second pass outputs and both CNC outputs [23]: A majority vote among all CNC results and second stage system outputs gave the best results.

The segmentation of the test data was used as is. For simplicity reasons no extra speaker clustering was performed, assuming one speaker per test recording. Table 7 shows the performance of the system combination on the development set ("tst2010") in WER, and Table 8 shows the summary of the final system combination results on various development and test sets in WER. The results shown on test set ("tst2011" and "tst2012") are based on IWSLT 2012 evaluation feedback.

System	WER
Single 1st pass	17.6%
CNC 1st pass (CNC$_1$)	17.1%
Single 2nd pass	16.1%
CNC 2nd pass (CNC$_2$)	14.5%
ROVER (CNC$_1$ + CNC$_2$ + 7 $*$ 2nd pass)	14.4%

Table 7: Comparison of the single system performance and the system combination results on the development set (“tst2010”) in WER.

Test set	WER
test2010	14.4%
test2011	12.3%
test2012	12.6%

Table 8: Summary of final system performance performed with ROVER (CNC$_1$ + CNC$_2$ + 7 $*$ 2nd pass). The results shown on test set (“tst2011” and “tst2012”) are based on IWSLT 2012 evaluation feedback.

8. Conclusion

In this paper we described our English speech-to-text system with which we participated in the IWSLT 2012 TED task evaluation on the ASR track. Besides utilizing already existing systems by adjusting them to the new domain, we trained a completely new system by including annotated audio data extracted from TED talks into acoustic model training. Furthermore, we built a dictionary and trained a language model specific to the TED task of this year’s evaluation campaign. Our final system utilizes a three-stage, multipass system combination framework. On the IWSLT 2010 test set our single system introduced in this work achieves a WER of 17.6%, and our final combination achieves a 14.4% WER.

9. Acknowledgements

This work was supported in part by an interACT student exchange scholarship. The research leading to these results has in part received funding from the European Union Seventh Framework Programme (FP7/2007-2013) 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.

10. References

[1] M. Federico, L. Bentivogli, M. Paul, and S. Stüker, “Overview of the IWSLT 2011 evaluation campaign,” in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2011*, San Francisco, CA, USA, December 8-9 2011.

[2] ——, “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.

[3] C. Saam, C. Mohr, K. Kilgour, M. Heck, M. Sperber, K. Kubo, S. Stüker, S. Sakti, G. Neubig, T. Toda, S. Nakamura, and A. Waibel, “The 2012 KIT and KIT-NAIST english ASR systems for the IWSLT evaluation,” in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2012*, Hong Kong, December 6-7 2012.

[4] C. Gollan, M. Bisani, S. Kanthak, R. Schluter, and H. Ney, “Cross domain automatic transcription on the TC-STAR EPPS corpus,” in *Proc. of ICASSP*, Philadelphia, USA, 2005, pp. 825–828.

[5] P. Zhan and M. Westphal, “Speaker normalization based on frequency warping,” in *Proc. of ICASSP*, Munich, Germany, 1997, pp. 1039–1042.

[6] T. Kaukoranta, P. Fränti, and O. Nevalainen, “Iterative split-and-merge algorithm for VQ codebook generation,” *Optical Engineering*, vol. 37, no. 10, pp. 2726–2732, 1998.

[7] M. Gales, “Semi-tied covariance matrices for hidden markov models,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[8] A. Venkataraman and W. Wang, “Techniques for effective vocabulary selection,” in *Proc. of EUROSPEECH*, Geneva, Switzerland, 2003, pp. 245–248.

[9] I. Witten and T. Bell, “The zero-frequency problem: estimating the probabilities of novelevents in adaptive text compression,” *IEEE Transactions on Information Theory*, vol. 37, no. 4, pp. 1085–1094, 1991.

[10] A. Stolcke, “SRILM – an extensible language modeling toolkit,” in *Proc. of ICSLP*, Denver, USA, 2002, pp. 901–904.

[11] R. Kneser and H. Ney, “Improved backing-off for m-gram language modeling,” in *Proc. ICASSP*, 1995, pp. 181–184.

[12] X.-H. Phan and C.-T. Nguyen, “GibbsLDA++: A C/C++ implementation of latent dirichlet allocation (LDA),” http://jgibblda.sourceforge.net/, 2007.

[13] F. Liu and Y. Liu, “Unsupervised language model adaptation incorporating named entity information,” in *Proc. of ACL*, Prague, Czech Republic, 2007, pp. 672–679.

[14] “The carnegie mellon university pronouncing dictionary,” http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[15] M. Bisani and H. Ney, “Joint-sequence models for grapheme-to-phoneme conversion,” *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[16] S. Jiampojamarn and G. Kondrak, “Online discriminative training for grapheme-to-phoneme conversion,” in *Proc. INTERSPEECH*, Beijing, China, 2009, pp. 1303–1306.

[17] C. C. S. Jiampojamarn and G. Kondrak, “Integrating joint n-gram features into a discriminative training framework,” in *Proc. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, Beijing, China, 2010, pp. 697–700.

[18] C. F. H. Soltau, F. Metze and A. Waibel, “A one-pass decoder based on polymorphic linguistic context assignment,” in *Proc. of ASRU*, Madonna di Campiglio, Italy, 2001.

[19] S. Stüker, C. Fügen, S. Burger, and M. Wölfel, “Cross-system adaptation and combination for continuous speech recognition: The influence of phoneme set and acoustic front-end,” in *Proc. of INTERSPEECH*, Pittsburg, PA, USA, 2006, pp. 521–524.

[20] L. Mangu, E. Brill, and A. Stolcke, “Finding consensus in speech recognition: Word error minimization and other applications of confusion networks,” *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[21] 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, no. 2, pp. 171–185, 1995.

[22] M. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[23] J. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER),” in *Proc. of ASRU*, Santa Barbara, USA, 1997, pp. 347–354.