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

SHENG LI[1,a)] YUYA AKITA[2,b)] TATSUYA KAWAHARA[3,c)]

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.

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 huge quantity of unlabeled data, particularly for the deep neural network (DNN) acoustic model. As described in [1, 2, 3, 4, 5, 6], the complete procedure of unsupervised training with unlabeled data includes preprocessing (e.g. speech segmentation, non-speech removal, speaker diarization, etc.), automatic transcription generation, and data selection before model training. Some recent studies [7, 8] extend the multi-task learning method for the multilingual acoustic modeling tasks [9, 10, 11] to the unsupervised training purpose, but the improvement is limited. In this paper, we focus on the automatic transcription generation and data selection as the most crucial part of this task, trying to solve several issues of the conventional paradigm of unsupervised training method.

For data selection, the most commonly used method is based on the confidence measure scores (CMS) computed by the ASR system [1, 2, 3, 4, 5, 6]. 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, we investigate a discriminative approach that uses dedicated classifiers to select usable data for model training. In recent years, conditional random fields (CRF) models [16], which can combine multiple sources such as acoustic, lexical and linguistic features with contextual information, are used for a variety of classification tasks including confidence estimation [17, 18].

We have applied the scheme to the lightly supervised training setting, where closed caption text is available and combined with an ASR hypothesis [20]. 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 complimentary ASR systems with similar accuracy and enough diversity on recognition patterns [12]. Deng et al. [13] 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 [14] has been used, but it is not robust to the small number of complementary systems with different distributions of CMS. 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 performance

2.1. Data Preparation

We have designed and constructed the Corpus of Chinese "Lecture Room" (百家講壇) [19], 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) [20]. The CCLR-USV set is totally unlabeled, and are used for additional training in this work. It has 184 lectures in total 248 speakers and 114.7 hours. All these data sets are listed in Table 1.

[1] School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan.
a) lisheng@ar.media.kyoto-u.ac.jp

Table 1 *Data sets in CCLR.*

Data Set	Corpora	#Lecture	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 system

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) [21]. For decoding, we use Julius ver.4.3.1 (DNN version[1]) [22] 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 Fig.1. The flowchart is as follows:

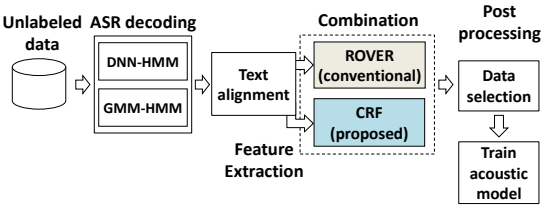

Fig.1 *Flow-chart of proposed method.*

3.1. Process flowchart

Preprocessing and Hypotheses Generation

For pre-processing, we first conduct speech segmentation to the utterance unit based on the BIC (Bayesian Information Criterion) method [26] 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.

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 [29, 30]. 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.

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**	√	**沦**	√	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)

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

3.3. Classifier design

We use CRF [16] as the classifier for this task. It can model the relationship between the features and labels by considering sequential dependencies of contextual information. For this reason, it is used for many applications such as confidence measuring [17, 18], ASR error detection [23], and automatic narrative retelling assessment [24].

Our objective is to accept effective 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 decisions 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 is for selection between the hypotheses, and the verifier CRF 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 word hypothesis. On the other hand, ***C3***, ***C4*** and ***C5*** are the mismatching cases between these two ASR hypotheses. We train a binary classifier to make a choice between these ASR hypotheses. Then, we apply the other classifier to verify it.

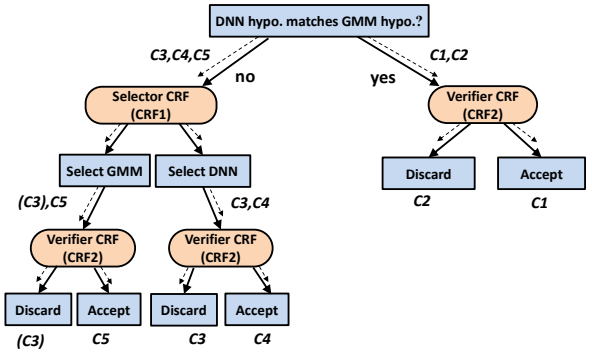

Fig.2 *Cascaded classifiers for data selection.*

The classification is organized by the two 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 word hypothesis with highest credibility either from the DNN hypothesis or the GMM hypothesis, and the other one is 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 it can make the data distribution more balanced. Erroneous patterns in ***C3*** (i.e. GMM hypothesis is incorrect) will be rejected by the verifier-CRF (**CRF-2**).

3.4. Feature design

The input features used in **CRF-1** and **CRF-2** are listed in Table 3 and Table 4. We categorize these features into two groups: ASR-based features and text-based features.

Table 3 *Feature sets for CRF-1.*

Feature Type	Features
ASR-based feature	1. Confidence measure score **(CMS).** 2. Duration of the current word (**DUR**). 3. Word trigram LM score **(WLM).** 4. Averaged acoustic model score **(AM).** 5. Number of left competing words (**NLW**). 6. Number of right competing words (**NRW**). 7. Density within word duration **(DEN).**
Text-based feature	1. Lexical feature **(LEX).** 2. Part-of-Speech **(POS).** 3. 5-gram char LM probability **(CLM).** 4. 5-gram char LM back-off behavior **(BO).**

Table 4 *Feature sets for CRF-2.*

Feature Type	Features
ASR-based feature	1. Confidence measure score of DNN system and posterior output of CRF-1 (**CMS**)
Text-based feature	1. Lexical feature **(LEX)** 2. Part-of-Speech **(POS)** 3. 5-gram char LM probability **(CLM)** 4. 5-gram char LM back-off behavior **(BO)**

These features are explained below.

The ASR-based features are extracted for word unit, and distributed to each character in the word. They are numeric features:

- The confidence measure score (**CMS**) is output by the Julius decoder [15] of the baseline ASR system. The value is between [0, 1] approximating a posterior probability of the hypothesis word.
- The word duration (**DUR**) feature is the number of frames of the word.
- The word trigram LM (**WLM**) feature is the word trigram language model score of the word while decoding.
- Averaged acoustic model score (**AM**) feature is the acoustic likelihood score averaged for each frame.
- The left competing words (**NLW**) feature is the number of the competing words to the left side of the current word in the word graph.
- The right competing words (**NRW**) feature is the number of the competing words to the right side of the current word in the word graph.
- The density (**DEN**) feature is how many words overlapping between the start time and the end time of the current word in the word graph.

The text-based features are extracted by rescoring and syntactic analysis in the character level:

- The lexical feature (**LEX**) is a lexical entry (ID) of the current character. It is a symbolic feature.
- The Part-of-Speech (**POS**) feature is obtained for each character unit by a CRF classifier trained with a character based Chinese-Tree-Bank (CTB) 4 [25]. This feature is symbolic.
- The language model probability feature (**CLM**) is a negative log probability of the current character rescored by a character 5-gram language model. This feature is numeric. When back-off is used, it is recorded as back-off behavior feature (**BO**). This feature is symbolic.

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 features of the preceding two characters and the following two characters.

For the selector CRF (**CRF-1**), 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.

For the verifier CRF (**CRF-2**), we recalculate the text-based features after the classification by selector CRF (**CRF-1**). Another feature we use is the posterior probability output of **CRF-1** (for the mismatching cases) and the confidence measure score of the DNN system (for the matching cases) as shown in Table 4.

3.5. Utterance selection for acoustic model training

The ASR hypotheses are merged into a single character sequence after the matching and selection process, and every character in the sequence will have a label, either "*accept*" or "*discard*", based on the verification process according to Fig.2.

Then, we need to make a decision whether or not this sequence of the data by the utterance unit is used for acoustic model training. We calculate the frame acceptance rate of each utterance, because the parameters of DNN are updated on the frame-level mini-batches. Using forced-alignment, we get the state-level label and their boundaries. In this way, the character-level labels can be distributed to all frames.

A more simplified method is we compute the character acceptance rate (CA) for every utterance to approximate the frame acceptance rate. Since Chinese is syllabic language and each character is a syllable, the "CA" actually means the ratio of "accept" syllables over the total number of syllables in an utterance. Considering spoken Chinese is highly homophonic, we tolerant a maximum character rejection rate up to 30% in each utterance.

4. Experimental evaluations

The proposed method is applied to CCLR-USV to make an enhanced acoustic model, which are tested on CCLR-TST.

4.1. Classifier implementation

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

Since the feature of **CRF-2** is depend on the result of **CRF-1**, We can use a five-fold cross validation method to get the features of **CRF-2**. Specifically, we partition the training data into five subsets, and train an individual **CRF-1** using 4/5 of the data to work on the rest 1/5 data.

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 too 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%. For model generalization, we also incorporate data from CCLR-LSV to enlarge the training data.

In the experiment, we use liner-chain CRF implemented in the CRFSuite package[1]. The standard Limited-memory BFGS (L-BFGS) [27] algorithm and L2 regularization are used to train the CRF models with the sparse features of a high dimension. Cut-off threshold for the occurrence frequency of feature is 1. The maximum number of iterations for L-BFGS optimization is 100. To minimize the information loss in the quantization, these numeric values are discretized with the method[2] described in [28]. The same kind of numeric features from DNN and GMM based system can have different quantization levels.

4.2. Classifier performances

Classification performance with various feature sets is evaluated on CCLR-DEV, as shown in Table 5 and Table 6. Performance is measured by *Precision*, *Recall* and *F-score*.

$$Precision = TP / FP$$
$$Recall = TP / (FP + FN)$$
$$F - score = 2 \times Precision \times Recall / (Precision + Recall)$$

where *TP* is true positives (correct output), *FP* is false positives (false alarm), and *FN* is false negatives (miss).

Table 5 *Feature Set Evaluation of CRF-1 on CCLR-DEV*

	CRF-1					
	Select GMM (C3 + C5)			*Select DNN (C4)*		
Feature	Recall	Precision	F-score	Recall	Precision	F-score
LEX	0.504	0.498	0.501	0.711	0.716	0.713
POS	0.458	0.449	0.453	0.681	0.689	0.685
CLM	0.471	0.530	0.499	0.763	0.717	0.739
BO	0.300	0.481	0.370	0.816	0.673	0.738
All Text	0.546	0.560	0.553	0.756	0.746	0.751
CMS	0.518	0.541	0.529	0.750	0.733	0.741
DUR	0.491	0.511	0.501	0.733	0.717	0.725
WLM	0.410	0.485	0.444	0.753	0.692	0.721
AM	0.468	0.498	0.483	0.732	0.708	0.720
NLW	0.491	0.455	0.472	0.667	0.697	0.682
NRW	0.491	0.465	0.478	0.679	0.701	0.690
DEN	0.483	0.458	0.470	0.677	0.697	0.687
All ASR	0.572	0.569	0.570	0.754	0.756	0.755
All Features	**0.610**	**0.617**	**0.613**	**0.785**	**0.780**	**0.782**

Table 6 *Feature Set Evaluation of CRF-2 on CCLR-DEV*

	CRF-2					
	Discard (C2+ C3)			*Accept (C1+C4+C5)*		
Feature	Recall	Precision	F-score	Recall	Precision	F-score
LEX	0.044	0.697	0.082	0.996	0.832	0.907
POS	0.002	0.730	0.003	0.999	0.826	0.905
CLM	0.088	0.684	0.155	0.992	0.838	0.908
BO	0.013	0.679	0.025	0.999	0.828	0.905
All Text	0.237	0.662	0.350	0.975	0.859	0.913
CMS (ASR)	0.631	0.588	0.609	0.907	0.921	0.914
All Features	0.621	0.627	**0.624**	0.922	0.920	**0.921**

We observe the overall performance of **CRF-2** (Table 6) is higher than that of **CRF-1** (Table 5). It suggests selection of the hypothesis is more difficult than verification of the hypothesis.

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

Among the feature sets, the text-based features and their combinations are generally less effective than ASR-based feature in **CRF-1** and **CRF-2**. But for both classifiers, combination of both feature sets shows further improvement. As an individual feature, the **CMS** feature is the most effective for **CRF-1** and **CRF-2**.

From these results, we adopt the complete feature set. Although errors by **CRF-1** in the first stage of the classification is inevitable, part of them are detected and discarded in the second stage of classification by **CRF-2**, as shown in Fig. 2.

4.3. ASR performance with enhanced model training

Then, we conduct DNN acoustic model training by adding the data selected from CCLR-USV to the CCLR-SV and CCLR-LSV. ASR performance of the model enhanced by the selected data 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** model: 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 [20].
- **Combine-ROVER**: combine the ASR hypotheses of CCLR-USV from the baseline GMM and the baseline DNN systems using ROVER [14]. 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.

In this experiment, we use the same setting with the baseline system described in Section 2 for DNN acoustic model training and testing as well as the lexicon and the language model.

ASR performance in CER is listed in Table 7. 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. This confirms the data selection with the verifier CRF has some effect for further improvement.

Table 7 ASR Performance (CER%) by Unsupervised Training
(Measured with NIST SCLite Scoring Tool)

	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	22.8	26.2
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**

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. IEEE-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. IEEE-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. IEEE-ICASSP, 2014.
[6] P. Zhang, Y. Liu, and T. Hain, “Semi-supervised dnn training in meeting recognition,” in Proc. SLT, 2014.
[7] H. Su, H. Xu, “Multi-softmax Deep Neural Network for Semi-supervised Training”, In Proc. INTERSPEECH, 2015.

[8] V. Manohar, D. Povey and S. Khudanpur, “Semi-supervised Maximum Mutual Information Training of Deep Neural Network Acoustic Models”, In Proc. INTERSPEECH, 2015.
[9] M. Harper, “IARPA Babel Program,” 2014.
[10] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, “Multilingual acoustic models using distributed deep neural networks,” In Proc. IEEE-ICASSP, pp. 8619–8623, 2013.
[11] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” In Proc. IEEE-ICASSP, pp. 7304–7308, 2013.
[12] 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.
[13] L. Deng and J. Platt, "Ensemble deep learning for speech recognition." In Proc. INTERSPEECH, 2014.
[14] J. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER), ” in IEEE-ASRU, 1997.
[15] 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.
[16] J. Lafferty, A. McCallum, and F. Pereira. “Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ” In Proc. ICML, 2001.
[17] M. Seigel and P. Woodland, Combining Information Sources for Confidence Estimation with CRF Models,” In Proc. INTERSPEECH, 2011.
[18] 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.
[19] S. Li, Y. Akita, and T. Kawahara, “Corpus and transcription system of Chinese lecture room,” In Proc. ISCSLP, 2014.
[20] 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.
[21] 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,” In Proc. IEEE-ASRU. 2011.
[22] A. Lee and T. Kawahara. “Recent development of open-source speech recognition engine Julius.” In Proc. APSIPA ASC, pp.131-137, 2009.
[23] W. Chen, S.Ananathakrishnan, R.Kumar, R.Prasad, and P.Natarajan, “ASR error dection in a conversational spoken language translation system,” In Proc. IEEE-ICASSP, 2013.
[24] 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.
[25] 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.
[26] M. Mimura and T. Kawahara, “Fast Speaker Normalization and Adaptation Based on BIC for Meeting Speech Recognition,” IEICE TRANSACTIONS on Information and Systems, vol.J95-D No.7.2012.
[27] J. Nocedal. “Updating Quasi-Newton Matrices with Limited Storage”. Mathematics of Computation. 35. 151. 773-782. 1980.
[28] U. Fayyad and K. Irani, “Multi-interval discretization of continuous attributes for classification learning,” In Proc. IJCAI, pp1022-1027, 1993.
[29] J. Luo, L. Lamel and J-L. Gauvain,“Modeling Characters versus Words for Mandarin Speech Recognition.” In Proc. IEEE-ICASSP, Taipei, Taiwan, 2009.
[30] 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.
[31] M. Hwang, X. Lei, W. Wang, and T. Shinozaki, “Investigation on Mandarin broadcast news speech recognition,” In Proc. INTERSPEECH, 2006.

