Data Selection from Multiple ASR Systems' Hypotheses for Unsupervised Acoustic Model Training

Sheng Li, Yuya Akita, Tatsuya Kawahara

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
Email: lisheng@sap.ist.i.kyoto-u.ac.jp

Abstract

This paper addresses unsupervised training of DNN acoustic model, by exploiting a large amount of unlabeled data with CRF-based classifiers. In the proposed scheme, we obtain ASR hypotheses by complementary GMM and DNN based ASR systems. Then, a set of dedicated classifiers are designed and trained to select the better hypothesis and verify the selected data. It is demonstrated that the classifiers can effectively filter usable data from unlabeled data for acoustic model training. The proposed method achieved significant improvement in the ASR accuracy from the baseline system, and it outperformed the models trained from the data selected based on the confidence measure scores (CMS) and also from the simple ROVER-based system combination.

Index Terms: speech recognition, acoustic model, unsupervised training, lecture transcription

1. Introduction

While the performance of acoustic model for speech recognition depends on the size of the training data, it is very costly to prepare accurate and faithful transcripts. We investigate an unsupervised training scheme, which takes the advantage of a large amount of unlabeled data, particularly for the deep neural network (DNN) acoustic model. As described in [1][2][3][4][5], the complete procedure of unsupervised training with unlabeled data includes pre-processing (e.g. speech segmentation, non-speech removal, speaker diarization, etc.), automatic transcription generation, and data selection before model training. We focus on the automatic transcription generation and data selection as the most crucial part of this task.

For data selection, the most commonly used method is based on the confidence measure scores (CMS) computed by the ASR system [8][9][10][11][12][13]. The word-level CMS is averaged over the utterance unit for data selection. When tuning the threshold of CMS, there is a trade-off between the data increase and the growth of noise in the label. It is not straightforward to find the optimal threshold and it is not practical to conduct exhaustive searching. Moreover, the optimum threshold depends on the available data size. This means that we need to tune the threshold every time the data size is increased and the ASR system is updated. Instead of using CMS, context-dependent state distribution [6] and global entropy reduction [7] can also be used for data selection. We investigate a discriminative approach that uses dedicated classifiers to select usable data for model training. In recent years, conditional random fields (CRF) models [18], which can combine multiple sources such as acoustic, lexical and linguistic features with contextual information, are used for confidence estimation [19][20] and a variety of other classification tasks, e.g. [24][25][29][30].

We have applied the scheme to the lightly supervised training [26] setting, where closed caption text is available and combined with an ASR hypothesis [27]. However, the assumption of closed caption text limits the applicability of the method. In this work, we extend to the more general unsupervised setting. We can leverage the text quality by combining hypotheses from a set of complementary ASR systems with similar accuracy and enough diversity on recognition patterns [14]. Deng et al. [15] demonstrated enough diversity exists between GMM and DNN systems. Conveniently, we can reuse the GMM-HMM system that is produced in the process of the DNN-HMM acoustic model training as a complementary system. Conventionally, ROVER-based system combination [16] has been used, but it is not robust to the small number of complementary systems with different distributions of CMS. The hypothesis combination can be formulated as a classification problem [21][22], but conventionally it is not integrated with hypothesis verification. In this study, the problem is solved by using a cascade of CRF classifications. In the proposed method, the CRF-based classifiers are prepared for two sub-tasks: selector CRF and verifier CRF. The selector CRF is trained to select a correct (or better) hypothesis either from GMM-HMM or DNN-HMM on the character/word level. The verifier CRF is then used to determine whether the selected result is correct or wrong. Data selection for acoustic model training is conducted according to the verification result.

In the remainder of the paper, we first describe the corpus of Chinese spoken lectures and the baseline ASR system in Section 2. Next, the proposed scheme for unsupervised training is formulated in Section 3. Then, the implementation of the method and experimental results are presented in Section 4. The paper is concluded in Section 5.

2. Corpus and Baseline ASR System

2.1. Data Preparation

We have designed and constructed the Corpus of Chinese "Lecture Room" (百家講壇) [23], which is a popular academic lecture program of China Central Television (CCTV) Channel 10. Since 2001, a series of lectures have been given by prominent figures from a variety of areas. The closed caption text is also provided by CCTV and free-download from the official website for a part of the lectures.

For the experimental purpose, we select 58 annotated lectures as the training set (CCLR-SV) and 19 annotated lectures as the test set (CCLR-TST). Additionally, 12 annotated lectures are held out as a development set (CCLR-DEV). Another set of 126 lectures that have closed caption texts only are used for lightly supervised training (CCLR-LSV)

[27]. The CCLR-USV set is totally unlabeled, and are used for additional training in this work. It has 184 lectures (35 multi-speaker and 149 single-speaker) in total 248 speakers and 114.7 hours. All these data sets are listed in Table 1.

Table 1 *Data sets in CCLR.*

	Data Set	#Lectures	Duration (hours)
Train	CCLR-SV	58	35.2
	CCLR-LSV	126	62.0
	CCLR-USV	184	114.7
Dev	CCLR-DEV	12	7.2
Test	CCLR-TST	19	11.9

2.2. Baseline ASR Systems

The dictionary for ASR consists of 53K lexical entries extracted from CCLR-SV together with Hub4 and TDT4. The OOV rate on CCLR-TST is 0.368%. The pronunciation entries were derived from the CEDICT open dictionary.

A word trigram language model (LM) was built for decoding. We interpolated the faithful annotation of CCLR-SV and closed caption texts of CCLR-LSV with related LDC corpora (Hub4, TDT, GALE) and the Phoenix lecture archive.

We adopt 113 phonemes (consonants and 5-tone vowels) as the basic HMM unit. We first built GMM-HMM and then DNN-HMM systems. The GMM system uses PLP features, consisting of 13 cepstral coefficients (including C0), plus their first and second derivatives, leading to a 39-dimensional feature vector. For each speaker, cepstral mean normalization (CMN) and cepstral variance normalization (CVN) are applied to the features. The DNN system uses 40-dimensional filterbank features plus their first and second derivatives with splicing 5 frames on each side of the current frame. It has 1320 nodes as input, 3000 nodes as output, and 7 hidden layers with 1024 nodes per layer. Training of DNN consists of the unsupervised pre-training step and the supervised fine-tuning step. They are implemented with Kaldi toolkit (nnet1) [28]. For decoding, we use Julius ver.4.3.1 (DNN version[1]) using the state transition probabilities of the GMM-HMM. This baseline system achieved an average Character Error Rate (CER) of 24.2% and 27.5% with the MLLR speaker-adapted GMM system, and 22.7% and 25.7% with the DNN system for CCLR-DEV and CCLR-TST, respectively.

3. CRF-based Hypothesis Combination and Data Selection

We propose an effective system combination and data selection scheme with CRF-based classifiers as shown in Figure 1. The flowchart is as follows:

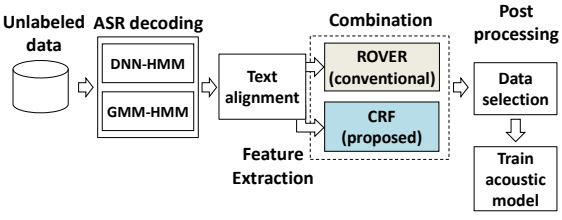

Fig. 1 *Flowchart of proposed method.*

[1]Available at http://julius.osdn.jp/en_index.php#latest_version

3.1. Process Flow

1) Preprocessing and Hypothesis Generation

For pre-processing, we first conduct speech segmentation to the utterance unit based on the BIC (Bayesian Information Criterion) method [32] and speaker clustering to remove non-speech segments and speech from other than the main lecturer in CCLR-USV. And then the unlabeled data is decoded by the DNN system and the speaker adapted GMM system, respectively.

2) Hypotheses Combination and Verification

Since different recognition patterns are observed between GMM and DNN based recognition hypotheses, we use CRF models to combine these diversities with their contextual information and determine which hypothesis should be selected for acoustic model training. At first, features are extracted from pair-wise aligned texts on the character level. Note that each Chinese character represents a syllable and has a corresponding meaning [35][36]. We adopt the character unit in order to avoid the mis-alignment due to different word segmentations and OOV problem. Moreover, as the size of characters is much smaller than the vocabulary size, we can train CRF models more efficiently. Then, a correct (or better) hypothesis is selected from complementary hypotheses and verified.

3) Post-processing and Acoustic Model Training

Data selection for acoustic model training is conducted by aggregating the result of the CRF classifications in the utterance level. The DNN system is retrained by adding the selected data.

3.2. Category of Alignment Patterns

We automatically transcribed the CCLR-SV data and made a three-way character alignment among these two ASR hypotheses by the GMM system and the DNN system and also the faithful transcripts (reference). By analyzing the aligned character sequence, we can categorize patterns into five classes, as shown in Table 2. The insertion and deletion cases are handled by using a null token. The definition of the category is as follows:

- *C1*: the DNN hypothesis is matched with the GMM hypothesis and also the correct transcript.
- *C2*: although the DNN hypothesis is matched with the GMM hypothesis, neither of them is correct.
- *C3*, *C4* and *C5*: the DNN hypothesis is different from the GMM hypothesis. In *C3*, neither of them is correct. In *C4*, the DNN hypothesis is correct. In *C5*, the GMM hypothesis is correct.

Table 2 *Category of alignment patterns.*

Category	DNN hypothesis		GMM hypothesis		reference text	Percent %
C1	**发**	√	**发**	√	**发**	**75.2%**
C2	**沦**	x	**沦**	x	论	**6.8%**
C3	雪	x	学	x	发	**6.6%**
C4	**法**	√	发	x	**法**	**7.7%**
C5	雪	x	**学**	√	**学**	**3.7%**

(√ means matching with reference, x means mismatching)

3.3. Design of Classifiers

We use CRF [18] as the classifier for this task. It can model the relationship between the features and labels by considering sequential dependency of contextual information.

Our objective is to accept useful data (*C1*, *C4* and *C5*) and remove erroneous data (*C2* and *C3*). We initially tried to design a flat classifier and cast the data selection and verification problem as a five-class classification problem, but it turned to be difficult because of the complex decision and the data imbalance. Therefore, we adopt a cascaded approach.

In the cascaded approach, we design two kinds of binary classifiers: selector CRF and verifier CRF. The selector CRF (CRF-1) is for selection between the hypotheses, and the verifier CRF (CRF-2 and 3) is for verification of the selected hypothesis. As described in the previous subsection, *C1* and *C2* are the matching cases between two different ASR hypotheses. In these cases, the data selection problem is reduced to whether to accept or discard the hypothesis. This is done by CRF-2. On the other hand, *C3*, *C4* and *C5* are the mis-matching cases between these two ASR hypotheses. We train a binary classifier (CRF-1) to make a choice between these ASR hypotheses. Then, we apply another classifier (CRF-3) to verify it. This classifier (CRF-3) should be different from the one used for *C1* and *C2* (CRF-2), because different kinds of information from GMM and DNN based systems are used.

The classification is organized by the three binary classifiers in a cascaded structure as illustrated in Fig. 2. The binary classifiers are focused on specific classification problems, so they are easily optimized. This design also mitigates the data imbalance problem. In Fig. 2, one classifier is used for selection of the hypothesis with highest credibility either from the DNN hypothesis or the GMM hypothesis, and the other two are used for verification of the selected (or matched) hypothesis. To make binary classification in the selector CRF (CRF-1), we merge *C3* into *C5*, because we observed the recognition accuracy of the DNN hypothesis is higher than that of the GMM hypothesis in the samples of *C3*. Erroneous patterns in *C3* will be rejected by the verifier-CRF (CRF-3).

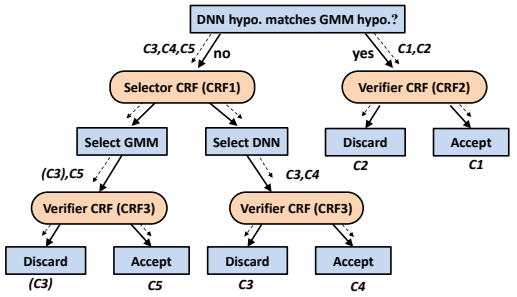

Fig. 2 *Cascaded classification scheme for data selection.*

3.4. Feature Design

We design the features for CRFs based on two groups: ASR-based features and text-based features listed in Table 3.

The ASR-based features are extracted for word unit, and distributed to each character in the word. They are numeric features output by the Julius decoder. The text-based features are extracted by rescoring and syntactic analysis in the character level.

Because most of the CRF implementations are designed to work with symbolic features, we need to convert the numeric features (**CMS, DUR, WLM, AM, NLW, NRW, DEN, CLM**) into discrete features. Moreover, for the symbolic features (**LEX, POS, BO**), the contextual information of the current unit (character) is also incorporated by adding the features of the preceding two characters and the following two characters.

Table 3 *Feature design.*

Categorize	Features
ASR-based feature	1. Confidence measure score of current word (**CMS**) [17]. 2. Duration of the current word (**DUR**). 3. Word trigram LM score (**WLM**). 4. Acoustic model score averaged per frame (**AM**). 5. Number of words connecting with current word to the left side of the lattice (**NLW**). 6. Number of words connecting with current word to the right side of the lattice (**NRW**). 7. Number of words overlapping with current word in the lattice (**DEN**).
Text-based feature	1. Lexical entry of current character (**LEX**). 2. Part-Of-Speech for each character unit (**POS**)[31]. 3. 5-gram char LM probability (**CLM**). 4. 5-gram char LM back-off behavior (**BO**).

For the selector CRF (CRF-1) and the verifier CRF (CRF-2), features from the GMM hypothesis and the DNN hypothesis are concatenated together, and the complementary information from both independent ASR systems can help make better classification. After a preliminary evaluation of the feature set, we adopt the complete feature set for the CRF-1 and CRF-2.

For the verifier CRF (CRF-3), we add the posterior probability output of CRF-1 and re-generate text-based features based on the context determined by CRF-1.

4. Experimental Evaluations

4.1. Classifier Implementation and Performance

In our implementation, we used the CRFSuite package[2] to train classifiers using CCLR-SV: CRF-2, which is trained to discriminate *C1* vs. *C2*, CRF-1, which is trained to discriminate *C3*+*C5* vs. *C4*, and CRF-3, which is trained to verify the output of CRF-1.

In the training data set (CCLR-SV), there is serious imbalance in training samples between classes. The distribution of these patterns in CCLR-SV is shown in Table 2. It is observed that 75.2% of them are categorized into *C1*. Other four classes are 6.8% (*C2*), 6.6% (*C3*), 7.7% (*C4*) and 3.7% (*C5*), respectively. This distribution will bias the training of the classifiers. Thus, we introduce a re-sampling technique. Specifically, we discarded part of samples which appear very frequently in *C1*. As a result, the calibrated distributions are as follows: *C1*:60.3%, *C2*:10.9%, *C3*+*C5*:16.6% and *C4*:12.2%.

We partition the CCLR-SV data into five segments, and derived the training data of CRF-3 using five-fold cross validation. In the validation, we trained an individual CRF-1 for each data partition. The ratio of positive samples (with the label "*accept*") against negative samples (with the label "*reject*") is 87.2% versus 12.8%.

To minimize the information loss in quantization, the numeric values are discretized with the method[3] described in [34]. The same kind of numeric features from the DNN and GMM systems can have different quantization levels.

[2] Available at http://www.chokkan.org/software/crfsuite/
[3] Available at http://www.irisa.fr/texmex/people/raymond/Tools/tools.html

In the experiment, we use a linear-chain CRF. The standard Limited-memory BFGS (L-BFGS) [32] algorithm and L2 regularization are used to train the CRF models with the sparse features of a high dimension.

Classification performance is measured by precision and recall. The confusion matrix with all features is shown in Table 3 and Table 4, and the classification rate is *C1*: 96.16%, *C2*: 49.13%, *C3+C5*: 61.01%, *C4*: 78.45%. Although the error rate by CRF-1 in the first stage of classification is not small, part of them are detected and discarded in the second stage of classification by CRF-3. We also test the performance of CRF-3 as shown in Table 5. We notice the false acceptance rate in CRF-2 and CRF-3 (errors in *C2* and *reject*) is relatively high, but we can tolerate many of these classification errors caused by the homophonic characters, which widely exist in Mandarin Chinese.

Table 3 *Confusion matrix of CRF-1 on CCLR-DEV.*

REF\HYP	*C3+C5*	*C4*	Sum	Recall
C3+C5	5575	3563	9138	61.01%
C4	3468	12624	16092	78.45%
Sum	9043	16187	25230	/
Precision	61.65%	77.99%	/	/

Table 4 *Confusion matrix of CRF-2 on CCLR-DEV.*

REF\HYP	*C1*	*C2*	Sum	Recall
C1	70485	2812	73297	96.16%
C2	3525	3404	6929	49.13%
Sum	74010	6216	80226	/
Precision	95.24%	54.76%	/	/

Table 5 *Confusion matrix of CRF-3 on CCLR-DEV.*

REF\HYP	*Reject*	*Accept*	Sum	Recall
Reject	9333	8028	17361	53.76%
Accept	4880	77647	82527	94.09%
Sum	14213	85675	99888	/
Precision	65.67%	90.63%	/	/

4.2. DNN Acoustic Models Enhanced by Selected Data

Then, we make utterance selection based on the character acceptance rate (CA) as a result of the previous classification. It is not practical to tune the CA threshold by using the development set, as it would take so long to train the DNN model for each CA threshold value. Considering Spoken Chinese is highly homophonic, we tolerant some character errors in utterances and accept the utterances with their CA no lower than 70%. Further relaxing the threshold only degrades the ASR performance due to the increase of errors.

The proposed method is applied to CCLR-USV to train an enhanced acoustic model, which are tested on CCLR-TST. The DNN acoustic model is retrained by adding the data selected from unlabeled data (CCLR-USV) to the labeled data (CCLR-SV and CCLR-LSV). ASR performance of the enhanced model is evaluated on both of CCLR-DEV and CCLR-TST. The proposed data selection method is compared with other methods as follows:

- **Baseline GMM** and **baseline DNN**: the models are trained by only using CCLR-SV and CCLR-LSV as described in Section 2.
- **DNN (CMS)**: we select utterances from CCLR-USV using the baseline DNN system based on a threshold of averaged CMS score (CMS≥0.6). The optimal threshold was determined by using GMM (MLE) models and CCLR-DEV [27].
- **Combine-ROVER**: combine the ASR hypotheses of CCLR-USV from the baseline GMM and the baseline DNN systems using ROVER [16]. We select utterances according to the optimal threshold of the averaged CMS score (CMS≥0.6). It is the conventional method for leveraging hypotheses and data selection. We also use all of the combined ASR hypotheses of CCLR-USV without any selection (CMS≥0.0).
- **Combine-CRFs**: combine the ASR hypotheses of CCLR-USV from two different baseline systems by using a set of CRF models. This is our proposed method for leveraging hypotheses and data selection. Effect of data selection is investigated on three thresholds: CA≥0.0 (no selection), CA=1.0 (use utterances with all characters accepted), and CA≥0.7.

Table 6 *ASR performance by unsupervised training.*

	Amount of data (hours)		CER%	
	labeled	unlabeled	DEV	TST
Baseline GMM	97.2	0	24.2	27.5
Baseline DNN	97.2	0	22.7	25.7
DNN (CMS≥0.6)	97.2	97.1	22.2	25.4
Combine-ROVER (CMS≥0.0)	97.2	114.7	21.9	24.9
Combine-ROVER (CMS≥0.6)	97.2	82.3	21.9	25.0
Combine-CRFs (CA≥0.0)	97.2	114.7	21.5	24.4
Combine-CRFs (CA=1.0)	97.2	38.9	21.3	24.5
Combine-CRFs (CA≥0.7)	**97.2**	**78.3**	**21.1**	**24.2**

ASR performance in CER is listed in Table 6. In this experiment, we use the same setting with the baseline system described in Section 2 for the DNN model specification as well as the lexicon and the language model. The results show that our proposed unsupervised training method significantly improved from the baseline. It also outperforms all other methods on both evaluation data sets.

We observe that both of Combine-CRFs and Combine-ROVER outperform DNN (CMS ≥ 0.6). This suggests the system combination effectively leverages the quality of automatic generated transcription texts. The fact that our proposed method Combine-CRFs (CA ≥ 0.0) further outperforms the Combine-ROVER (CMS≥0.0) demonstrates the effectiveness of the CRF models using many features. The Combine-ROVER (CMS≥0.6) and Combine-ROVER (CMS≥0.0) has no significant difference, while the improvement by Combine-CRFs (CA≥0.7) is statistically significant compared with the other two models (CMS≥0.0 and CA=1.0) among our proposed method and the improvement by Combine-CRFs (CA=1.0) is also statistically significant compared with Combine-ROVER (CMS≥0.6). This confirms the data selection with the verifier CRF has some effect for further improvement.

5. Conclusions

We have proposed a new scheme for hypotheses leveraging and data selection for unsupervised training of DNN acoustic model. The method uses dedicated classifiers, which are trained with the training database of the baseline acoustic model, to combine complementary ASR hypotheses and select usable data for model training. We designed a cascaded classification scheme based on a set of binary classifiers, which incorporates a variety of features. Experimental evaluations show that the proposed unsupervised training method effectively filters usable data, and improves the ASR accuracy from the baseline model and in comparison with the conventional ROVER-based method.

References

[1] K. Yu, M. Gales, L. Wang and P. Woodland, Unsupervised training and directed manual transcription for LVCSR. Speech Communication, Vol52(7), pp.652-663, 2010.

[2] H. Liao, E. McDermott and A. Senior, Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. In Proc. ASRU, pp. 368-373, 2013.

[3] Y. Huang, D. Yu, Y. Gong and C. Liu, Semi-supervised GMM and DNN acoustic model training with multi-system combination and confidence re-calibration. In Proc. InterSpeech, pp. 2360-2364, 2013.

[4] K. Vesely, M. Hannemann and L. Burget, Semi-supervised training of deep neural networks, In Proc. ASRU, pp267-272, 2013.

[5] D. Imseng, B. Potard, P. Motlicek, A. Nanchen, H. Bourlard, Exploiting un-transcribed foreign data for speech recognition in well-resourced languages. In Proc. ICASSP, 2014.

[6] O. Siohan, "Training data selection based on context-dependent state matching," in Proc. ICASSP, pp. 3316–3319, 2014.

[7] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," Computer Speech & Language, vol. 24, no. 3, 2010.

[8] H. Jiang, Confidence measures for speech recognition: a survey, Speech Communication, vol. 45, no. 4, pp. 455-470, Apr. 2005.

[9] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," in Proc. EUROSPEECH, September 1997, vol. 2, pp. 827–830.

[10] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," IEEE Trans. Speech & Audio Process., vol. 9, no. 3, pp. 288–298, March 2001.

[11] A. Lee, K. Shikano, and T. Kawahara. Real-time word confidence scoring using local posterior probabilities on tree trellis search, In Proc. IEEE-ICASSP, Vol.1, pp.793--796, 2004.

[12] L. Mangu, E. Brill, and A. Stolcke. Finding Consensus Among Words: Lattice-Based Word Error Minimization. In Proc. Eurospeech'99, pp. 495-498, Budapest.

[13] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," In NIST Speech Transcription Workshop, 2000.

[14] K. Audhkhasi, A Zavou, P Georgiou, and S. Narayanan, "Theoretical analysis of diversity in an ensemble of automatic speech recognition systems," IEEE/ACM Transactions on Audio, Speech, and Language Processing, volu.22, no. 3, March 2014.

[15] L. Deng and J. Platt, "Ensemble deep learning for speech recognition." In Proc. INTERSPEECH, 2014.

[16] J. Fiscus, A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER), in IEEE-workshop ASRU, 1997.

[17] A. Lee, K. Shikano, and T. Kawahara. Real-time word confidence scoring using local posterior probabilities on tree trellis search, In Proc. IEEE-ICASSP, Vol.1, pp.793--796, 2004.

[18] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. ICML, 2001.

[19] M. Seigel and P. Woodland, Combining Information Sources for Confidence Estimation with CRF Models, In Proc. INTERSPEECH, 2011.

[20] J. Fayolle, F. Moreau, C. Raymond, and G. Gravier, "CRF-based combination of contextual features to improve a posteriori wordlevel confidence measures," Proc. Interspeech, 2010.

[21] D. Hillard, B. Hoffmeister, M. Ostendorf, R. Schlüter, and H. Ney, iROVER: Improving system combination with classification, in Proc. NAACL-HLT Companion Volume Short Papers, 2007, pp. 65–68.

[22] B. Hoffmeister, R. Schluter and H. Ney, "iCNC and iROVER: The Limits of Improving System Combination with Classification?" in Proc. Interspeech 2008, pp.232-235.

[23] S. Li, Y. Akita, and T. Kawahara, Corpus and transcription system of Chinese lecture room, In Proc. ISCSLP, 2014.

[24] J. Zhang, H. Chan, P. Fung and L. Cao. A Comparative Study on Speech Summarization of Broadcast News and Lecture Speech. In Proc. INTERSPEECH, pp. 2781-2784, 2007.

[25] S. Kong, M. Wu, C. Lin, Y. Fu, and L. Lee. Learning on Demand - Course Lecture Distillation by Information Extraction and Semantic Structuring for Spoken Documents. In Proc. INTERSPEECH, pp. 4709-4712, 2009.

[26] L. Lamel, J.L. Gauvain and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training", Computer Speech and Language, vol.16, pp. 115-129, January 2002.

[27] S. Li, Y. Akita, and T. Kawahara. Discriminative data selection for lightly supervised training of acoustic model using closed caption texts. In Proc. INTERSPEECH, pp.3526-3530, 2015.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, The Kaldi speech recognition toolkit, IEEE-ASRU. 2011.

[29] W. Chen, S. Ananathakrishnan, R. Kumar, R. Prasad, and P. Natarajan, ASR error dection in a conversational spoken language translation system, In Proc. ICASSP, 2013.

[30] M. Lehr, I. Shafran, E. Prud'hommeaux, and B. Roark, Discriminative Joint Modeling of Lexical Variation and Acoustic Confusion for Automated Narrative Retelling Assessment, In Proc. NAACL, 2013.

[31] M. Shen, H. Liu, D. Kawahara, and S. Kurohashi. 2014. Chinese Morphological Analysis with Character-level POS Tagging. In proceedings of the 52th Annual Meeting of the Association for Computational Linguistics (ACL), Short Paper, pages 253–258, Baltimore, USA, 2014.

[32] M.Mimura and T.Kawahara,"Fast speaker normalization and adaptation based on BIC for meeting speech recognition." In Proc. APSIPA ASC, 2011.

[33] J. Nocedal. "Updating Quasi-Newton Matrices with Limited Storage". Mathematics of Computation. 35. 151. 773-782. 1980.

[34] U. Fayyad and K. Irani, "Multi-interval discretization of continuous attributes for classification learning," In Proc. IJCAI, pp1022-1027, 1993.

[35] J. Luo, L. Lamel and J-L. Gauvain,"Modeling Characters versus Words for Mandarin Speech Recognition." In Proc. ICASSP, Taipei, Taiwan, 2009.

[36] X. Liu, J. L. Hieronymus, M. J. F. Gales and P. C. Woodland. Syllable Language Models for Mandarin Speech Recognition: Exploiting Character Sequence Models, Journal of the Acoustical Society of America, Volume 133, Issue 1, 519-528, January 2013.