Semi-Supervised Acoustic Model Training by Discriminative Data Selection From Multiple ASR Systems' Hypotheses

Sheng Li, Yuya Akita, *Member, IEEE*, and Tatsuya Kawahara, *Senior Member, IEEE*

***Abstract*—While the performance of ASR systems depends on the size of the training data, it is very costly to prepare accurate and faithful transcripts. In this paper, we investigate a semisupervised training scheme, which takes the advantage of huge quantities of unlabeled video lecture archive, particularly for the deep neural network (DNN) acoustic model. In the proposed method, we obtain ASR hypotheses by complementary GMM- and DNN-based ASR systems. Then, a set of CRF-based classifiers is trained to select the correct hypotheses and verify the selected data. The proposed hypothesis combination shows higher quality compared with the conventional system combination method (ROVER). Moreover, compared with the conventional data selection based on confidence measure score, our method is demonstrated more effective for filtering usable data. Significant improvement in the ASR accuracy is achieved over the baseline system and in comparison with the models trained with the conventional system combination and data selection methods.**

***Index Terms*—acoustic model, lecture transcription, semi-supervised training, Speech recognition.**

I. Introduction

AUTOMATIC speech recognition of spoken lectures has been investigated for almost a decade in many institutions world-wide [1]–[7], but there are still technically challenging issues for the system to be of practical use, including modeling of acoustic and pronunciation variations, speaker adaptation and topic adaptation. One of the biggest obstacles is the high expense to prepare accurate and faithful transcripts for spoken lectures (labeled data), since the performance of ASR systems depends on the size of the training data. In this work, we investigate a semi-supervised training scheme, which takes the advantage of huge quantities of unlabeled video lecture archive, particularly for the deep neural network (DNN) acoustic model.

Semi-supervised training combines a small set of labeled data with a large set of unlabeled data. The conventional paradigm of semi-supervised acoustic model training dealing with the unlabeled data includes preprocessing (e.g. speech segmentation, non-speech removal, speaker diarization, etc.), automatic transcription generation, data selection and model training. A number of studies have been conducted to address these processes [8]–[13]. However, they still do not solve the crucial part of automatic transcription generation and data selection. In this paper, we focus on these issues of the conventional paradigm of the semi-supervised training method.

For data selection, the most commonly used method is based on the confidence measure scores (CMS) computed by the ASR system [22]–[28], [60] with some post-processing or calibration [10], [30], [31], [67]. 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 search. 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 [29], 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 [30], [31].

We have applied the approach to the lightly supervised training [32] setting, where closed caption text is available and combined with an ASR hypothesis [33]. However, the assumption of the closed caption text limits the applicability of the method. In this work, we extend to the more general semi-supervised 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 [34]. Deng and Platt [35] mentioned 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 [36] 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 [63], [64], 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

Manuscript received January 06, 2016; revised April 28, 2016; accepted April 28, 2016. Date of publication May 03, 2016; date of current version June 21, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shinji Watanabe.

The authors are with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: lisheng@sap.ist.i.kyoto-u.ac.jp; akita@econ.kyoto-u.ac.jp; kawahara@i.kyoto-u.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2562505

selected result is reliable or not. Data selection for acoustic model training is conducted according to the verification result.

In the remainder of the paper, we first make a brief review on the semi-supervised training of DNN acoustic model in Section II. We describe the corpus of Chinese spoken lectures and the baseline ASR system in Section III. Next, the proposed method of semi-supervised training is formulated in Section IV. Then, the implementation of the method on the lecture transcription task is explained and experimental results are presented in Section V. The paper is concluded in Section VI.

II. Semi-Supervised Training of DNN Acoustic Model

Typical supervised training of DNN acoustic model [37] requires faithful labels for the fine-tuning, during which the pre-trained network [38], [39] is supervised-trained by the error back-propagation (BP) algorithm [40].

Semi-supervised training of DNN acoustic model [8]–[13] is developed, when the size of labeled training data is limited and huge quantities of unlabeled data on holding. It usually takes following steps:

1) transcribe unlabeled data with a seed model that was trained with the labeled data.
2) use the automatically generated transcript (ASR result) as a label.
3) retrain the model by adding the newly transcribed data to the existing labeled data.

However, taking use of the unlabeled data without data selection will make the model training less effective, because the DNN model training is more sensitive to the noise in the state label compared to the GMM model training, especially in sequence discriminative training [11], [41], [42].

Yu *et al.* [8] described the most commonly used data selection method, in which utterance-level CMS is adopted in semi-supervised training of GMM-based acoustic models from unlabeled data. We can sort the utterances by utterance-level CMS and select a certain percentage of top utterances to be used for model training.

For semi-supervised training of DNN-based acoustic models, the similar data filtering method has been used [9], [13], [10].

Liao *et al.* [9] showed that the high-confidence data are usually clustered like "island of confidence", by alternatively adopting binary word confidence scores. Applying an "island of confidence" filtering heuristic to select useful training segments, they achieved significantly improved performance for transcribing YouTube videos.

Zhang *et al.* [13] explored semi-supervised training of DNN in a meeting recognition task. They introduced improved DNN-based CMS estimators. Together with the error resolution, the CMS-based data selection achieved significant WER reduction.

Huang *et al*. [10] investigated semi-supervised GMM and DNN acoustic model training. They proposed a multi-system combination to improve the transcription accuracy and a confidence re-calibration approach to improve the data selection. Experiments showed significant improvement of retrained acoustic model on the mobile data.

Thomas *et al.* [19] selected the untranscribed data based on the utterance-level CMS, which was a log-linear combination of the ASR-based confidence and MLP posteriogram-based confidence. In their experiments, the method yielded a good result in a low-resource LVCSR setting.

In the fine-tuning step of DNN training, the gradients are used to update network parameters (of the weight matrix and bias) over frame-level mini-batches. It is possible to perform frame-level data selection, when we have frame-level CMS.

Vesely *et al.* [11] found it beneficial to conduct frame selection based on per-frame CMS derived from confusion network, as well as to reduce the disproportion in the amount of transcribed and untranscribed data by including the transcribed data several times in a low-resourced setting.

Imseng *et al.* [12] exploited untranscribed data of multiple European languages during semi-supervised DNN training. The resultant ASR system outperformed the baseline system trained with transcribed data only. They also revealed that CMS-based frame selection effectively reduced the size of the training data without degrading the ASR performance.

When DNN is regarded as a log-linear classifier (softmax output layer) upon a feature extractor (lower layers), unreliable data may help boost the training of lower layers, but is harmful for training the output softmax layer. Some recent studies [14], [15] introduced a multi-task training architecture for semi-supervised training without confidence filtering. In [16]–[18], [65], [66], multi-lingual training data share the same hidden layers but use different softmax layers for language-dependent senone classification. This architecture is used for semi-supervised training by viewing the transcribed and untranscribed data as different languages. After training, the softmax layer for unlabeled data is thrown away and only the softmax layer for labeled data is preserved.

In summary, the objective of these methods is to avoid the unfaithful label "polluting" the softmax layer of the network. In this paper, we focus on more effective data selection based on the above-mentioned methods. There are also other machine learning methods for semi-supervised training of acoustic model, e.g. graph-based method [20], submodular-based method [21] and data selection based on context-dependent state distribution [61] or global entropy reduction [62]. However, we will not discuss them in this paper.

III. Corpus and Baseline System

A. Data Preparation

We have compiled the Corpus of Chinese Lecture Room (CCLR) [43] as shown in Table I. While Chinese is one of the major languages for which ASR has been investigated, studies on Chinese lecture speech recognition are limited [44], [45], and a large-scale lecture corpus has not been made. We have designed and constructed CCLR based on the CCTV program of "Lecture Room" (百家講壇), 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

TABLE I
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

TABLE II
COMPONENT AND INTERPOLATED LMS

Language model	Corpora	#Words	PPlex.	Weight
Component LMs	CCLR	1.07 M	374	0.31
	HUB4	0.34 M	710	0.01
	TDT4	4.75 M	923	0.04
	GALE	1.03 M	426	0.16
	Phoenix	4.12 M	352	0.48
Interpolated LM	/	11.31 M	248	/

also provided by CCTV and available at the official website for a part of the lectures.

For the experimental purpose, we select 58 annotated lectures as the training set (CCLR-SV). In addition, 126 un-annotated but captioned lectures are used for lightly supervised training (CCLR-LSV) [33]. We use 19 annotated lectures as the test set (CCLR-TST). Additionally, 12 annotated lectures are held out as the development set (CCLR-DEV). 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 I.

B. Baseline ASR Systems

The dictionary for ASR consists of 53K lexical entries from CCLR-SV together with Hub4 and TDT4 distributed through LDC. The OOV rate on CCLR-TST is 0.368%. The pronunciation entries were derived from the CEDICT[1] open dictionary.

A word trigram language model (LM) was built for decoding. We complemented the small-sized text of CCLR-SV and CCLR-LSV with lecture texts collected from the web, whose size is 1.07M words. Then, this lecture corpus was interpolated with the corpora (Hub4 of 0.34M, TDT4 of 4.75M and GALE of 1.03M) and the Phoenix lecture archive (4.12M, text recordings of 1,300 broadcasted lectures from the Phoenix-HK official website[2]). The interpolated weights were determined to get the lowest perplexity on CCLR-DEV as shown in Table II.

We adopt 113 phonemes (consonants and 5-tone vowels) as the basic HMM unit. We first build a GMM-HMM system and then a DNN-HMM system.

The GMM system uses PLP features 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 and cepstral variance normalization are applied to the features. It is trained with the MPE criterion. Moreover, we conduct unsupervised speaker adaptation using MLLR for each lecture, which is effective for long lecture speech.

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, and has 1320 nodes as input, 3000 nodes as output and 7 hidden layers with 1024 nodes per layer. The activation function is sigmoidal function. Training of DNN consists of the unsupervised pre-training step and the supervised fine-tuning step. We use Kaldi toolkit (nnet1) [46], which implements SGD to minimize the cross-entropy between the supervision labels and network output. The SGD uses mini-batches of 256 frames, and a default "Newbob" learning rate schedule which starts with an initial learning rate of 0.008 and halves the rate when the improvement in frame accuracy on a cross-validation set between two successive epochs falls below 0.5%. The training terminates when the frame accuracy increases by less than 0.1%. The cross-validation set is held out from the training data by 10%. To accelerate the training time, we use single GPU (Tesla K20m). On this stage, the training is based with the CE criterion, and sequential discriminative training is not conducted. For decoding, we use Julius ver.4.3.1 (DNN version[3]) [47] using the state transition probabilities of the GMM-HMM.

Since the data size of CCLR-SV is not large enough to train a baseline lecture transcription system, we introduced a lightly-supervised training method [33] to enhance the model training by exploiting usable data in another large data set CCLR-LSV with closed caption texts.

This baseline system achieved an average Character Error Rate (CER) of 24.2% and 27.5% with the MLLR speaker adapted GMM-HMM model, and 22.7% and 25.7% with the DNN-HMM model for CCLR-DEV and CCLR-TST, respectively.

Hypothesis combination requires a set of complimentary ASR systems with similar accuracy and enough diversity on recognition patterns [34]. We trained two other DNN systems with the different feature types. One uses 13-dimensional MFCC features (with the first and second derivatives) and the other uses 13-dimensional PLP features (with the first and second derivatives). For these complementary systems, we calibrated their CMS before ROVER-based system combination by using a four-system committee-based recalibration algorithm [10]. The ASR performance (CER%) is listed in Table III. The pairwise edit distances of these systems are listed in Table IV. The largest diversity exists between GMM and DNN systems with similar accuracy (difference on the CER% less than 2% in Table III) as mentioned in [35], and their ROVER result outperforms other two-system ROVER combinations and also the four-system ROVER combination.

[1] Available at http://cc-cedict.org/wiki/

[2] Available at http://v.ifeng.com/gongkaike/sjdjiangtang/

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

TABLE III
ASR PERFORMANCE OF SINGLE SYSTEM AND ROVER COMBINATION ON CCLR-DEV

	Complementary Systems for ROVER				ASR Performance
	DNN (fbank)	DNN (MFCC)	DNN (PLP)	GMM (MPE + MLLR)	CER (%)
1-System	√				22.7
		√			23.5
			√		24.0
				√	24.2
2-system	√	√			21.8
	√		√		21.8
	√			√	**20.8**
		√	√		22.9
		√		√	21.7
			√	√	21.7
4-system	√	√	√	√	21.1

TABLE IV
PAIR-WISE EDIT DISTANCE OF ASR RESULTS ON CCLR-DEV (CHARACTER LEVEL)

	GMM (MPE+MLLR)	DNN (PLP)	DNN (MFCC)	DNN (fbank)
GMM (MPE+MLLR)	/	/	/	/
DNN (PLP)	**24.5%**	/	/	/
DNN (MFCC)	**24.6%**	14.7%	/	/
DNN (fbank)	**24.3%**	17.3%	16.1%	/

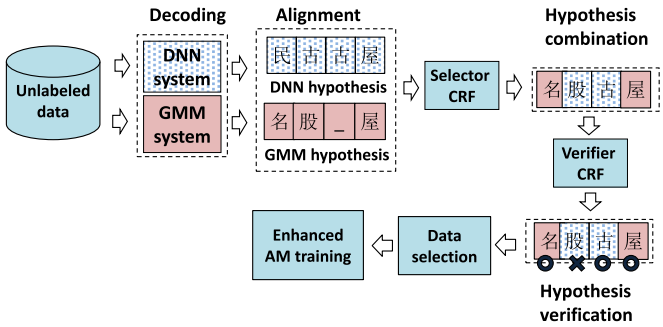

Fig. 1. Flowchart of the proposed method.

Conveniently, we can reuse the GMM-HMM system that is produced in the process of the DNN-HMM (filterbank feature) acoustic model training as a complementary system.

IV. CRF-BASED HYPOTHESIS COMBINATION AND DATA SELECTION

We propose an effective system combination and data selection method with CRF-based classifiers as shown in Fig. 1. The process flow is as follows.

A. Process Flow

1) Preprocessing and Hypothesis Generation: For preprocessing, we first conduct speech segmentation to the utterance unit based on the Bayesian information criterion method [48] and speaker clustering to remove non-speech segments and speech from other than the main lecturer. Then the unlabeled data in CCLR-USV is decoded by the DNN system and the speaker-adapted GMM system, respectively.

2) Hypothesis 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 [49]–[51]. 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 classification in the utterance level. The DNN system is retrained by adding the selected data.

TABLE V
CATEGORY OF ALIGNMENT PATTERNS

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

(√ means matching with reference, means mismatching)

B. Categories 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-based system and the DNN-based system and also the faithful transcripts (reference). By analyzing the aligned character sequence, we can categorize patterns into five classes, as listed in Table V. The insertion and deletion cases are handled by using a null token.

The definitions of the categories are as follows:

1) ***C1***: the DNN hypothesis is matched with the GMM hypothesis and also the correct transcript. A majority of the samples falls in this category.
2) ***C2***: although the DNN hypothesis is matched with the GMM hypothesis, neither of them is correct. This case is rare.
3) ***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.

C. Classifier Design

We use CRF [29] as the classifier for this task. It can model the relationship between the features and labels by considering

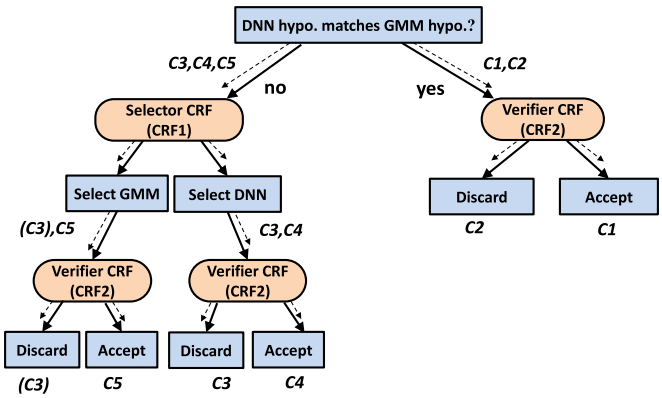

Fig. 2. Cascaded classification scheme for data selection.

sequential dependencies of contextual information. For this reason, it is used for many applications such as confidence measuring [30], [31], ASR error detection [52], and automatic narrative retelling assessment [53].

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 (see Table V). 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 character 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. For general purpose, this classifier is the same as the one used for ***C1*** and ***C2***. We do not have enough training samples to train individual classifiers.

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 character 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 makes the data distribution more balanced. Erroneous patterns in ***C3*** (i.e. GMM hypothesis is incorrect) will be rejected by the verifier CRF (**CRF-2**).

D. Feature Design

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

TABLE VI
FEATURE DESIGN 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 VII
FEATURE DESIGN FOR CRF-2

Feature Type	Features
ASR-based feature	1. Confidence measure score of baseline 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 the word unit, and distributed to each character in the word. They are numeric features:

1) The **CMS** is output by the Julius decoder [25] of the baseline ASR system. The value is between [0, 1] approximating a posterior probability of the hypothesis word.
2) The word duration (**DUR**) feature is the number of frames of the word.
3) The word trigram LM (**WLM**) feature is the word trigram LM score of the word while decoding.
4) Averaged acoustic model score (**AM**) feature is the acoustic likelihood score averaged for each frame.
5) 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.
6) 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.
7) 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:

1) The lexical feature (**LEX**) is a lexical entry (ID) of the current character. It is a symbolic feature.
2) 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 [54]. This feature is symbolic.
3) The LM probability feature (**CLM**) is a negative log probability of the current character rescored by a

character 5-gram LM. 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**), it is difficult to use the ASR-based features for the selected hypothesis, because the features from two different ASR systems have different dynamic ranges [55], [56]. We also recalculate the text-based features after classification by the selector CRF (**CRF-1**) because of the context change. Additional feature we use is the posterior probability output of **CRF-1** (for the mismatching cases) and the CMS of the DNN system (for the matching cases) as shown in Table VII.

E. Data Selection for Acoustic Model Training

The ASR hypotheses are merged into a single character sequence after the matching and selection processes, 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 should be used for acoustic model training. Two kinds of data selection scheme are investigated as follows:

1) Utterance-Level Selection: The most commonly used utterance-level selection is based on utterance-level CMS, which is formulated as follows:

$$C_{\text{sent}} = \frac{1}{N}\sum_{i=1}^{N} C_{wi}$$

where C_{wi} is the CMS of word w_{i} obtained by confusion network decoding [26] and N is the number of words in the utterance.

Then we can sort the utterances by utterance-level CMS and select a certain percentage of top utterances for model training.

In our proposed method, we compute the character acceptance rate (CA) for every utterance. Since Chinese is a syllabic language and each character is a syllable, the "CA" actually means the ratio of "accepted" syllables over the total number of syllables in an utterance.

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 tolerate some character errors existing in the utterances and select these utterances with their CA no lower than 70%.

2) Frame-Level Selection: We also implement frame-level data selection based on frame dropping and multi-task training methods.

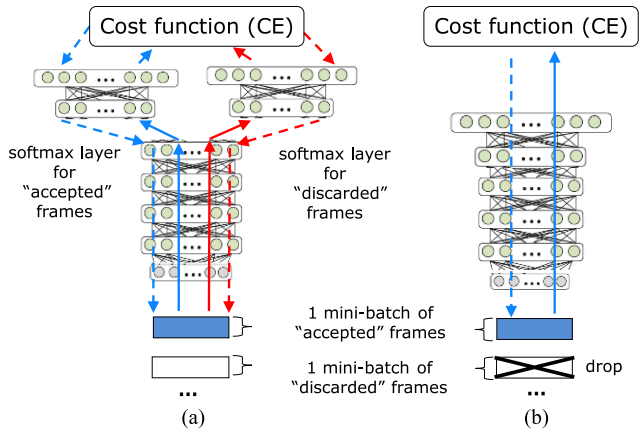

Fig. 3. Frame-level data selection methods.

We determine acceptance of each frame, so the parameters of DNN are updated on the selected 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. With the frame-level selection, we can train DNN model by either multi-task training method shown in Fig. 3(a) or frame dropping method shown in Fig. 3(b).

We make each mini-batch (256 frames) consisting of either "accepted" frames or "discarded" frames, and then shuffle all of the mini-batches. In the multi-task training method, the "accepted" mini-batches and the "discarded" mini-batches update the shared hidden layers but update different softmax layers. And we only preserve the softmax layer for "accepted" frames after training. In the frame dropping method, we only use the "accepted" mini-batches to update the whole network.

V. Experimental Evaluations

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

A. Classifier Implementations

1) Training and Testing Data for Classifiers: In our implementation, we train CRF classifiers using CCLR-SV: **CRF-1**, which is trained to discriminate ***C3*** + ***C5*** versus ***C4***, and **CRF-2**, which is trained to verify the output of **CRF-1** ($C4 + C5$ vesus ***C3***) and to discriminate ***C1*** versus ***C2***.

Since the feature of **CRF-2** depends on the result of **CRF-1**, we 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 be applied to on the rest 1/5 data.

2) Training Data Resampling: 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 V. 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 training of the classifiers. Thus, we introduce a re-sampling

TABLE VIII
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**

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%.

3) Incorporating Data From Captioned Data: For improved training, we also incorporate data from CCLR-LSV to enlarge the training data. This process is not straightforward, because we only have closed caption texts instead of faithful transcripts.

We made a three-way character alignment among the two ASR hypotheses by the GMM-based system and the DNN-based system and also the closed caption texts. We regard the all-matching cases as positive samples and the all-mismatching cases as negative samples, and add them to the training data of **CRF-2**.

4) Training Settings of CRF Classifiers: In the experiment, we use a linear-chain CRF implemented in the CRFSuite package.[4] The standard Limited-memory BFGS (L-BFGS) [57] algorithm and L2 regularization are used to train the CRF models with the sparse features of a high dimension. The 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[5] described in [58]. The same kind of numeric features from the DNN and GMM based systems can have different quantization levels.

B. Classification Accuracy of CRF Classifiers

Classification performance with various feature sets is evaluated on CCLR-DEV, as shown in Tables VIII and IX. Performance is measured by precision, recall and F-score:

$$\text{Precision} = TP/FP$$

$$\text{Recall} = TP/(FP + FN)$$

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

TABLE IX
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**

$$F - \text{score} = 2 \times \text{Precision} \times \text{Recall}/(\text{Precision} + \text{Recall})$$

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

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

Among the feature sets, the text-based features and their combinations are generally less effective than the ASR-based feature in **CRF-1** and **CRF-2**. However, 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.

C. Performance of Hypothesis Selection and Verification

Next, we evaluate the performance of selection and verification of ASR hypotheses using CCLR-DEV and CCLR-TST.

The GMM-HMM and DNN-HMM baseline systems are described in Section III. Other methods compared with the proposed method are as follows:

1) **Combine-ROVER:** the hypothesis and CMS derived from the ROVER-based system combination (the conventional method).
2) **Combine-single-CRF:** we trained a five-class CRF model to combine the ASR hypothesis.
3) **Combine-CRFs:** we train two classifiers for system combination (the proposed method). We will test different stages of our proposed cascade classification: **Combine-CRFs (CRF-1)** for evaluating the effectiveness of the selection process only and **Combine-CRFs (CRF-1+CRF-2)** to evaluate the effectiveness of the verification process.

We use following metrics for evaluation:

1) **CER**: ASR evaluation measure after the hypothesis combination.
2) **Normalized Cross Entropy (NCE):** It assigns the information gain to each of the hypothesis word to evaluate the quality of the confidence score distribution [59]. Higher

TABLE X
EVALUATION OF THE DATA SELECTION AND VERIFICATION

	CCLR-DEV			CCLR-TST		
	CER (%)	NCE	EER (%)	CER (%)	NCE	EER (%)
GMM-HMM (MPE + MLLR)	24.2	0.30	18.3	27.5	0.30	18.6
DNN-HMM (fbank)	22.7	0.32	20.7	25.7	0.28	21.8
Combine-ROVER	20.8	0.26	22.7	24.5	0.26	23.3
single-CRF	21.9	0.28	21.7	25.7	0.25	22.8
Combine-CRFs (CRF-1)	**20.5**	0.28	18.2	**24.0**	0.25	19.3
Combine-CRFs (CRF-1+CRF-2)	**20.5**	**0.37**	**17.1**	**24.0**	**0.34**	**18.5**

values of NCE indicate better ASR confidence estimation. Perfect ASR confidence estimates give an NCE of 1. The definition of NCE is as follows:

$$\mathrm{NCE} = \left\{ H_{\max} + \sum_{\text{correct}} \log_2 \left(\hat{p}\left(w\right)\right) + \sum_{\text{incorrect}} \log_2 \left(1-\hat{p}\left(w\right)\right) \right\} / H_{\max}$$

$$H_{\max} = -n\log_2\left(p_c\right) - \left(N-n\right)\log_2\left(1-p_c\right)$$

where n is the number of correct hypothesis words, N is the total number of hypothesis words, p_c is the average probability that an output word is correct $(= n/N)$, $\hat{p}(w)$ is the confidence measure output of output word w.

3) **Equal Error Rate (EER):** the false alarm rate or the miss rate at the confidence score threshold where the false alarm and miss rate get equal. Lower values of EER indicate better ASR confidence estimation. Perfect ASR confidence estimates give an EER of 0.

The results are listed in Table X. The proposed method **Combine-CRFs** outperforms the other methods. We observed that combination of hypotheses by ROVER method (**Combine-ROVER**) can effectively reduce the recognition error rate (around absolute 2%) from the best single system (**DNN-HMM**), but it does not improve the confidence estimation. Using a single CRF classifier (**single-CRF**) can improve the confidence estimation, but it does not lead to the reduction of the recognition error rate. Our proposed method (**Combine-CRFs**) shows robustness to the small number of complementary systems and different distributions of CMS between the DNN-based system and GMM-based system. The **CRF-1** improves the recognition result of the ROVER method (around absolute $0.3\% \sim 0.5\%$). Note that iROVER [63], [64] is similar to the case using only **CRF-1**. Moreover, **CRF-2** further improves the confidence estimation quality based on the **CRF-1** classification result.

D. ASR Performance of DNN Acoustic Model Enhanced by Selected Data

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:

1) **Baseline GMM** and **baseline DNN**: the models are trained by only using CCLR-SV and CCLR-LSV as described in Section III.
2) **DNN (CMS)**: we select utterances from CCLR-USV using the baseline DNN system based on a threshold of averaged CMS score (CMS $\geq$ 0.6). The optimal threshold was determined by using GMM (MLE) models and CCLR-DEV [33]. It is the most commonly used method. We also use all of the ASR hypotheses of CCLR-USV from the DNN based system without any selection (CMS $\geq$ 0.0).
3) **Combine-ROVER**: combine the ASR hypotheses of CCLR-USV from the baseline GMM and the baseline DNN systems using ROVER [36]. We select utterances according to the optimal threshold of the averaged CMS score (CMS $\geq$ 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 $\geq$ 0.0). We derive the hypothesis and CMS from the ROVER-based system combination.
4) **Combine-CRFs (CA = 1.0, CA $\geq$ 0.0 and CA $\geq$ 0.7)**: 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 $\geq$ 0.0 (no selection), CA = 1.0 (use utterances with all characters accepted), and CA $\geq$ 0.7.

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

TABLE XI
ASR PERFORMANCE (CER%) OF CROSS-ENTROPY DNN MODEL BY UTTERANCE-LEVEL SELECTION

	Amount of data (hours)		CER%	
	labeled	unlabeled	DEV	TST
Baseline GMM (MPE + MLLR)	97.2	0	24.2	27.5
Baseline DNN (fbank)	97.2	0	22.7	25.7
DNN (CMS $\geq$ 0.0)	97.2	114.7	22.3	25.4
DNN (CMS $\geq$ 0.6)	97.2	83.9	22.0	25.1
Combine-ROVER (CMS $\geq$ 0.0)	97.2	114.7	22.0	24.9
Combine-ROVER (CMS $\geq$ 0.6)	97.2	68.7	21.9	24.9
Combine-CRFs (CA $\geq$ 0.0)	97.2	114.7	21.5	24.4
Combine-CRFs (CA = 1.0)	97.2	32.5	21.6	24.7
Combine-CRFs (CA $\geq$ 0.7)	**97.2**	**71.5**	**21.3**	**24.2**

ASR performance in CER is listed in Table XI. The results show that our proposed semi-supervised training method significantly improved the baseline DNN system. It also outperforms all other methods on both evaluation data sets.

We observe that both of Combine-CRFs and Combine-ROVER outperform the simple CMS-based selection DNN

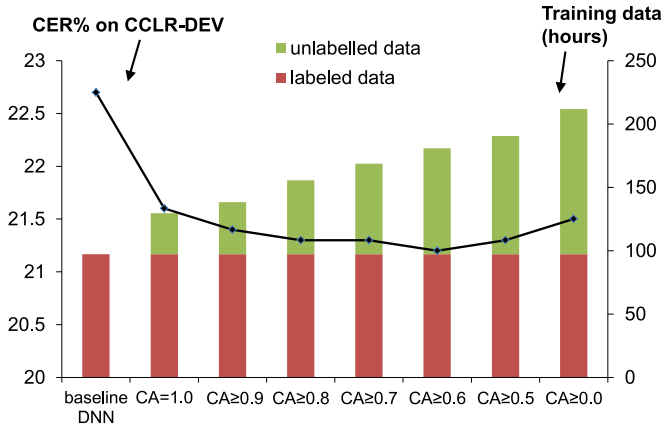

Fig. 4. Training data amount and resulting accuracy on CCLR-DEV.

TABLE XII
ASR PERFORMANCE (CER%) OF CROSS-ENTROPY DNN MODEL BY FRAME-LEVEL SELECTION

	Amount of data (hours)		CER%	
	labeled	unlabeled	DEV	TST
Combine-CRFs (CA ≥ 0.7)	97.2	71.5	21.3	24.2
Combine-CRFs (multi-task)	97.2	114.7	21.3	24.3
Combine-CRFs (drop-frames)	97.2	90.4	21.4	24.3

(CMS ≥ 0.0 and CMS ≥ 0.6). This suggests the system combination effectively leverages the quality of automatically generated transcription. The fact that our proposed method Combine-CRFs (CA ≥ 0.0) further outperforms the Combine- ROVER (CM ≥ 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.

It is observed during the training that the proposed method (CA ≥ 0.7) results in better CE and frame accuracy than other methods (DNN (CMS ≥ 0.6) and ROVER (CMS ≥ 0.6)).

We also conducted the proposed method with different threshold values (0.5 ∼ 1.0) to show the relationship of the training data amount and the resulting model accuracy on the CCLR-DEV set in Fig. 4. We can see there is no significant difference in the range of 0.5 to 0.8.

Finally, we conduct the frame-level verification result as described in Section IV-E2, where "accepted" frames are used for supervised learning. We implement the frame dropping and the multi-task training methods. We refer these two different methods to Combine-CRFs (multi-task) and Combine-CRFs (drop-frames) respectively. Their ASR performance shows no significant difference compared with Combine-CRFs (CA ≥ 0.7) in Table XII. However, the frame-level selection methods do not require any threshold tuning.

TABLE XIII
ASR PERFORMANCE (CER%) OF SMBR DNN MODEL

	Amount of data (hours) for sMBR training		CER%	
Seed CE DNN model	labeled	unlabeled	DEV	TST
Baseline DNN (fbank)	97.2	0	21.9	24.7
Combine-CRFs (CA ≥ 0.0)	97.2	114.7	20.9	23.3
Combine-CRFs (CA = 1.0)	97.2	32.5	21.0	23.6
Combine-CRFs (CA ≥ 0.7)	**97.2**	**71.5**	**20.7**	**23.1**
Combine-CRFs (CA ≥ 0.7)	**97.2**	**71.5 (CE only)**	**20.3**	**23.0**

On the other hand, utterance-level selection is advantageous for conducting sequence discriminative training. We train the sMBR DNN models by using four Cross-Entropy (CE) DNN models as the seed model listed in Table XI: Baseline DNN (fbank), Combine-CRFs (CA ≥ 0.7, CA ≥ 0.0 and CA = 1.0). Their ASR performance is shown in Table XIII. We can see a significant improvement by Combine-CRFs (CA ≥ 0.7) over the other three models (Baseline DNN, CMS ≥ 0.0 and CA = 1.0). The effectiveness of the proposed method is still maintained after sMBR training. That means our data selection method also works for sequence discriminative DNN training. However, the sMBR training based on Combine-CRFs (CA ≥ 0.7) CE model using the labeled data only shows further improvement. The result suggests that the sMBR training is sensitive to errors in the label.

VI. CONCLUSION

We have proposed a new method for hypothesis leveraging and data selection for semi-supervised 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 semi-supervised training method effectively filters usable data, and improves the ASR accuracy from the baseline model and in comparison with the conventional ROVER-based method and the CMS-based selection method.

Since our baseline systems have large room for improvement, we will investigate combining different types of deep learning based acoustic model. We also hope we can investigate with a larger data set. The proposed method does not depend on the character/syllable level modeling. Therefore, it will hopefully be ported to other languages such as English.

REFERENCES

[1] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. IEEE ISCA IEEE Workshop Spontaneous Speech Process. Recog.*, 2003, pp. 7–12.

[2] H. Nanjo and T. Kawahara, "Language model and speaking rate adaptation for spontaneous presentation speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 391–400, Jul. 2004.

[3] I. Trancoso, R. Nunes, L. Neves, C. Viana, H. Moniz, D. Caseiro, and A. I. Mata, "Recognition of classroom lectures in European Portuguese," in *Proc. INTERSPEECH*, 2006, pp. 281–284.

[4] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the MIT spoken lecture processing project," in *Proc. INTERSPEECH*, 2007, pp. 2553–2556.

[5] H. Yamazaki, K. Iwano, K. Shinoda, S. Furui, and H. Yokota, "Dynamic language model adaptation using presentation slides for lecture speech recognition," in *Proc. INTERSPEECH*, 2007, pp. 2349–2352.

[6] T. Kawahara, Y. Nemoto, and Y. Akita, "Automatic lecture transcription by exploiting slide information for language model adaptation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4929–4932.

[7] M. Paul, M. Federico, and S. Stucker, "Overview of the IWSLT 2010," in *Evaluation Campaign. Proc. Int. Workshop Spoken Language Translation*, 2010, pp. 3–27.

[8] K. Yu, M. Gales, L. Wang, and P. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Commun.*, vol. 52 no. 7, pp. 652–663, 2010.

[9] 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. Automat. Speech Recog. Understanding*, 2013, pp. 368–373.

[10] 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*, 2013, pp. 2360–2364.

[11] K. Vesely, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *Proc. Automat. Speech Recog. Understanding*, 2013, pp. 267–272.

[12] D. Imseng, B. Potard, P. Motlicek, A. Nanchen, and H. Bourlard, "Exploiting un-transcribed foreign data for speech recognition in well-resourced languages," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 2322–2326.

[13] P. Zhang, Y. Liu, and T. Hain, "Semi-supervised DNN training in meeting recognition," in *Proc. IEEE Spoken Language Technol. Workshop*, 2014, pp. 141–146.

[14] H. Su and H. Xu, "Multi-softmax deep neural network for semi-supervised training," presented at the INTERSPEECH, Dresden, Germany, 2015.

[15] V. Manohar, D. Povey, and S. Khudanpur, "Semi-supervised maximum mutual information training of deep neural network acoustic models," presented at the INTERSPEECH, Dresden, Germany, 2015.

[16] M. Harper, "IARPA babel program," 2014.

[17] 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. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8619–8623.

[18] 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. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7304–7308.

[19] S. Thomas, M. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 6704–6708.

[20] Y. Liu and K. Kirchhoff, "Graph-based semi-supervised acoustic modeling in DNN based speech recognition," in *Proc. IEEE Workshop Spoken Language Technol.*, USA, 2014, pp. 177–182.

[21] K. Wei, Y. Liu, K. Kirchhoff, C. Bartels, and J. Bilmes. "Submodular subset selection for large-scale speech training data," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Italy, 2014, pp. 3311–3315.

[22] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Commun.*, vol. 45, no. 4, pp. 455–470, Apr. 2005.

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

[24] 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, Mar. 2001.

[25] A. Lee, K. Shikano, and T. Kawahara, "Real-time word confidence scoring using local posterior probabilities on tree trellis search," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 1, pp.793–796.

[26] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," in *Proc. EUROSPEECH*, Budapest, Hungary, 1999, pp. 495–498.

[27] H. Xu, D. Povey, L. Mangu, and J. Zhu, "An improved consensus-like method for Minimum Bayes Risk decoding and lattice combination," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4938–4941.

[28] V. Goel and W. J. Byrne, "Minimum Bayes-risk automatic speech recognition," *Comput. Speech Language*, vol. 14, no. 2, pp. 115–135, 2000.

[29] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

[30] M. Seigel and P. Woodland, "Combining information sources for confidence estimation with CRF models," presented at the INTERSPEECH, Florence, Italy, 2011.

[31] J. Fayolle, F. Moreau, C. Raymond, and G. Gravier, "CRF-based combination of contextual features to improve a posteriori word level confidence measures," presented at the INTERSPEECH, Makuhari, Japan, 2010.

[32] L. Lamel, J. L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Comput. Speech Language*, vol. 16, pp. 115–129, Jan. 2002.

[33] S. Li, Y. Akita, and T. Kawahara, "Discriminative data selection for lightly supervised training of acoustic model using closed caption texts," in *Proc. INTERSPEECH*, 2015, pp. 3526–3530.

[34] K. Audhkhasi, A. Zavou, P. Georgiou, and S. Narayanan, "Theoretical analysis of diversity in an ensemble of automatic speech recognition systems," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 711–726, Mar. 2014.

[35] L. Deng and J. Platt, "Ensemble deep learning for speech recognition," presented at the INTERSPEECH, Singapore, 2014.

[36] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. IEEE Workshop Automat. Speech Recog. Understanding*, 1997, pp. 347–354.

[37] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no.1, pp. 30–42, Jan. 2012.

[38] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, 2006.

[39] G. Hinton, "A practical guide to training restricted Boltzmann machines," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. UTML TR 2010-003, 2010.

[40] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.

[41] H. Su, G. Li, D. Yu, and F. Seide, "Error back propagation for sequence training of context-dependent deep neural networks for conversational speech transcription," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 6664–6668.

[42] K. Vesely, A. Ghoshal, L. Burget and D. Povey, "Sequence-discriminative training of deep neural networks," presented at the INTERSPEECH, Lyon, France, 2013.

[43] S. Li, Y. Akita, and T. Kawahara, "Corpus and transcription system of Chinese lecture room," in *Proc. Int. Symp. Chin. Spoken Language Process.*, 2014, pp. 2016442–445.

[44] J. Zhang, H. Chan, P. Fung, and L. Cao, "A comparative study on speech summarization of broadcast news and lecture speech," in *Proc. INTERSPEECH*, 2007, pp. 2781–2784.

[45] 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*, 2009, pp. 4709–4712.

[46] D. Povey *et al.*, "The Kaldi speech recognition toolkit," presented at the IEEE Automatic Speech Recognition Understanding, Big Island, HI, USA, 2011.

[47] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," in *Proc. Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit Conf.*, 2009, pp. 131–137.

[48] M. Mimura and T. Kawahara, "Fast speaker normalization and adaptation based on BIC for meeting speech recognition," presented at the Asia-Pacific Signal Information Processing Association Annual Summit Conf., Xi'an, China, 2011.

[49] J. Luo, L. Lamel, and J.-L. Gauvain, "Modeling characters versus words for mandarin speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, 2009, pp. 4325–4328.

[50] X. Liu, J. L. Hieronymus, M. J. F. Gales, and P. C. Woodland, "Syllable language models for mandarin speech recognition: Exploiting character sequence models," *J. Acoust. Soc. Amer.*, vol. 133, no. 1, pp. 519–528, Jan. 2013.

[51] M. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on mandarin broadcast news speech recognition," presented at the INTERSPEECH, Pittsburgh, PA, USA, 2006.
[52] W. Chen, S. Ananathakrishnan, R. Kumar, R. Prasad, and P. Natarajan, "ASR error detection in a conversational spoken language translation system," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7418–7422.
[53] 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. North Amer. Chapter Assoc. Comput. Linguistics*, 2013, pp. 211–220.
[54] M. Shen, H. Liu, D. Kawahara, and S. Kurohashi, "Chinese morphological analysis with character-level POS tagging," in *Proc. 52th Annu. Meet. Assoc. Comput. Linguistics., Short Paper*, Baltimore, MD, USA, 2014, pp. 253–258.
[55] Y. C. Tam, Y. Lei, J. Zheng, and W. Wang, "ASR error detection using recurrent neural network language model and complementary ASR," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 2331–2335.
[56] J. Kim, J. Chong, and I. Lane, "Efficient on-the-fly hypothesis rescoring in a hybrid GPU/CPU-based large vocabulary continuous speech recognition engine," in *Proc. INTERSPEECH*, 2012, pp. 1035–1038.
[57] J. Nocedal, "Updating quasi-Newton matrices with limited storage," *Math. Comput.*, vol. 35, no. 151, pp. 773–782, 1980.
[58] U. Fayyad and K. Irani, "Multi-interval discretization of continuous attributes for classification learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 1993, pp. 1022–1027.
[59] M. Siu, H. Gish, and F. Richardson, "Improved estimation, evaluation and applications of confidence measures for speech recognition," in *Proc. EUROSPEECH*, 1997, pp. 831–834.
[60] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," presented at the NIST Speech Transcription Workshop, College Park, MD, USA, 2000.
[61] O. Siohan, "Training data selection based on context-dependent state matching," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 3316–3319.
[62] 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," *Comput. Speech Language*, vol. 24, no. 3, pp. 433–444, 2010.
[63] 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.
[64] 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.
[65] S. Scanzio, P. Laface, L. Fissore, R. Gemello, and F. Mana, "On the use of a multilingual neural network front-end," presented at the INTERSPEECH, Brisbane, QLD, Australia, 2008.
[66] K. Vesely, M. Karafiát, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proc. IEEE Spoken Language Technol. Workshop*, 2012, pp. 336–341.
[67] D. Yu, J. Li, and L. Deng, "Calibration of confidence measures in speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2461–2473, Nov. 2011.

Sheng Li received the B.S. degree in 2006 and the M.E. degree in 2009 from the Computer Science and Software Institution, Nanjing University (National Central University, before 1949), Nanjing, China, and the Ph.D. degree in information science from Kyoto University, Kyoto, Japan, in 2016. From 2009 to 2012, he served for the Chinese Academic of Sciences, doing research on LVCSR, CALL, and multimodal speech synthesis. He is currently a Researcher at Kyoto University working on acoustic modeling for speech recognition.

Yuya Akita received the B.E., M.Sc., and Ph.D. degrees in 2000, 2002, and 2005, respectively, from Kyoto University. From 2005 to 2015, he was an Assistant Professor at Academic Center for Computing and Media Studies, Kyoto University. He is currently a Senior Lecturer at the Graduate School of Economics, Kyoto University.

His research interests include spontaneous speech recognition and spoken language processing. He is a member of IEICE, IPSJ, and ASJ. He received the Awaya Memorial Award from ASJ in 2007, the Yamashita SIG Research Award from IPSJ in 2010, the Commendation for Science and Technology by the Minister of MEXT, and the Kiyasu Special Industrial Achievement Award from IPSJ, both in 2012.

Tatsuya Kawahara received the B.E., M.E., and Ph.D. degrees in 1987, 1989, and 1995, all in information science, from Kyoto University, Kyoto, Japan. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA.

He is currently a Professor at the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR and NICT.

He has published more than 300 technical papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several speech-related projects in Japan including free large vocabulary continuous speech recognition software (http://julius.sourceforge.jp/) and the automatic transcription system for the Japanese Parliament (Diet).

He received the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2012. From 2003 to 2006, he was a member of the IEEE SPS Speech Technical Committee. He was a General Chair of the IEEE Automatic Speech Recognition & Understanding Workshop (ASRU 2007). He also served as a Tutorial Chair of INTERSPEECH 2010 and a Local Arrangement Chair of ICASSP 2012. He is an Editorial Board Member of the Elsevier *Journal of Computer Speech and Language*, *APSIPA Transactions on Signal and Information*, and the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He is VP-Publications (BoG member) of APSIPA.