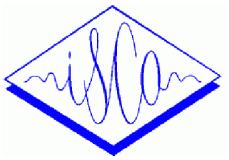

International Workshop on
Spoken Language Translation 2012
Hong Kong, December 6-7, 2012

The 2012 KIT and KIT-NAIST English ASR Systems for the IWSLT Evaluation

Christian Saam[1], Christian Mohr[1], Kevin Kilgour[1], Michael Heck[1], Matthias Sperber[1], Keigo Kubo[2], Sebastian Stüker[1], Sakriani Sakti[2], Graham Neubig[2], Tomoki Toda[2], Satoshi Nakamura[2] and 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,kevin.kilgour,christian.mohr,christian.saam,alex.waibel}@kit.edu
{keigo-k,ssakti,neubig,tomoki,s-nakamura}@is.naist.jp

Abstract

This paper describes our English *Speech-to-Text* (STT) systems for the 2012 IWSLT TED ASR track evaluation. The systems consist of 10 subsystems that are combinations of different front-ends, e.g. MVDR based and MFCC based ones, and two different phone sets. The outputs of the subsystems are combined via confusion network combination. Decoding is done in two stages, where the systems of the second stage are adapted in an unsupervised manner on the combination of the first stage outputs using VTLN, MLLR, and cMLLR.

Index Terms: speech recognition, IWSLT, TED talks, evaluation system, system development

1. Introduction

The *International Workshop on Spoken Language Translation* (IWSLT) offers a comprehensive evaluation campaign on spoken language translation. One part of the campaign focuses on the translation of TED Talks[1], short 5-25min presentations by people from various fields related in some way to Technology, Entertainment, and Design (TED) [1]. In order to evaluate different aspects of this task IWSLT organizes several evaluation tracks on this data covering the aspects of automatic speech recognition (ASR), machine translation (MT), and the full-fledged combination of the two of them into speech translation systems.

The goal of the TED ASR track is the automatic transcription of TED lectures on a given segmentation, in order to interface with the machine translation components in the speech-translation track. The quality of the resulting transcriptions are measured in word error rate (WER).

In this paper we describe our English ASR systems with which we participated in the TED ASR track of the 2012 IWSLT evaluation campaign. This year, our system is a further development of our last year's evaluation system [2] and makes use of system combination and cross-adaptation, by utilising acoustic models which are trained with different acoustic front-ends and employ two different phoneme sets. In addition to last year, we also included TED talks available via TED's website by training on them in a slightly supervised manner.

We submitted two primary systems. One was solely developed by KIT, the other one was developed in cooperation with NAIST in Japan. A description of the additional work done by NAIST on the KIT-NAIST (contrastive) submission can be found in [3].

On the 2011 evaluations set, which serves as a progress test set, we were able to reduce the word error rate of our transcription systems from 17.1% to 12.0%, a relative reduction of 29.8%. On the 2012 evaluation set, the KIT-NAIST primary system reached a WER of 12.4%.

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

Text corpus	Word Count	sources
IWSLT training data transcripts	3 million	2
News (+news commentary)	2114 million	4
Parallel Giga Corpus	523 million	1
LDC English Gigaword 4	1800 million	6
UN + Europarl documents	376 million	1
Google Books Ngrams (subset)	1000 million ngrams	1
total	4816 million	15

Table 1: *Language Model training data word count per corpus after cleaning and data selection and number of text sources included in corpus. The total word count does not include the Google Books Ngrams.*

The rest of this paper is structured as follows. Section 2 describes the data that our system was trained on. This is followed by Section 3 which provides a description of the two acoustic front-ends used in our system. An overview of the techniques used to build our acoustic models is given in Section 4. We describe the language model used for this evaluation in Section 5 and our decoding strategy and results are presented in Section 6.

2. Training Data

For acoustic model training we used the following data sources:

- 237 hours of Quaero training data from 2010 to 2012.
- 157 hours of data downloaded from the TED talks website, including the subtitles provided by the TED conferences archive

For the language model and vocabulary selection we used the subtitles of the TED talks and text data from various sources (see Table 1) totalling about 4816 million words.

3. Front-Ends

We trained systems for two different kinds of acoustic front-ends. One is based on the widely used *mel-frequency cepstral coefficients* (MFCC) obtained from a discrete Fourier transform and the other on the *warped minimum variance distortionless response* (MVDR). The second front-end replaces the Fourier transformation by a warped MVDR spectral envelope [4], which is a time domain

technique to estimate an all-pole model using a warped short time frequency axis such as the mel-scale. The use of the MVDR eliminates the overemphasis of harmonic peaks typically seen in medium and high pitched voiced speech when spectral estimation is based on linear prediction.

For training, both front-ends provided features every 10 ms. During decoding this was changed to 8 ms after the first stage. The altered frame-shift introduces a slight variation in the decoding results which can be exploited in the ROVER stage of the decoding process.

For the MVDR front-end we used a model order of 22 without any filter bank since the warped MVDR already provides the properties of the mel-scale filter bank, namely warping to the mel-frequency and smoothing. The advantage of this approach over the use of a higher model order and a linear filter bank for dimensionality reduction is an increase in resolution in low frequency regions which cannot be attained with traditionally used mel-scale filter banks. Furthermore, with the MVDR we apply an unequal modelling of spectral peaks and valleys that improves noise robustness, due to the fact that noise is mainly present in low energy regions.

Both front-ends apply *vocal tract length normalization* (VTLN) [5]. For MFCC this is done in the linear domain, for MVDR in the warped frequency domain. The MFCC front-end uses 13 or 20 cepstral coefficients, the MVDR front-end uses 15. The mean and variance of the cepstral coefficients were normalized on a per-utterance basis. For both front-ends 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). Through the temporal context present in the stacked super-vectors the LDA can implicitly perform an approximation of dynamic spectral features. The dimensionality of the final feature vectors was empirically proven to work well and coincides with the dimensionality of a 14 dimensional static feature vector augmented with first and second order dynamic features.

In recent years neural network based features have been shown to improve ASR systems [6]. A typical setup involves training a neural network to recognize phones (or phone-states) from a window of ordinary (e.g. MFCC) feature vectors. With the help a hidden bottleneck layer the trained network can be used to project the input features onto a feature vector with an arbitrarily chosen dimensionality [7]. The input vector is derived from a 15 frame context window with each frame containing 20 MFCC or MVDR coefficients. So far, we used LDA to reduce the dimensionality of this input vector, which limits the resulting LDA-features to linear combinations of the input features. A *multi layer perceptron* (MLP) with the bottleneck in the 2nd hidden layer can make use of non-linear information.

For our IWSLT systems we used bottleneck features for both our MVDR and MFCC front ends.

4. Acoustic Modeling

4.1. Data Preprocessing

For the TED data only subtitles were available so the data had to be split into sentence-like chunks. Therefore the data was decoded to discriminate speech and non-speech and a forced alignment given the subtitles was done where only the relevant speech parts detected by the decoding were used. All this preprocessing was done at NAIST.

4.2. AM Training

We used a context dependent quinphone setup with three states per phoneme, and a left-to-right topology without skip states. All acoustic models initially used 8,000 distributions and codebooks derived from decision-tree based clustering of the states of all possible quinphones. The models were trained using *incremental splitting of Gaussians* (MAS) training, followed by *optimal feature space* training and 2 iterations of Viterbi training. All models use *vocal tract length normalization* (VTLN). After training the continuous density tied state models we further split the state clusters to arrive at 24000 distributions over the 8000 codebooks again based on a decision-tree. Then we trained these semi-continuous models with two iterations of Viterbi training. For some systems the semi-continuous models were worse than the fully-continuous ones, so for the final decoding we used the ones that achieved lower WER on the development data.

We used two different phoneme sets. The first one is based on the CMU dictionary [2] and is the same phoneme set as the one used in last year's system. It consists of 45 phonemes and allophones. The second phoneme set is derived from the BEEP dictionary [3] and contains 52 phonemes and allophones. For the CMU phoneme set we generated missing pronunciations with the help of FESTIVAL [8], while for the beep dictionary we used Sequitur [9] for this. Both grapheme to phoneme converters were trained on subsets of the respective dictionaries.

In total we trained 9 different acoustic models, combining different front-ends and different phoneme sets, which were combined for decoding as described in 6. We found that not all possible combinations need to be trained. The improvements of adding models with new combinations of techniques already used in other systems in different combinations is very small especially when the number of single systems is large.

5. Language Modeling

A 4gram case sensitive language model with modified Kneser-Ney smoothing was built for each of the text sources listed in Table 1. This was done using the SRI Language Modelling Toolkit [10]. Only half the transcripts of the IWSLT develpoment data were used to build a language model, the other half was used as our tuning set. The aforementioned language models built from the text sources in Table 1 were interpolated using interpolation weights estimated on this tuning set resulting in a 4 GB language model with $56,300k$ 2grams, $330,488$ 3grams and $909,927k$ 4grams. The NAIST language model [3] used in KIT-NAIST primary was built with the same sources and tools but applied more thorough data selection strategies for the LDC Gigaword texts.

5.1. Vocabulary Selection

To select the vocabulary the development data text was randomly split in half. For each of our text sources, except the Gigaword and Google Books ngrams (see Table 1) we built a Witten-Bell smoothed unigram language model using the union of the text sources' vocabulary as the language models' vocabulary (global vocabulary). With the help of the maximum likelihood count estimation method described in [11] we found the best mixture weights for representing the tuning set's vocabulary as a weighted mixture of the sources' word counts thereby giving us a ranking of all the words in global vocabulary by their relevance to the tuning set. The top 130k words were selected as our vocabulary. Unknown pronunciations were automatically generated using the aforementioned grapheme to phones conversion.

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[3] ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz

6. Decoding Strategy and Results

The decoding was performed with the *Janus Recognition Toolkit* (JRTk) developed at Karlsruhe Institute of Technology and Carnegie Mellon University [12]. Our decoding strategy is based on the principle of system combination and cross-system adaptation. System combination works on the principle that different systems commit different errors that cancel each other out. Cross-system adaptation profits from the fact that the unsupervised acoustic model adaptation works better when performed on output that was created with a different system that works approximately equally well [13]. The final step in our system decoding set-up is the ROVER combination of several outputs [14].

We trained 9 different acoustic models as described in section 4 and a language model as described in section 5. An additional acoustic model and an additional language model was trained at NAIST. For the IWLST ASR track 3 different submissions were done, which are described in the following.

6.1. KIT Primary Submission

The decoding strategy of the KIT primary submission is described in Figure 1. The set-up used for our evaluation system consists of two stages. In each stage multiple systems are run, and their output is combined with the help of *confusion network combination* (CNC) [15]. On this output the acoustic models of the next stage are then adapted using *Vocal Tract Length Normalization* (VTLN) [5], *Maximum Likelihood Linear Regression* (MLLR) [16], and *feature space constrained MLLR* (fMLLR) [17]. Finally the ten second pass decodings and the CNC outputs of the first pass results as well as the CNC outputs over the second pass decodings are combined using ROVER.

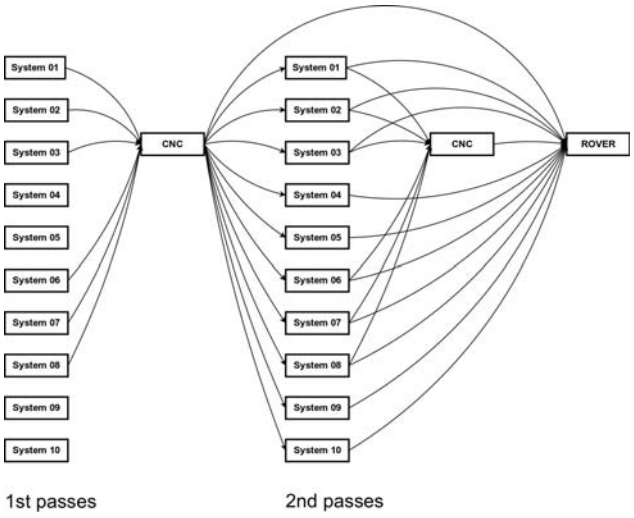

Figure 1: Decoding Strategy of the KIT Primary Submission

6.2. KIT-NAIST Primary and Contrastive Submission

Further to the KIT primary submission we submitted the outputs of two more systems in the IWLST ASR track namely the KIT-NAIST primary and contrastive submissions. Figure 2 shows the principal decoding strategies for all submissions done. The three submissions are depicted as the two rightmost rectangles and the central rectangle.

The KIT-NAIST contrastive submission differs from the KIT primary submission in the fact that a different language model and pronunciation dictionary was used for the decoding which were trained in cooperation with NAIST.

The KIT-NAIST primary submission then is a combination of the KIT primary and the KIT-NAIST contrastive submissions. We combined a subset of outputs of the second passes and CNCs done for both the KIT primary submission and for the KIT-NAIST contrastive submission. In order to let the ROVER combine the most diverse outputs we selected ten second pass systems using the most diverse techniques plus two CNCs. That is the five most diverse of the ten KIT systems and the five most diverse of the ten KIT-NAIST systems respectively, together with the CNC of the KIT-NAIST first pass outputs and the CNC of the KIT second pass outputs. The final system output for the KIT-NAIST primary submission is depicted in Figure 2 by the central rectangle.

System	WER
KIT 2011	17.4%
KIT 2012	12.0%

Table 2: *WER on tst2011 with KIT's system for the evaluation campaign of 2011 compared to the system for the campaign of 2012.*

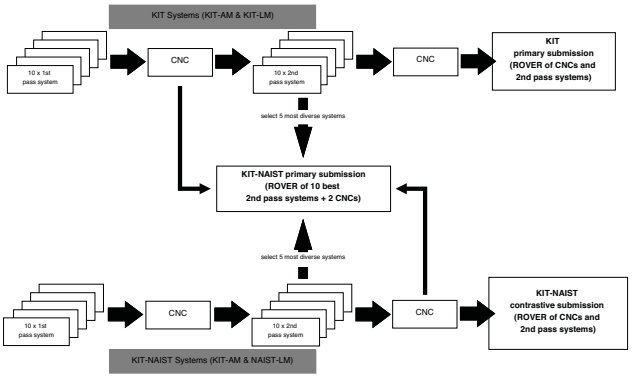

Figure 2: Decoding Strategy of the KIT Primary, KIT-NAIST Primary and Contrastive Submissions

6.3. Results

We evaluated our systems on the IWSLT test sets from 2010 (tst2010), 2011 (tst2011) and 2012 (tst2012). We used the tst2010 set as development set and for parameter optimization. Sets tst2011 and tst2012 were used for this years evaluation campaign, set tst2011 also for last years campaign.

Since the tst2011 set was used for this years and last years evaluation campaign we can indicate our progress over the last year. The compared results are shown in Table 2.

Table 3 shows the results of the KIT primary decoding strategy and its intermediate steps on the development set tst2010.

Table 4 shows the results of all our submissions on all described test sets.

7. Conclusion

In this paper we presented our English LVCSR systems, with which we participated in the 2012 IWSLT evaluation.

System	WER
Single best 1st pass system	17.8%
CNC 1st pass	16.6%
Single best 2nd pass system	15.3%
CNC 2nd pass	14.7%
ROVER	14.3%

Table 3: *WER of the decoding strategy for the KIT primary submission and its intermediate steps on the development set.*

	KIT primary	KIT-NAIST primary	KIT-NAIST contrastive
tst2010	14.3%	14.0%	14.4%
tst2011	12.0%	12.0%	12.3%
tst2012	12.7%	12.4%	12.6%

Table 4: *WER for our three submissions for the three different test sets.*

8. Acknowledgements

This work was supported in part by an interACT student exchange scholarship. '*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. The work leading to these results has received funding from the European Union under grant agreement $n \circ 287658$. This work was partly realized within the Quaero Programme, funded by OSEO, French State agency for innovation.

9. References

[1] M. Federico, 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] S. Stüker, K. Kilgour, C. Saam, and A. Waibel, "The 2011 kit english asr system for the iwslt evaluation," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2011*, San Francisco, December 8-9 2011.

[3] M. Heck, K. Kubo, M. Sperber, S. Sakti, S. Stüker, K. Kilgour, C. Mohr, C. Saam, G. Neubig, T. Toda, S. Nakamura, and A. Waibel, "The kit-naist (contrastive) english asr system for iwslt 2012," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2012*, Hong Kong, December 6-7 2012.

[4] M. Wölfel and J. McDonough, "Minimum variance distortionless response spectralestimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, September 2005.

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

[6] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On using mlp features in lvcsr," in *Proceedings of ICSLP*. Citeseer, 2004.

[7] K. Kilgour, C. Saam, C. Mohr, S. Stüker, and A. Waibel, "The 2011 kit quaero speech-to-text system for spanish," 2011.

[8] A. W. Black and P. A. Taylor, "The Festival Speech Synthesis System: System documentation," Human Communciation Research Centre, University of Edinburgh, Edinburgh, Scotland, United Kongdom, Tech. Rep. HCRC/TR-83, 1997.

[9] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, May 2008.

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

[11] A. Venkataraman and W. Wang, "Techniques for effective vocabulary selection," *Arxiv preprint cs/0306022*, 2003.

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

[13] 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 *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006, ICSLP)*. Pittsburgh, PA, USA: ISCA, September 2006, pp. 521–524.

[14] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proceedings the IEEE Workshop on Automatic Speech Recognition and Understanding*. Santa Barbara, CA, USA: IEEE, December 1997, pp. 347–354.

[15] 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, October 2000.

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

[17] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 5, pp. 357–366, 1995.