The NAIST English Speech Recognition System for IWSLT 2013

Sakriani Sakti, Keigo Kubo, Graham Neubig, Tomoki Toda, Satoshi Nakamura

Augmented Human Communication Laboratory,
Graduate School of Information Science,
Nara Institute of Science and Technology, Japan
{ssakti,keigo-k,neubig,tomoki,s-nakamura}@is.naist.jp

Abstract

This paper describes the NAIST English speech recognition system for the IWSLT 2013 Evaluation Campaign. In particular, we participated in the ASR track of the IWSLT TED task. Last year, we participated in collaboration with Karlsruhe Institute of Technology (KIT). This year is our first time to build a full-fledged ASR system for IWSLT solely developed by NAIST. Our final system utilizes weighted finite-state transducers with four-gram language models. The hypothesis selection is based on the principle of system combination. On the IWSLT official test set our system introduced in this work achieves a WER of 9.1% for tst2011, 10.0% for tst2012, and 16.2% for the new tst2013.

1. Introduction

Similar to the IWSLT 2012 Evaluation Campaign [1], IWSLT 2013 featured an Automatic Speech Recognition (ASR) track in which systems are required to recognize the recordings made available by TED on their website[1]. TED talks bring together the world's most fascinating thinkers and doers, who are challenged to give the talk of their lives in about 5-25 minutes covering topics related to technology, entertainment and design (TED). Spanning everything, from internet trends to solving the world's water supply problems, today TED is a global movement "riveting talks by remarkable people free to the world".

This paper describes the NAIST English speech recognition system. The main challenge of this ASR track is to develop a system that is capable of recognizing spontaneous and open-domain speeches. Last year, we participated in collaboration with Karlsruhe Institute of Technology (KIT). This year is our first time to build a full-fledged ASR system for IWSLT solely developed by NAIST. Our system utilizes weighted finite-state transducers which is based on the Kaldi speech recognition toolkit [2]. Basically, our strategy in this year is to explore and investigate various acoustic features (MFCC, PLP, FBANK), front-end processing (LDA, STC, fMLLR, SAT), and acoustic models (HMM/GMM, SGMM, DNN) provided in the Kaldi toolkit, as well as various grapheme-to-phoneme strategy (Sequitur G2P, DirectTL+, Structured ARROW). However, in case of language models, only traditional four-gram language models were performed at the moment.

The final submission is based on the principle of system combination. The underlying assumption of system combination is that different systems commit different errors which may cancel each other out. However due to limited time, we were not able to submit the full-set combination system. The submitted system was a combination of (HMM/GMM MFCC + SGMM MFCC + HMM/GMM FBANK + SGMM FBANK + DNN FBANK). Furthermore, only half of data were used to train DNN-FBANK model. Nevertheless, experiment results reveal that in comparison with last year best system on the 2011 and 2012 evaluations set which serves as a progress test set, we were still able to reduce the word error rate of our transcription systems from 10.9% to 9.1% for tst2011 and from 12.1% to 10.0% for tst2012, giving a relative reduction of 16.5% and 17.4% respectively. And on the new official 2013 evaluation set, the final system reached a WER of 16.2%.

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 vocabulary and language model used for this evaluation in Section 5 and pronunciation lexicon including grapheme-to-phoneme conversion 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, we used TED talks released before the cut-off date of 31 December 2010, downloaded from the TED websites with the corresponding subtitles. The collected talk resulting in a total of 157 hours of speech.

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

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

Translation (WMT 2011), including EuroParl (EPPS), News Commentary (NC), and NEWS.

Table 1: Total size (word count) and vocabulary size of the individual text corpora.

Data	Size	Vocabulary
TED	2.7m	45k
EPPS	54m	82k
NC	4.5m	50k
NEWS	2,402m	1,047k

We normalized the text corpora of TED, EPPS, NC, and NEWS, in a case-insensitive fashion. Table 1 shows the resulting text corpora along with their total size (word count) and vocabulary size.

2.2. Test Corpora

Concerning the test corpora, the development and evaluation data sets (dev2010, tst2010, dev2012) used in past editions, were provided by IWSLT organizer for development purposes. As for evaluation purposes, evaluation data sets of tst2011 and tst2012 were used as the progress test set to compare the results of this year against the best results achieved in 2011 and 2012. Then, a new released test set for official test set of 2013 (tst2013) were used for final evaluation of our systems.

3. Front-End Processing

We investigated the use of three different kinds of acoustic front-ends: (1) mel-frequency cepstral coefficients (MFCC), (2) perceptual linear prediction (PLP)[3] and (3) log mel-filter bank (FBANK). The frontend provides features every 10ms with 25ms width. For each utterance in the speech training data, 13 static of acoustic features (MFCCs, PLPs, or FBANKs) including zeroth order for each frame are extracted and normalized with cepstrum mean normalization in order to have zero mean per speaker.

To incorporate the temporal structures and dependencies, 9 adjacent frames (4 frames on each left-right side of the current frame) of the acoustic features (MFCCs, PLPs, or FBANKs) are spliced together into one single feature vector leading to 117 dimensional super vectors (9x13 dimensions). These are then projected down to an optimum 40 dimensions by applying a linear discriminant analysis (LDA). After that, the resulting features are further de-correlated using maximum likelihood linear transformation (MLLT)[4], which is also known as global semi-tied covariance (STC)[5] transform. Moreover, speaker adaptive training (SAT)[6] is performed using a single feature-space maximum likelihood linear regression (fMLLR)[7] transform estimated per speaker.

4. Acoustic Model

Acoustic models are trained on the LDA+STC+fMLLR features describe above. We employed 39 phonemes of English based on CMU dictionary without stress information. Additionally, we added 9 special phoneme of non-speech sounds derived from TED speech sources. These include *SIL* for silence, *SENTSTART* and *SENTEND* for head and tail TED's sound effect, and *APPLAUSE*, *BEEP*, *LAUGHTER*, *MUSIC*, *NOISE*, and *VOICENOISE* for sound that appeared in TED speech sources.

Here we investigated the use three different kinds of acoustic models: (1) Hidden Markov Model/Gaussian Mixture Model (HMM/GMM) (2) Subspace Gaussian Mixture Models (SGMM) (3) Deep Neural Network (DNN) which are described below.

- **HMM/GMM**
  Three-state left-to-right HMM topology without skip states. The HMM units are derived from 39 phonemes of English. Each phoneme is classified by its position in word (4 classes: begin, end, internal and singleton). Context-dependent cross-word triphone HMMs were first trained with GMM output probability. The final model totally include 320K Gaussians trained with boosted maximum mutual information (MMI)[8] criterion of discriminative training.

- **SGMM**
  For SGMM, the Kaldi toolkits provides an implementation of the approach described in [9]. In this case, HMMs are builts with subspace GMM output probability. The final model consists of 9.1K states, which is also trained with boosted maximum mutual information (MMI) [8] criterion of discriminative training.

- **DNN**
  Here, we performed HMM/DNN hybrid framework, in which the network is trained with 7 layers, where each hidden layer has 2048 neurons. This DNN is initialized with stacked restricted Boltzmann machines (RBMs) that are pretrained in a greedy layerwise fashion.

5. Vocabulary and Language Model

5.1. Vocabulary

For the vocabulary selection, we followed an approach proposed by Venkataraman et al. [10]. We built unigram language models from all text sources, and combined them to satisfy unigram probabilities that maximize the likelihood of a held-out TED data set dev2010, by using the SRILM toolkit [11]. We then defined the 100k most probable unigrams from the combined unigram language model as the vocabulary.

5.2. LM Training

We constructed a 3-gram language model for decoding a utterance, and a 4-gram language model for rescoring hypotheses. At first, we built 3-gram and 4-gram language models with modified Kneser-Ney smoothing [12] from each of the text corpora by using kaldi LM toolkit[2]. These were then combined per n-gram language model 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 dev2010. Additionally, we pruned the n-gram entries that have a lower probability than 5e-10 in the combined 3-gram language model. For combining and pruning the language model, we employed the SRILM toolkit. The combined and pruned 3-gram language model contains 20 million bigrams, 45 million trigrams. The combined 4-gram language model contains 35 million bigrams, 194 million trigrams, and 397 million 4-grams. Perplexities on tst2010 for each 3-gram and 4-gram language model is shown in Table 2.

Table 2: Language model perplexities on tst2010 for each 3-gram and 4-gram language models. The n-gram entries that have a lower probability than 5e-10 in the 3-gram language model is pruned.

Data	3-gram	4-gram
TED	174.38	170.82
EPPS	450.38	429.14
NC	413.97	410.51
NEWS	200.63	192.30
Combined	138.58	127.72

6. Dictionary

6.1. G2P conversion

G2P conversion is employed to obtain a pronunciation of words that does not exist in a dictionary. We try three G2P conversion methods, (1) joint n-gram model [13] as implemented in Sequitur G2P (*Sequitur*), (2) DirecTL+ (*DirecTL+*) which is an online discriminative training based on MIRA for G2P conversion [14, 15] and (3) Structured AROW [16] which is also an online discriminative training that extends AROW [17] to structured learning (*SAROW*).

Table 3: Phoneme error rate (PER), word error rate (WER) and learning time (Time) for each G2P conversion methods in the CMU dictionary.

	PER(%)	WER(%)	Time(hr.)
Sequitur	6.77	28.55	17.5
DirecTL+	6.19	26.38	55.4
SAROW	6.15	26.48	28.5

We have compared these methods in a preliminary experiment in term of phoneme error rate (PER) and word error rate (WER). In the CMU dictionary, we have employed 10% as test data, 5% as development data and the reminder as training data. As showing in Table 3, *DirecTL+* and *SAROW* significantly improved over *Sequitur* in terms of PER and WER. The *SAROW* was almost the same performance as the *DirecTL+* in terms of PER and WER, while the *SAROW* improved the learning time of the *DirecTL+*. From the learning time of the *SAROW*, We determined to employ Structured AROW as our G2P conversion method in dictionary construction.

6.2. Dictionary construction

We first constructed a G2P model with Structured AROW as described above. Here, all data in the CMU dictionary were employed as training data. For some training parameters such as learning iteration, we re-used parameters employed in the preliminary experiment. After that, we applied the trained G2P model to a word that appears in the language model but does not appear in the CMU dictionary, except abbreviation words with all capitalized letters. The pronunciation of abbreviation words were constructed based on rule in which in each letter is converted to the corresponding single-letter pronunciation. The number of the converted word was 36k words in the 100k vocabulary.

7. Decoding Strategy and Results

Our decoding algorithms use weighted finite state transducers (WFSTs)[18] based on Kaldi speech recognition toolkit[2], a free, open-source toolkit for speech recognition research. The decoding-graph construction process is basically based on the conventional recipe described in [18] with slight modification to allow different phones to share the same context-dependent states.

7.1. Single System

Figure 1 shows the results given various configurations on the use of different acoustic features and acoustic models. For comparison we evaluated the performance on the development set. The results reveal that on each development set, DNN models with MFCCs, PLPs, or FBanks always outperformed HMM/GMM and SGMM. On the most left "dev2010" is the ASR performance on development set of 2010 given the segmentation data, while on the second one "dev2010 (no seg)" is the ASR performance on development set without time segmentation information. As can be seen, without time segmentation, the performance of ASR systems slightly reduced.

7.2. Combination System

Here, we investigate model combination system, feature combination system and full combination described below.

- **Model Combination System**
  Here, we focus to investigate the ASR performance

[2] http://merlin.fit.vutbr.cz/kaldi/kaldi_lm.tar.gz

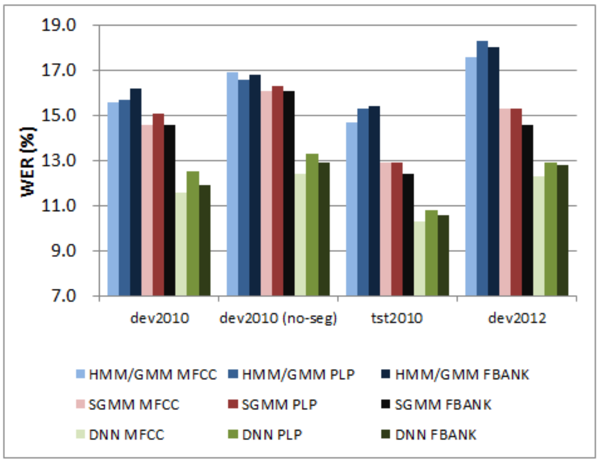

Figure 1: Performance of the single system on the development set and test set in WER.

of each acoustic features of MFCCs, PLPs, and FBANKs. Figure 2 shows the results of those acoustic features in combination of all acoustic models (HMM/GMM+SGMM+DNN). In average, the performance of those features are mainly the same. In most systems, the combination with optimum weight provide an improvement of the performances. Unfortunately, MFCC (HMM/GMM+SGMM+DNN) combination system performed worse than the best MFCC (DNN) single system. This is because the optimum weight was calculated at once (globally) based on the results of all single systems in all development sets, which may not be effective for all cases.

- **Feature Combination System**
  Here, we focus to investigate the ASR performance of each acoustic models of HMM/GMM, SGMM, and DNN. Figure 3 shows the results of those acoustic models in combination of all acoustic features (MFCCs+PLPs+FBANKs). The HMM/GMM always performed the worst. The best performance was achieve with DNN. However, the combination with optimum weight does not provide any significant improvement of the performances.

- **Full Combination System**
  Here, we perform feature and model combination system from 4-combination system to the full 9-combination system. Figure 4 shows the results of those acoustic models in combination of various acoustic features (MFCCs+PLPs+FBANKs) and various acoustic models (HMM/GMM+SGMM+DNN). The results reveal that the full 9-combination system provide a better performance than others. However, it is quite surprising that there is no significant improvement from 4-combination system to the full 9-combination system.

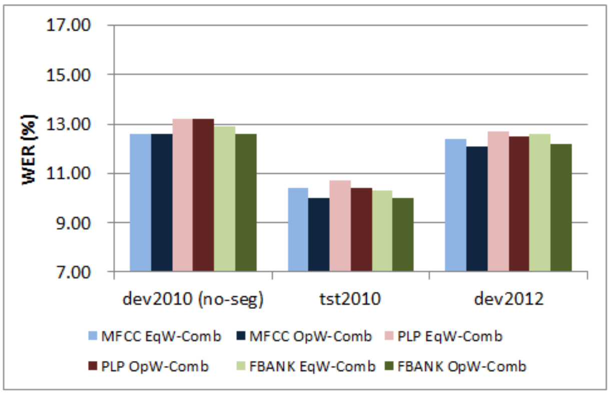

Figure 2: Performance of each acoustic features of MFCCs, PLPs, and FBANKs with acoustic model combination (HMM/GMM+SGMM+DNN) on the development set in WER.

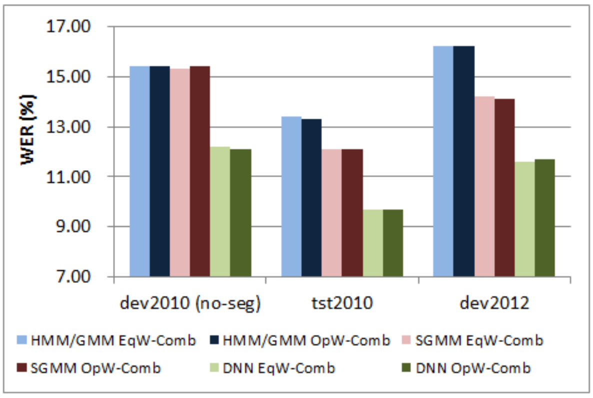

Figure 3: Performance of each acoustic models of HMM/GMM, SGMM, and DNN with acoustic features combination (MFCCs+PLPs+FBANK) on the development set and test set in WER.

7.3. Final Submission System

As we described previously, due to a limited time, we were not able to submit the full-set of 9-combination system. Our submitted primary system was based on a combination of (HMM/GMM MFCC + SGMM MFCC + HMM/GMM FBANK + SGMM FBANK + DNN FBANK). Table 4 shows the summary of our final system based on IWSLT 2013 evaluation feedback in comparison with the best system from feature combination, model combination, and the full 9-combination system.

In comparison with last year best system, experiment results reveal that the performance of the submitted system on the 2011 and 2012 evaluations set which serves as a progress test set, were still able to reduce the word error rate of our transcription systems from 10.9% to 9.1% for tst2011 and from 12.1% to 10.0% for tst2012, giving a relative reduction of 16.5% and 17.4% respectively. And on the new official 2013 evaluation set, the submitted system reached a WER of

16.2%. However, the best performance was provide by full 9-combination system which reached a WER of 15.6% giving another 3.7% relative reduction from the submitted system.

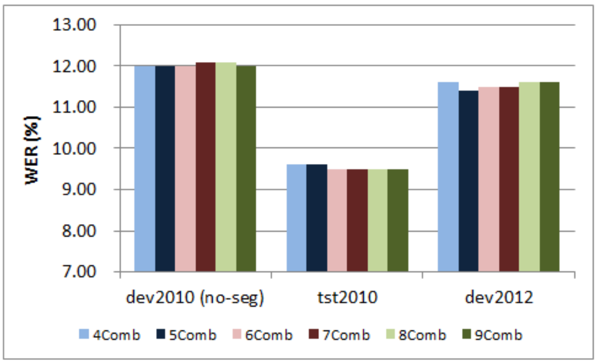

Figure 4: Performance of 4-combination system to the full 9-combination system on the development set and test set in WER.

ASR System	tst2011	tst2012	tst2013
Model Combination System	9.4%	10.4%	16.1%
Feature Combination System	9.2%	10.1%	16.0%
Full 9-Combination System	9.0%	9.7%	15.6%
Official Submitted System	9.1%	10.0%	16.2%

Table 4: Summary of final system performances based on IWSLT 2013 evaluation feedback in comparison with the best system from feature combination, model combination, and the full 9-combination system.

8. Conclusion

In this paper we described our English speech-to-text system with which we participated in the IWSLT 2013 TED task evaluation on the ASR track. The decoding strategy for the final submission is based on the principle of system combination. The underlying assumption of system combination is that different systems commit different errors which may cancel each other out. However due to a limited time, we are not able to submit the full-set combination system. The submitted system was a combination of (HMM/GMM MFCC + SGMM MFCC + HMM/GMM FBANK + SGMM FBANK + DNN FBANK). Nevertheless, experiment results reveal that on the 2011 and 2012 evaluations set which serves as a progress test set, we were still able to reduce the word error rate of our transcription systems from 10.9% to 9.1% for tst2011 and from 12.1% to 10.0% for tst2012, giving a relative reduction of 16.5% and 17.4% respectively. And on the new official 2013 evaluation set, the final system reached a WER of 16.2%. The best performance was provided by full 9-combination system which reached a WER of 15.6% giving another 3.7% relative reduction from the submitted system.

9. References

[1] M. Federico, L. Bentivogli, M. Paul, and S. Stueker, "Overview of the IWSLT 2012 evaluation campaign," in *Proc. IWSLT 2012*, Hong Kong, 2012, pp. 12–33.

[2] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Moticek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, Hawaii, USA, 2011.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1998.

[4] R. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proc. of ICASSP*, 1998, pp. 661–664.

[5] 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.

[6] R. S. T. Anastasakos, J. Mcdonough and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

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

[8] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. of ICASSP*, Las Vegas, USA, 2008, pp. 4057–4060.

[9] D. Povey, L. B. D. Povey, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "The subspace Gaussian mixture model – a structured model for speech recognition," in *Proc. of ICASSP*, Las Vegas, USA, 2008, pp. 4057–4060.

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

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

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

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

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

[15] 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.

[16] K. Kubo, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Grapheme-to-phoneme conversion based on adaptive regularization of weight vectors," in *Proc. INTERSPEECH*, 2013, pp. 1946–1950.

[17] K. Crammer, A. Kulesza, and M. Dredze, "Adaptive regularization of weight vectors," *Advances In Neural Information Processing Systems*, pp. 414–422, 2009.

[18] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, vol. 20, no. 1, pp. 69–88, 2002.