PAPER

Automatic Lecture Transcription Based on Discriminative Data Selection for Lightly Supervised Acoustic Model Training

Sheng LI[†a)], ***Nonmember***, **Yuya AKITA**[†], ***and*** **Tatsuya KAWAHARA**[†], ***Members***

SUMMARY The paper addresses a scheme of lightly supervised training of an acoustic model, which exploits a large amount of data with closed caption texts but not faithful transcripts. In the proposed scheme, a sequence of the closed caption text and that of the ASR hypothesis by the baseline system are aligned. Then, a set of dedicated classifiers is designed and trained to select the correct one among them or reject both. It is demonstrated that the classifiers can effectively filter the usable data for acoustic model training. The scheme realizes automatic training of the acoustic model with an increased amount of data. A significant improvement in the ASR accuracy is achieved from the baseline system and also in comparison with the conventional method of lightly supervised training based on simple matching.

key words: *speech recognition, acoustic model, lightly supervised training, lecture transcription*

1. Introduction

Automatic transcription of lectures is one of the promising applications of automatic speech recognition (ASR), since many courses of audio and video lectures are being digitally archived and broadcasted. Captions to the lectures are needed not only for hearing-impaired persons but also for non-native viewers and elderly people. ASR is also useful for indexing the content.

ASR of 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 reach a practical level, including modeling of acoustic and pronunciation variations, speaker adaptation and topic adaptation. In this work, we address effective acoustic model training targeted on Chinese spoken lectures.

There is a large amount of audio and video data of lectures, but it is very costly to prepare accurate and faithful transcripts for spoken lectures, which are necessary for training acoustic and language models. We observed that, even given a caption text, a lot of work is needed to make a faithful transcript because the caption text is much different from what is actually spoken, and phenomena of spontaneous speech such as fillers and repairs need to be included.

In order to increase the training data for an acoustic model, a scheme of lightly supervised training, which does not require faithful transcripts but exploits available verbatim texts, has been explored for broadcast news [10]–[12] and parliamentary meetings [13]. In the case of parliamentary meetings, verbatim texts are made by stenographers, and thus can be used to predict faithful transcripts. However, in the case of TV programs, closed caption texts are not so verbatim because of the space constraint, and thus can be used in an indirect manner for lightly supervised training. A typical method [10], [11] consists of two steps. In the first step, a biased language model is constructed based on the closed caption text of the relevant program to guide the baseline ASR system to decode the audio content. The second step is to filter the reliable segments of the ASR output, usually by matching it against the closed caption. In this simple method, only matched segments are selected.

The conventional filtering method, however, has a drawback that it significantly reduces the amount of usable training data. Moreover, it is presumed that the unmatched or less confident segments of the data are more useful than the matched segments because the baseline system failed to recognize them and may be improved with additional training [12]. Recent work by Long et al. [14] proposed methods to improve the filtering by considering the phone error rate and confidence measures. Other studies, e.g. [15], introduced an improved alignment method for lightly supervised training.

Instead of simple sequence matching [10]–[12] and heuristic measure-based selection [14], [15], in this work, we propose to train a set of dedicated classifiers to select the usable data for acoustic model training. Given an aligned sequence of the ASR hypothesis and the closed caption text (and also reference text in the training phase), a set of classifiers is trained based on a discriminative model to select between the ASR result and the closed caption text, or reject both if they are not matched. It is trained with a database of a relatively small size used for training the baseline acoustic model and applied to a large-scale database that has closed caption texts but not faithful transcripts.

In the remainder of the paper, we first describe the corpus of Chinese spoken lectures and the baseline ASR system in Sect. 2. Next, our proposed scheme of classifier design for lightly supervised training is formulated in Sect. 3. Then, the implementation of the method on the lecture transcription task is explained and experimental results are presented in Sect. 4. The paper is concluded in Sect. 5.

2. Corpus and Baseline ASR System

For a comprehensive study on ASR of spontaneous Chinese

Manuscript received February 6, 2015.
Manuscript revised March 13, 2015.
Manuscript publicized April 28, 2015.

[†]The authors are with the Graduate School of Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan.
a) E-mail: lisheng@ar.media.kyoto-u.ac.jp

DOI: 10.1587/transinf.2015EDP7047

Table 1 Organization of CCLR corpus.

	#lectures	Duration (hours)	Text size		Text type
			#words	#chars	
CCLR-TRN	58	35.2	0.31M	0.50M	caption faithful
CCLR-TST	19	11.9	0.10M	0.17M	faithful
CCLR-LSV	126	62.0	0.54M	0.81M	caption
CCLR-DEV	12	7.2	0.06M	0.10M	faithful

language, we compile a corpus of Chinese spoken lectures and investigate the ASR technology using it.

2.1 Corpus of Chinese Lecture Room

While Chinese is one of the major languages for which ASR has been investigated, studies on Chinese lecture speech recognition are limited [8], [9], and a large-scale lecture corpus for this study has not been made. We have designed and constructed a corpus of Chinese spoken lectures 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 also provided by CCTV and free-download from the official website.

As of the end of 2013, we made annotation (segmentation of the lecture part and faithful transcription) to the selected 98 lectures (90 speakers: 21 female, 69 male), which amount to 61.6 hours of speech and 1.2M characters of text. They are categorized into three general topics: 38 lectures about history-culture-art, 29 lectures about society-economy-politics, and 31 lectures about science-technology. We have also collected 126 lectures with closed captions, which are not annotated (faithfully transcribed) so far. We call all of the data both annotated and unannotated as the Corpus of Chinese Lecture Room (CCLR) [29]. For the experimental purpose, we select 58 annotated lectures as the training set (CCLR-TRN), and 19 annotated lectures as the test set (CCLR-TST). The 126 un-annotated lectures are used for lightly supervised training (CCLR-LSV). Additionally, 12 annotated lectures are held out as a development set (CCLR-DEV).

All these data sets are listed in Table 1.

2.2 Baseline ASR System and Performance

For a baseline lecture transcription system, we used CCLR-TRN of 35.2 hours as the training set, and tested on CCLR-TST.

The baseline system uses PLP-based features, consisting of 13 cepstral coefficients (including C0), plus their first and second derivatives, leading to a 39-dim feature vector. For each speaker, cepstral mean normalization (CMN) and cepstral variance normalization (CVN) are applied to the features. We build both GMM (Gaussian Mixture Model)-HMM and DNN (Deep Neural Network)-HMM systems. We adopt 113 phonemes (consonants and 5-tone vowels) as the basic HMM unit. The total number of tied triphone states is 3000, and each state has 16 Gaussian mixture components. GMM-HMM is trained with both maximum likelihood (ML) and minimum phone error (MPE) criteria.

For the DNN model training, we use the same PLP features, and the only difference from GMM-HMM is the features are globally normalized to have a zero mean and a unit variance. We use the baseline GMM-HMM (MPE) to generate the state alignment label. The network has 429 nodes as input (5 frames on each side of the current frame), 3000 nodes as output and 6 hidden layers with 1024 nodes per layer.

Training of DNN consists of the unsupervised pre-training step and the supervised fine-tuning step. In the unsupervised pre-training step, Restricted Boltzmann Machines (RBMs) are stacked in a greedy layer-wise fashion. The Gaussian-Bernoulli RBM is trained with an initial learning rate of 0.01 and the Bernoulli-Bernoulli RBMs with a rate of 0.4. The learning rate is decreased during pre-training. L2 regularization is applied to the weights with a penalty factor of 0.0002.

The supervised fine-tuning is based on frame-level cross-entropy training. We randomly select 1/8 of total utterances from CCLR-TRN for cross validation and the remaining 7/8 for supervised training. The utterance frames are presented in a randomized order while using SGD (stochastic gradient descent) to minimize the cross-entropy between the supervision labels and network output. The SGD uses mini-batches of 256 frames, and an exponentially decaying schedule that starts with an initial learning rate of 0.01 and halves the rate when the improvement in the frame accuracy on the held-out set between two successive epochs falls below 0.5%. The stopping condition is the frame accuracy increases by less than 0.1%. We used single GPU (Tesla K20m) to accelerate the training.

When testing, the PLP features are feed-forwarded through the DNN model to generate posterior probabilities of the triphone states, which are normalized by the state prior probabilities. The state prior probabilities are estimated from the training label. All these above are implemented with the Kaldi toolkit [26]. For decoding, we use Julius 4.3 (DNN version) [16] using the state transition probabilities of GMM-HMM (MPE).

The dictionary consists of 53K lexical entries from CCLR-TRN together with Hub4 and TDT4. Cut-off is not applied to define the lexicon. The OOV rate on CCLR-TST is 0.368%. The pronunciation entries were derived from the CEDICT open dictionary and the HKUST dictionary materials included in the Kaldi package. There are 1.7K English word entries and most of them are technical terms and persons' names. We converted the English phoneme set into the Mandarin phoneme set using the language transfer rules described in [27].

A word trigram language model (LM) was built for decoding. Since the size of the annotated text of CCLR-TRN is very small, we complemented it with lecture texts collected from the web, whose size is 1.07M words. Then, the lecture corpus was interpolated with other three corpora distributed

Table 2 Specification of language model training corpora.

	Corpora	#words	perplexity	weights
Component Language Models	TDT4	4.75M	1208	0.07
	HUB4	0.34M	1254	0.01
	GALE	1.03M	519	0.36
	CCLR-TRN+ web lecture text	1.07M	451	0.56
Interpolated language model		7.19M	371	/

through LDC. The interpolated weights were determined to get a lowest perplexity on CCLR-DEV. The text size, the perplexity on CCLR-DEV, and the interpolation weights are listed in Table 2.

This baseline system achieved an average Character Error Rate (CER) of 39.31% with the GMM (MLE) model, 36.66% with the GMM (MPE) model, and 31.60% with the DNN model for CCLR-TST. The main reason of the relatively low performance of the baseline system compared with the CSJ [1] and TED talks [7] is the small amount of faithful training data. Therefore, we investigate lightly supervised training to exploit unlabeled data.

3. Classifier Design for Data Selection

3.1 Lightly Supervised Training Framework

To perform lightly supervised training, we need a criterion to select data. The conventional lightly supervised training relies on simple matching between the caption text and the ASR hypothesis, and thus discards so much data which could be useful.

In this paper, we propose a data selection framework based on dedicated classifiers to replace the simple method as shown in Fig. 1. Training of the classifiers is conducted by using the training database of the baseline acoustic model (CCLR-TRN).

First, we generate an ASR hypothesis (1-best) using the baseline acoustic model and a biased language model. A biased language model is made for each lecture by interpolating the baseline model with the language model generated by the caption text of the lecture. The weights of these language models are 0.1 and 0.9. We conduct unsupervised MLLR speaker adaptation, which is also done in decoding CCLR-LSV.

Then, the ASR hypothesis is aligned with the corresponding caption text by using dynamic programming. By referring to the annotation (faithful transcript) of CCLR-TRN, both text-based and speech-based features are extracted from the alignment patterns between the ASR hypothesis and the caption text. They are used to train discriminative classifiers to select one of them or reject both.

Finally, for CCLR-LSV, an ASR hypothesis is also generated and aligned with the corresponding caption text in a similar manner. But there is no faithful annotation for this data set, so the derived classifiers are applied to select and verify word by word either from the ASR hypothesis or the caption text.

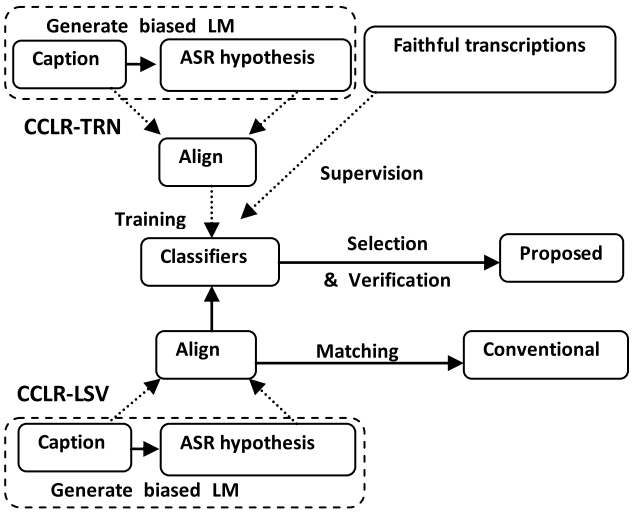

Fig. 1 Framework of proposed lightly supervised training.

Table 3 Category of alignment patterns.

	Caption		ASR hypothesis		Faithful transcriptions (reference)
C1	发表	√	发表	√	发表
C2	沦亡	X	沦亡	X	论文
C3	雪山	X	学说	X	学术
C4	雪辉	X	学会	√	学会
C5	法人	√	发热	X	法人

(X means mismatching with reference, √ means matching)

3.2 Category of Word Alignment Patterns

By analyzing the aligned word sequence between the ASR hypothesis and the caption text, we can categorize patterns by referring to the faithful transcript, as listed in Table 3. Here, insertion and deletion cases are handled by introducing a null token.

- *C1*: the ASR hypothesis is matched with the caption and also the correct transcript. A majority of the samples falls in this category.
- *C2*: although the ASR hypothesis is matched with the caption, it is not correct. This case is rare.
- *C3*, *C4* and *C5*: the ASR hypothesis is different from the caption. In *C3*, neither of them is correct. In *C4*, the ASR hypothesis is correct. In *C5*, the caption is correct.

Note that the conventional method [10], [11] is equivalent to simply using *C1* and *C2*. The objective of this study is to incorporate more effective data (*C4* and *C5*) while removing erroneous data (*C2* and *C3*).

The distribution of these patterns in CCLR-TRN is shown in Fig. 2. It is observed that 75.7% of them are categorized into *C1*. Among others, *C4* is the largest because the caption text is often edited from the faithful transcript for readability.

We initially tried to design a classifier to conduct classification of these five categories, but it turned to be difficult

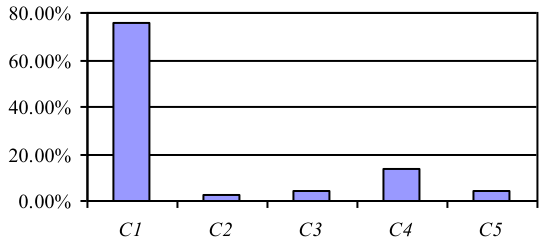

Fig. 2 Data distribution in CCLR-TRN.

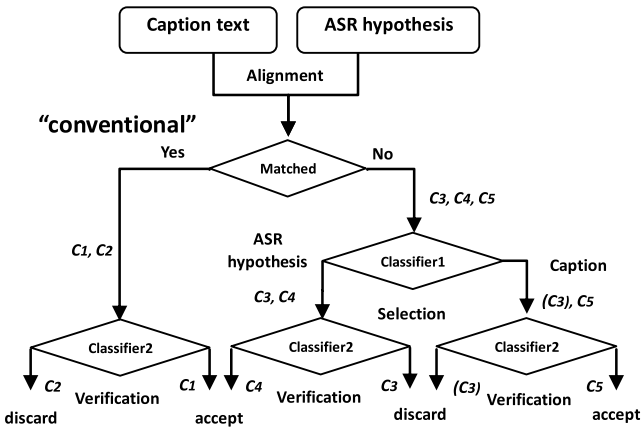

Fig. 3 Cascaded classification scheme for data selection.

because of the complex decision and the data imbalance. Therefore, we adopt a cascaded approach.

3.3 Cascaded Classifiers for Word-Level Data Selection

In the cascaded approach, we design two kinds of classifiers. One is for selection of the hypothesis and the other is for verification of the selected hypothesis.

C1 and *C2* are the matching cases between the ASR hypothesis and the caption. 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 the ASR hypothesis and the caption. We train a binary classifier to make a choice between the ASR hypothesis and the caption word. Then, we apply the other classifier to verify it. This classifier can be the same as the one used for *C1* and *C2*.

The classification is organized by the two binary classifiers in a cascaded structure as illustrated in Fig. 3. The binary classifiers are focused on specific classification problems, so they are easily optimized. This design also mitigates the data imbalance problem. In Fig. 3, one classifier is used for selection of the word hypothesis with highest credibility either from the ASR hypothesis or the caption text, and the other is used for verification of the selected (or matched) hypothesis.

To make binary classification, we merge *C3* into *C4*, because we observed the phone accuracy of the ASR hypothesis is higher than that of the caption text in *C3*. Erroneous patterns in *C3* will be rejected by the second classifier.

Table 4 Feature set for classification.

Feature Type	Features
Text-based	1. Lexical feature (LEX)
	2. Part-of-Speech (POS)
	3. Language model probability (LM)
	4. tf-idf (TF)
Speech-based	1. confidence measure by decoder (CMS)
	2. word duration (DUR)

Note that the conventional method [10], [11] simply accepts *C1* and *C2*, but our proposed method can also incorporate more effective data (*C4* and *C5*) and remove erroneous data (*C2*).

3.4 Feature Set Design for Classifiers

We use conditional random fields (CRF) [17] 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 [18], ASR error detection [19], and automatic narrative retelling assessment [20].

When training the classifiers and conducting data selection, we need to convert the alignment patterns into a feature vector. These features include both acoustic and linguistic information sources. They are selected by referring to the work on confidence measures and ASR error detection. The text-based features are defined for both ASR hypothesis and caption text while the speech-based features are computed for the ASR hypothesis only.

These features, listed in Table 4, are explained below.

- The lexical feature (LEX) is a lexical entry (ID) of the current word. It is a symbolic feature.
- The Part-of-Speech (POS) feature is obtained by a CRF classifier trained with Chinese-Tree-Bank (CTB) 4. We defined 15 POS tag symbols according to the CTB's guideline. This feature is symbolic.
- The language model probability feature (LM) is a negative log probability of the current word by unigram, bigram and trigram models. Back-off is not considered here. This feature set is numeric.
- The tf-idf (TF) feature is computed by multiplying the tf-value and the log idf-value. The tf-value is calculated from the word frequency in the caption text of the current lecture. The idf-value is computed from the caption text of entire CCLR-TRN and CCLR-LSV sets. This feature is numeric.
- The confidence measure score (CMS) is output by the Julius decoder [28] 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.

Because most of the CRF implementations are designed to work with symbolic features, we need to convert the numeric features into discrete features. To minimize the

Table 5 Number of classes of the features.

Feature	LEX	POS	LM	TF	CMS	DUR
#classes	53956	15	100*3	100	100	10

information loss in the quantization, we first normalize all of the numeric features to [0,1] and then use a step of 0.01 to get 100 binary features for the LM, TF and CMS features. The DUR feature is quantized into 10 bins. Table 5 lists the number of classes after discretization for each feature.

Moreover, for the symbolic features of LEX and POS, the contextual information of the current word is also incorporated by adding features of the preceding two words and the following two words.

3.5 Utterance Selection for Acoustic Model Training

For CCLR-LSV, the ASR hypothesis and the caption text are merged into a single word sequence after the matching and selection process, and every word in the sequence will have a label, either "accept" or "discard", based on the verification process according to Fig. 3.

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. Since the acoustic model is based on phone units, phone-based accuracy is a natural measure for selection of utterances [14]. In this work, we can compute the phone acceptance rate (PA) for every utterance by distributing the "accept" and "discard" classification results to all phones. The "PA" actually means the ratio of "accept" phones over the total number of phones in an utterance.

However, it is not easy to figure out the optimum point on the threshold of this measure between the growth of noise and the amount of training data [22]. It is affected by a number of factors and often determined a posteriori depending on the data set and the baseline performance. In this work, we will show that using only reliable utterances (PA=100%) is best for the proposed lightly supervised acoustic model training.

4. Experimental Evaluations

4.1 Classifier Implementation and Performance

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

We first conduct speech segmentation to the utterance unit based on the BIC (Bayesian Information Criterion) method [23] and speaker clustering to remove non-speech segments and speech from other than the main lecturer in CCLR-LSV.

In our implementation, we used the Wapiti CRF classifier [24] to train two classifiers using CCLR-TRN: CRF-2, which is trained to discriminate $C1$ vs. $C2$, and CRF-1, which is trained to discriminate $C3 + C4$ vs. $C5$. In the experiment, we use second-order CRF. Because of the sparse features with a high dimension (more than one thousand), L1 regularization and the Orthant-Wise Limited-memory Quasi-Newton (OWL-QN) algorithm is used to train the CRF models [25].

Table 6 Feature set evaluation of CRF-1 by 5-fold cross validation on CCLR-TRN.

	CRF-1					
	C3 + C4			*C5*		
Feature	Recall	Precision	F-score	Recall	Precision	F-score
LEX	0.831	0.818	0.825	0.709	0.727	0.718
POS	0.817	0.799	0.808	0.676	0.700	0.688
LM	0.773	0.815	0.794	0.724	0.669	0.695
TF	0.825	0.775	0.799	0.622	0.693	0.656
LEX+POS+LM+TF-IDF	0.828	0.834	0.831	0.740	0.732	0.736
CMS	0.789	0.783	0.786	0.655	0.663	0.699
DUR	0.785	0.810	0.797	0.709	0.676	0.692
CMS+DUR	0.810	0.807	0.808	0.694	0.698	0.696
All Features	**0.845**	**0.852**	**0.848**	**0.769**	**0.758**	**0.763**

Table 7 Feature set evaluation of CRF-2 by 5-fold cross validation on CCLR-TRN.

	CRF-2					
	C1			*C2*		
Feature	Recall	Precision	F-score	Recall	Precision	F-score
LEX	0.975	0.803	0.880	0.561	0.923	0.698
POS	0.960	0.828	0.889	0.634	0.897	0.743
LM	0.983	0.794	0.878	0.531	0.945	0.680
TF	0.906	0.762	0.828	0.480	0.737	0.581
LEX+POS+LM+TF-IDF	0.984	0.821	0.895	0.605	0.953	0.740
CMS	0.955	0.809	0.876	0.585	0.877	0.702
DUR	0.974	0.812	0.885	0.586	0.924	0.717
CMS+DUR	0.973	0.815	0.887	0.594	0.923	0.723
All Features	**0.985**	**0.833**	**0.903**	**0.639**	**0.958**	**0.766**

In the training dataset, there is serious imbalance between classes as observed in Fig. 2. This will bias the training of the classifiers. Thus, we introduce a re-sampling technique. Specifically, we duplicated the samples in $C2$, and discarded part of samples in $C1$ and $C3 + C4$. As a result, the calibrated distributions are as follows: $C1$: 44.1%, $C2$: 24.1%, $C3 + C4$: 19.5% and $C5$: 12.3%.

Classification performance with various feature sets is compared by 5-fold cross validation on CCLR-TRN, as shown in Table 6 and Table 7. Performance is measured by precision, recall and F-score:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

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 (Table 7) is higher than that of CRF-1 (Table 6). It suggests selection of the hypothesis is more difficult than verification of the hypothesis. In CRF-2 (Table 7), performance of $C1$ (verification) is higher than that of $C2$ (rejection), because the number of training samples of $C1$ is much larger than that of $C2$. The re-sampling technique does not essentially solve the problem of a smaller variety and coverage, though it mit-

Table 8 Confusion matrix.

REF\HYP	C1	C2	C3+C4	C5	Sum	Recall
C1	48416	752	0	0	49168	98.5%
C2	9688	17116	0	0	26804	63.9%
C3+C4	0	0	18301	3369	21670	84.5%
C5	0	0	3176	10553	13729	76.9%
Sum	58104	17868	21477	13922	111371	/
Precision	83.3%	95.8%	85.2%	75.8%	/	/

igates it.

Among the set of features, the text-based features are generally more effective than the speech-based features, but combination of both feature sets shows further improvement. As an individual feature, the lexical feature is the most effective for CRF-1, while the POS feature is the most effective for CRF-2, since more variety is needed for selection than verification of the hypothesis.

Note that the confidence measure score (CMS) is not so effective as expected. Its performance is comparable to that of the duration feature (DUR).

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. 3.

The confusion matrix with all features is shown in Table 8, and the classification rate is $C1$: 98.5%, $C2$: 63.9%, $C3+C4$: 84.5%, $C5$: 76.9%.

4.2 Utterance Selection for Model Training

Next, we investigate the effect of utterance selection based on the phone acceptance rate (PA). It is not practical to tune the threshold by using the development set, as it would take so long to train the DNN model for each PA threshold value. Therefore, the tuning is conducted with GMM-HMM (MLE) by adding the selected data to CCLR-TRN.

ASR performance (CER%) on CCLR-DEV is plotted in Fig. 4. Note that adding more data by relaxing the PA threshold only degrades the ASR performance due to the increase of errors. The best ASR performance is achieved at PA=100%. It shows the advantage of our proposed method that it can effectively select the most usable utterances and makes the data selection easy without tuning the threshold.

4.3 ASR Performance with Enhanced Model Training

Then, we conduct lightly supervised training of the acoustic model with the data selected from CCLR-LSV with PA=100%. ASR performance of the model enhanced by the selected data is evaluated on CCLR-TST. The proposed data selection method is compared with other three methods as follows:

- Baseline: the model trained by only using CCLR-TRN as described in Sect. 2. It is an expected lower bound of the proposed method.
- No selection: simply pool the CCLR-TRN lectures and entire CCLR-LSV lectures together, and directly use the ASR hypothesis of CCLR-LSV without any selection.
- Conventional: the conventional lightly supervised training method which selects the data based on simple matching of the ASR hypothesis and the caption text [10], [11] (upper part of Fig. 3).

Table 9 ASR performance (CER%) by lightly supervised training.

	Amount of data (hours)		CER% CCLR-TST		
	CCLR-TRN	CCLR-LSV	MLE	MPE	DNN
Baseline	35.2	0	39.31	36.66	31.60
No selection	35.2	62.0	38.50	34.42	28.80
Conventional	35.2	26.5	38.51	34.68	29.19
Proposed	**35.2**	**48.9**	**37.93**	**33.99**	**28.39**

Fig. 4 ASR performance (GMM-HMM on CCLR-DEV) for different PA threshold values.

In this experiment, we use the same setting with the baseline system described in Sect. 2 for GMM (MLE and MPE) and DNN acoustic model training as well as the lexicon and the language model.

ASR performance in CER is listed for GMM (MLE), GMM (MPE) and DNN models in Table 9. The results show that our proposed lightly supervised training method outperforms all other methods for MLE, MPE and DNN models. The improvement is statistically significant. The p-values from two-tailed t-test at 0.05 significant level of our proposed method compared with Baseline, No selection and Conventional methods are 0.0031, 0.0017 and 0.028 for GMM (MLE), 1.96e-07, 0.011 and 3.28e-04 for GMM (MPE) and 7.06e-09, 0.0183 and 0.0011 for DNN.

Another advantage of our method confirmed in this experiment is that it can significantly enlarge the training data by selecting usable data while discarding the erroneous segments effectively. As shown in Table 9, the percentage of the data selected from CCLR-LSV by our proposed method is 78.9%, which is almost double of the data by the conventional method (41.9%). In the conventional method, an utterance unit is discarded if it contains any words of C3, C4 or C5 cases. However, without any selection, ASR performance is degraded due to inclusion of erroneous segments. This phenomenon is also confirmed in Fig. 4. This result demonstrates that the classifiers work effectively for lightly supervised training.

5. Conclusions

We have proposed a new data selection scheme for lightly supervised training of an acoustic model. The method uses dedicated classifiers for data selection, which are trained with the training database of the baseline acoustic model. 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 lightly supervised training method effectively increases the usable training data and improves the accuracy from the baseline model and in comparison with the conventional method. This means our method can effectively identify the most credible data in huge archives of unfaithful data. This is very important for big data tasks.

References

[1] K. Maekawa, "Corpus of Spontaneous Japanese: Its Design and Evaluation," Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.7–12, 2003.
[2] H. Nanjo and T. Kawahara, "Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition," IEEE-TSAP, vol.12, no.4, pp.391–400, 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," Proc. INTERSPEECH, pp.281–284, 2006.
[4] J. Glass, T.J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent Progress in the MIT Spoken Lecture Processing Project," Proc. INTERSPEECH, pp.2553–2556, 2007.
[5] H. Yamazaki, K. Iwano, K. Shinoda, S. Furui, and H. Yokota, "Dynamic Language Model Adaptation Using Presentation Slides for Lecture Speech Recognition," Proc. INTERSPEECH, pp.2349–2352, 2007.
[6] T. Kawahara, Y. Nemoto, and Y. Akita, "Automatic Lecture Transcription by Exploiting Slide Information for Language Model Adaptation," Proc. ICASSP, pp.4929–4932, 2008.
[7] M. Paul, M. Federico, and S. Stucker, "Overview of the IWSLT 2010 Evaluation Campaign," Proc. IWSLT, pp.3–27, 2010.
[8] J. Zhang, H. Chan, P. Fung, and L. Cao, "A Comparative Study on Speech Summarization of Broadcast News and Lecture Speech," Proc. INTERSPEECH, pp.2781–2784, 2007.
[9] S.-Y. Kong, M.-R. Wu, C.-K. Lin, Y.-S. Fu, and L.-S. Lee, "Learning on Demand - Course Lecture Distillation by Information Extraction and Semantic Structuring for Spoken Documents," Proc. INTERSPEECH, pp.4709–4712, 2009.
[10] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," Computer Speech & Language, vol.16, no.1, pp.115–129, Jan. 2002.
[11] L. Nguyen and B. Xiang, "Light Supervision in Acoustic Model Training," Proc. ICASSP, vol.1, p.I-185, 2004.
[12] H.Y. Chan and P. Woodland, "Improving Broadcast News Transcription by Lightly Supervised Discriminative Training," Proc. ICASSP, vol.1, pp.737–740, 2004.
[13] T. Kawahara, M. Mimura, and Y. Akita, "Language Model Transformation Applied to Lightly Supervised Training of Acoustic Model for Congress Meetings," Proc. ICASSP, pp.3853–3856, 2009.
[14] Y. Long, M.J.F. Gales, P. Lanchantin, X. Liu, M.S. Seigel, and P.C. Woodland, "Improving Lightly Supervised Training for Broadcast Transcription," Proc. INTERSPEECH, 2013.
[15] J. Driesen and S. Renals, "Lightly supervised automatic subtitling of weather forecasts," Proc. IEEE-ASRU, pp.452–457, 2013.
[16] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," Proc. APSIPA ASC, pp.131–137, 2009.
[17] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Proc. ICML, 2001.
[18] M. Seigel and P. Woodland, "Combining Information Sources for Confidence Estimation with CRF Models," Proc. INTERSPEECH, 2011.
[19] W. Chen, S. Ananathakrishnan, R. Kumar, R. Prasad, and P. Natarajan, "ASR error dection in a conversational spoken language translation system," Proc. ICASSP, pp.7418–7422, 2013.
[20] M. Lehr, I. Shafran, E. Prud'hommeaux, and B. Roark, "Discriminative Joint Modeling of Lexical Variation and Acoustic Confusion for Automated Narrative Retelling Assessment," Proc. NAACL, 2013.
[21] H. Jiang, "Confidence measures for speech recognition: A survey," Speech Communication, vol.45, no.4, pp.455–470, April 2005.
[22] H. Lin, and J. Bilmes, "How to select a good training-data subset for transcription: submodular active selection for sequences," Proc. INTERSPEECH, pp.2859–2862, 2009.
[23] M. Mimura and T. Kawahara, "Fast Speaker Normalization and Adaptation Based on BIC for Meeting Speech Recognition," Proc. APSIPA, 2011.
[24] T. Lavergne, O. Cappé, and F. Yvon, "Practical Very Large Scale CRFs," Proc. ACL, pp.504–513, July 2010.
[25] N. Sokolovska, T. Lavergne, O. Cappé, and F. Yvon, "Efficient Learning of Sparse Conditional Random Fields for Supervised Sequence Labelling," IEEE J. Sel. Topics Signal Process., vol.4, no.6, pp.953–964, Dec. 2010.
[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," IEEE-ASRU, 2011.
[27] S. Li, and L. Wang, "Cross Linguistic Comparison of Mandarin and English EMA Articulatory Data," Proc. INTERSPEECH, 2012.
[28] A. Lee, K. Shikano, and T. Kawahara, "Real-time word confidence scoring using local posterior probabilities on tree trellis search," Proc. IEEE-ICASSP, vol.1, pp.793–796, 2004.
[29] S. Li, Y. Akita, and T. Kawahara, "Corpus and transcription system of Chinese lecture room," Proc. Chinese Spoken Language Processing (ISCSLP), pp.442–445, 2014.

Sheng Li received his B.S. degree in 2006 and M.E. degree in 2009 from computer science and software institution in Nanjing University (P.R.C.). From 2009 to 2012, he served for CAS (Chinese Academic of Sciences), doing research on LVCSR, CALL and multimodal speech synthesis. He is now a Ph.D. student in Kyoto University. His current research is spoken lecture transcription.

Yuya Akita received B.E., M.Sc. and Ph.D. degrees in 2000, 2002 and 2005, respectively, from Kyoto University. Since 2005, he has been an assistant professor at Academic Center for Computing and Media Studies, Kyoto University. His research interests include spontaneous speech recognition and spoken language processing. He is a member of IEICE, IPSJ, ASJ and IEEE. 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 B.E. in 1987, M.E. in 1989, and Ph.D. in 1995, all in information science, from Kyoto University, Kyoto, Japan. In 1990, he became a Research Associate in the Department of Information Science, Kyoto University. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor in the Academic Center for Computing and Media Studies and an Affiliated Professor in the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR and NICT. He has published more than 250 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). Dr. Kawahara received the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology (MEXT) in 2012. From 2003 to 2006, he was a member of IEEE SPS Speech Technical Committee. He was a general chair of 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 Elsevier Journal of Computer Speech and Language and APSIPA Transactions on Signal and Information. He is VP-Publications (BoG member) of APSIPA and a senior member of IEEE.