The NICT ASR System for IWSLT2011

Kazuhiko Abe, Youzheng Wu, Chien-lin Huang,Paul R. Dixon,
Shigeki Matsuda, Chiori Hori and Hideki Kashioka

National Institute of Information and Communication Technology
Kyoto, Japan
kazuhiko.abe@nict.go.jp

Abstract

In this paper, we describe NICT's participation in the IWSLT 2011 evaluation campaign for the ASR Track.

To recognize spontaneous speech, we prepared an acoustic model trained by more spontaneous speech corpora and a language model constructed with text corpora distributed by the organizer. We built the multi-pass ASR system by adapting the acoustic and language models with previous ASR results.

The target speech was selected from talks on the TED (Technology, Entertainment, Design) program. Here, a large reduction in word error rate was obtained by the speaker adaptation of the acoustic model with MLLR. Additional improvement was achieved not only by adaptation of the language model but also by parallel usage of the baseline and speaker-dependent acoustic models. Accordingly, the final WER was reduced by 30% from the baseline ASR for the distributed test set.

1. Introduction

In the IWSLT 2011 evaluation campaign [1], the Speech Communication Group of National Institute of Information and Communication Technology(NICT) participated in the English ASR track. This paper describes the speech recognition system developed for this campaign.

On this ASR track, the main challenge is spontaneous and open-domain speech recognition. After referring to the conventional research on similar spontaneous English speech recognition [2][3], we selected two publicly available corpora, European Parliament Plenary Session (EPPS) [4] and English broadcast news speech [5], for acoustic model training. For the pronunciation labeling, we use the CMU pronunciation dictionary [6] and Sequitur G2P [7]. Using these widely known speech corpora, tools and pronunciation dictionary, we made acoustic models, and using the text corpora determined by the organizer, we trained the language model. After making a search space with these models for our WFST-based speech recognition decoder, we built a baseline ASR system to obtain our initial results.

To reduce the word error rate (WER), our mutlti-pass ASR adapted the acoustic and language models for each talk using the initial results as trained label information. For the acoustic model, mean adaptation with maximum likelihood linear regression (MLLR) [8] was used. The baseline acoustic model and the applied model were combined to achieve more stable performance. For the language model, linear interpolation was carried out with the baseline language model and a trigram obtained from the previous stage in our multi-pass ASR system.

The organization of the remainder of the paper is as follows. Section 2 describes the implemented ASR system component. Then the performance of our ASR system is described in Section 3. A discussion and our conclusions are given in the final sections.

2. ASR System Components for IWSLT2011

This section gives an outline of our multi-pass ASR system, including the acoustic and language models as well as other components (Fig. 1).

The first ASR process used the baseline acoustic model (AMbase) and the baseline language model (LMbase) to obtain 1-best results. After initial decoding, an adapted acoustic model (AMadapt) with MLLR was made using these initial results. In the next process, AMadapt and AMbase were used in parallel. Adaptation of the language model also used the initial results. After making a small language model with the initial results (LMfirst-pass), LMbase and LMfist-pass were also combined. The following subsections describe the details of these adaptation methods.

2.1. Acoustic Modeling

2.1.1. Baseline acoustic models

The target of the ASR track is talks on the TED (Technology, Entertainment, Design) program. To build an ASR system for these spontaneous monologue speeches, we prepared three training sets. The first training set (train1) contains approximately 59 h of transcribed English EPPS speech data provided by TC-STAR project [4]. We eliminated the utterances which contain unclear parts or which have a low alignment score. The second set of training data (train2) consists of 198 h of broadcast news data from the HUB-4 corpus [5]. The third set of training data (train3) contains the sum total of the first and second sets.

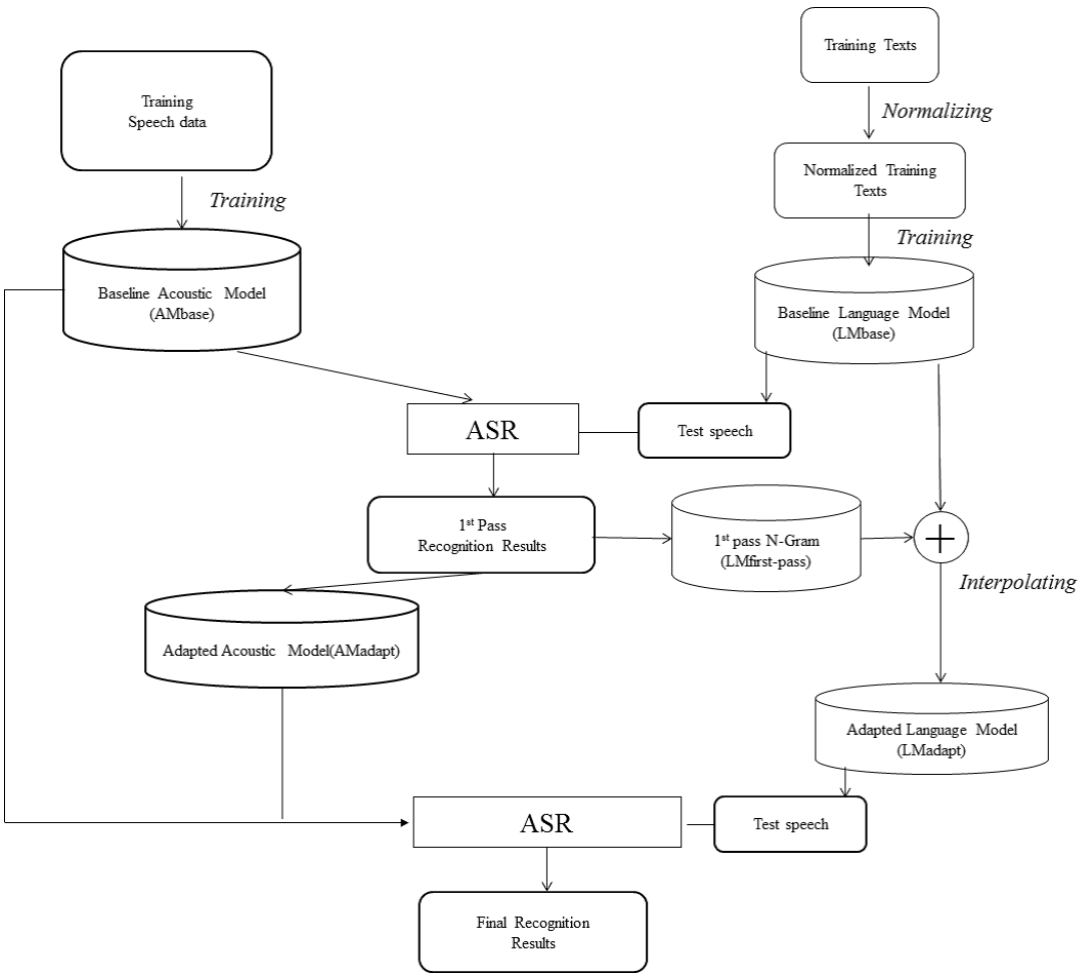

Figure 1: ASR system components.

Set	Data	Volume
train1	EPPS	59h
train2	HUB4	198h
train3	HUB4+EPPS	257h

Table 1: Training corpora for acoustic model

The acoustic model was trained by acoustic features based on mel-frequency cepstral coefficients (MFCC) and computed with a 25-ms frame length and a 10-ms frame shift. The features have 39 dimensions consisting of 13 static MFCCs (including energy) and their first- and second-order derivatives. For extracted features, a cepstral mean subtraction (CMS) technique was applied at the segment level. After applying CMS, histogram equalization (HEQ) was also applied. HEQ is commonly used in image processing, speech recognition and speaker recognition [10][11]. In HEQ, feature vector coefficients are considered independent of each other. HEQ provides a transformation that maps the histogram of each feature's vector component onto a reference histogram to achieve improved discrimination ability. The target distribution of HEQ is selected as a Gaussian in this study.

The acoustic modeling was based on across-word triphone states represented by left-to-right three-state hidden Markov models (HMMs). The number of triphone states was reduced by decision tree clustering. These models were trained using incremental splitting of Gaussians, followed by two iterations of Viterbi training.

Finally, we fixed the best training corpora and number of Gaussians based on the performance of the development set. As described in the following section, for the baseline acoustic model in particular, the models trained with the HUB4+EPPS (train3) were selected.

2.1.2. Unsupervised adaptation of Acoustic models

The baseline acoustic model was adapted with MLLR. For the labeling data, the ASR results of the previous stage were used. Only mean transforms were estimated; variances were not adapted. We applied a centroid splitting algorithm to construct a regression class tree, in which the number of base

clusters (classes) was set to 32. The adaptation data contained all utterances of a talk, and a single speaker-dependent model was created. However, ASR results sometimes contain errors. To reduce the influence by the mismatched label and problem of over-fitting, adapted acoustic model is used with the baseline acoustic model.

2.2. Language Modeling

2.2.1. Baseline language models

All of the monolingual training corpora were preprocessed before the models were trained. In this step, a non-standard-word-expansion tool was applied to convert non-standard words (such as CO2, 95%, and $3) to their pronunciations (CO two, ninety five percent, and three dollars). Statistics of the training corpora after preprocessing are shown in Table 2.

Corpus	Word Count
IWSLT11.TALK.train.en(TED)	2,063,299
europarl-v6	50,023,104
news-commentary-v6	3,880,801
news.2007	305,977,980
news.2008	759,301,946
news.2009	929,178,153
news.2010	361,006,759
news.2011	49,258,168
total	2,460,690,210

Table 2: Statistics of English LM training corpora

Then, the most frequent 100 k words were extracted from the preprocessed corpora, which, together with the CMU.v7 pronunciation dictionary, are used as the language models ' vocabulary. Finally, our vocabulary contains 157,753 entries and has an OOV rate of 0.78% in terms of the development data set. For each of the preprocessed corpora, a modified Kneser-Ney smoothed trigram language model was constructed using the MITLM toolkit [12]. The LMs were then interpolated by optimizing the perplexity of the development data set. For convenience of presentation, the interpolated LM is called LMbase.

2.2.2. Adapted language models

As the previous section described, language model adaptation was accomplished by training a smoothed trigram model (LMfirstpass) with the initial recognition results. To decrease the negative impact from errors in initial recognition results, those bigrams and trigrams that appear only once are discarded [13].

Finally, our adapted language model(LMadapt) can be expressed by

$$LMadapt = \alpha * LM_{base} + (1 - \alpha) * LM_{firstpass} \quad (1)$$

where α (interpolation weight) is set based on our experiment on the development data set.

2.3. Pronunciation Dictionary

We used the CMU dictionary v.7.1 [6], which consists of 39 phonemes, to provide the pronunciation definition for the acoustic model training data and recognition lexicon. For words not listed in the CMU dictionary, their pronunciation is defined by a statistical grapheme-to-phoneme (G2P) converter [7] that uses the converted model trained with the CMU dictionary.

2.3.1. WFST-based Speech Recognition Decoder

Our decoder is a general one-pass Viterbi decoder [14]. To use the speaker-independent and speaker-dependent acoustic models in parallel, we created a sets of lexicons and context-dependency transducers and these were combined and optimized according to

$$(C_{SI} \circ det(L)) \cup (C_{SD} \circ det(L)) \quad (2)$$

where $\cup$ is the union operation and SI and SD subscripts indicate speaker-independent and speaker-dependent, respectively. The final recognition cascade we used was

$$(C_{SI} \circ det(L)) \cup (C_{SD} \circ det(L)) \circ G \quad (3)$$

The composition of the language model G was performed on-the-fly using look-ahead composition. This construction allowed us to drive several search networks in parallel with the decoder in a memory-efficient manner. This highlights one of the advantages of the WFST framework for speech recognition. The search network was modified to allow parallel decoding, while requiring no code changes to the decoder core itself.

3. Experimental Results

We checked and tuned our ASR system for IWSLT with the distributed development and test sets, which were defined for IWSLT2011. The development set consists of 8 TED talks (dev2010) and the test set contains 11 of these talks (tst2010). After checking the performance of NICT ASR system, we executed it with the official test set (tst2011) to report our results.

In this section we report the performance for tst2010. Each talk is segmented for each utterance, and the number of utterances is 1,664 for the test set. This development set was used to tune not only the acoustic model and language model specifications but also the decoding parameters, especially search beam number, insertion penalty, and weights of acoustic and language model. The number of Gaussians per state was set to 24. The number of iterations for the adaptation stage was fixed at 2, and the interpolation weight was was fixed at 0.8. Accordingly, the performance of our ASR system was checked with this test. WER was calculated with the distributed tools and settings. In the next subsection, we

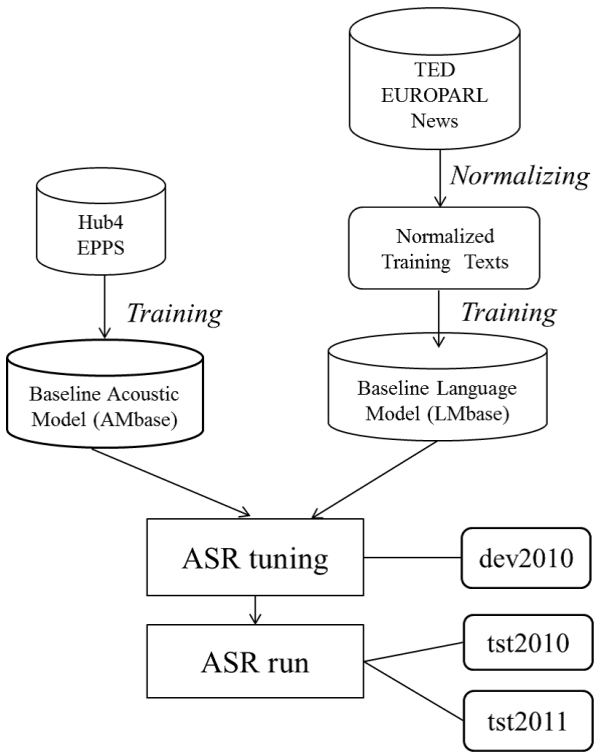

Figure 2: Development steps for IWSLT2011.

show the performance of our ASR system with the baseline acoustic and language models. Following that, we describe the performances obtained using the adapted models.

3.1. Performance of ASR System with Baseline Models

After setting the parameters for ASR, we compared the performance of the first stage.

The language model was trained with the distributed corpus and their perplexity for baseline language model (LMbase) was 144.643 (Table 3) Using this language model, the baseline decoding was processed.

Corpus	Perplexity
IWSLT11.TALK.train.en(TED)	184.126
+europarl-v6	172.551
+news-commentary-v6	170.637
+news.2007	156.599
+news.2008	148.994
+news.2009	145.733
+news.2010	145.733
+news.2011	144.665
total(LMbase)	144.643

Table 3: Perplexity on tst2010 by optimizing dev2010

Table 4 shows the experimental results for the acoustic model trained by each training data set (train1, train2 and train3).

Data set	WER(%)
train1	44.5
train2	39.2
train3	37.3

Table 4: Performance of training speech data for acoustic model

The acoustic model trained with EPPS speech data (train1) was worse than the model trained with HUB4 (train2). However, within the combination of these two sets (train3), train1 slightly contributed to the reduction of WER. Consequently, for our ASR system, this acoustic model trained with train3 was set as the baseline acoustic model.

3.2. ASR Results with Adapted Models

Table 5 shows the performances of the baseline and adapted models.

Models	WER(%)
AMbase	37.3
AMadapt	29.4
AMbase+AMadapt	26.0
AMbase+AMadapt, LMAdapt	25.7

Table 5: Performance of adaptation model

The average number of utterances and words per talk were 151.3 and 2511, respectively. In this condition, the adapted acoustic model effectively reduced the WER from the baseline, resulting in a reduction rate of 21.2%. Furthermore, the parallel usage of acoustic model was also effective, at an 11.6% reduction rate. Although the acoustic model adaptation process was fairly effective, the improvement of perplexity is not sufficient (Table 6) and the effect on the language model was small.

LM	Perplexity
LMbase	144.643
LMadapt	132.290

Table 6: Comparison of Perplexity

We also checked the performance for each talk as shown in Fig. 3. For all talks, MLLR adaptation and model merging were effective. In particular, the error reduction of adaptation was made ranging from 21.3% to 37.7%. However, we found no correlation with gender, the number of utterances, or speaking rate.

4. Discussion

From the results of the previous section, acoustic model adaptation in particular reduces the error rate of the baseline

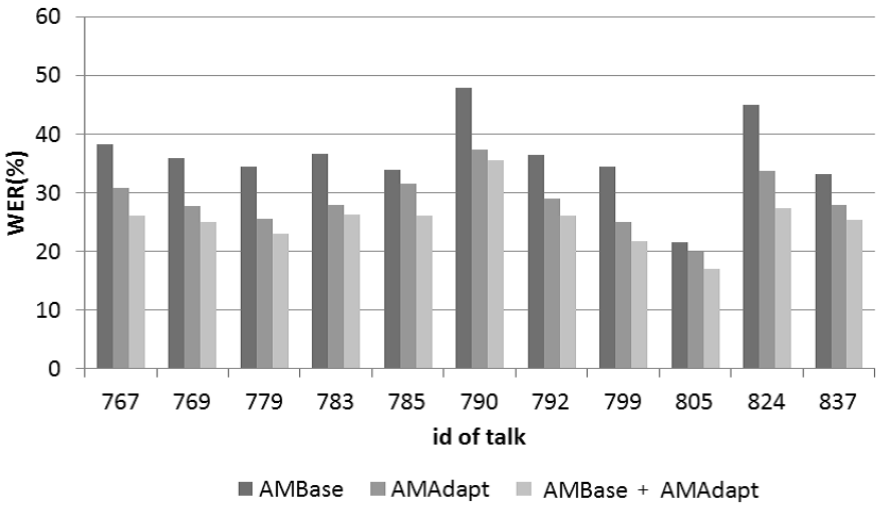

Figure 3: Performance of adapted acoustic models for each talk

system. A combination of techniques is also effective for this task. However, for a more stable ASR system, it is clear that we should consider many additional techniques. These might include, for example, such widely used approaches as training data clustering and discriminative training. Furthermore, we must also give careful attention to the quality of speech, especially noise and reverberation. For the acoustic adaptation, we should compare other adaptation methods and the current system. We need to clarify not only the performance but, also the relationship between effectiveness and the features of a given talk. The effect of language model adaption is smaller than that of acoustic model adaptation. Therefore, we should also consider state-the-art techniques to improve the language models and their adaptation method.

5. Conclusions

NICT ' s ASR system for IWSLT is a multi-pass ASR system. Acoustic model adaptation was effective in reducing the WER of a speaker-independent system. Moreover, the adapted acoustic model was more effective when combined with the baseline acoustic model.

6. Acknowledgements

The authors would like to thank Prof. Kawahara, Dr. S. Sakti and Dr. Y. Tsao for discussions on this research.

7. 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), San Francisco, CA, Dec. 2011

[2] Markus Nubaum-Thom, Simon Wiesler, Martin Sundermeyer, Christian Plahl, Stefan Hahn, Ralf Schlüter, Hermann Ney, "The RWTH 2009 Quaero ASR Evaluation System for English and German", In Proc. of Interspeech 2010, pp. 1517—1520

[3] M. Sundermeyer, M. Nubaum-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. of ICASSP 2011, pp. 2212-2215

[4] H. v. d. Heuvel, K. Choukri, C. Gollan, A. Moreno, and D. Mostefa, "TC-STAR: New language resources for ASR and SLT purposes", in ICASSP, Philadelphia, PA, USA, 2005

[5] Fiscus J, Garofolo J, Przybocki M, Fisher W, Pallett D (1998) 1997 English broadcast news speech (HUB4), catalog nbr LDC98S71, Linguistic Data Consortium, Philadelphia, PA

[6] R.L. Wide (1998, 31/7/2011). The CMU Pronunciation Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[7] Bisani, M. and Ney, H. "Joint-Sequence Models for Grapheme-to-Phoneme Conversion", Speech Communication, Volume 50, Issue 5, May 2008, Pages 434—451

[8] 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

[9] Christian Gollan, Maximilian Bisani, Stephan Kanthak,Ralf Schlter, and Hermann Ney. "Cross domain automatic transcription on the tc-star epps corpus",

in Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05), Philadelphia,PA, USA, March 2005

[10] A. Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Bentez and A. J. Rubio, "Histogram Equalization of Speech Representation for Robust Speech Recognition", IEEE Trans. on Speech and Audio Processing, vol.13, no. 3, pp. 355-366, 2005

[11] Chien-Lin Huang, Yu Tsao, Chiori Hori, and Hideki Kashioka, "Feature Normalization and Selection for Robust Speaker State Recognition", in Proc. Oriental COCOSDA, Hsinchu, Taiwan, 2011, pp. 102-105

[12] Bo-June (Paul) Hsu. "Language Modeling in Limited-Data Domains" PhD thesis, Massachusetts Institute of Technology, 2009

[13] Nanjo, H. Kawahara, T., "Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition", IEEE Transactions on Speech and Audio Processing (2004), 12(4): 391-400.

[14] Paul R. Dixon, Chiori Hori, Hideki Kashioka "A Comparison of Dynamic WFST Decoding Approaches", Submitted to ICASSP 2012