Discriminative Data Selection for Lightly Supervised Training of Acoustic Model using Closed Caption Texts

Sheng Li, Yuya Akita, Tatsuya Kawahara

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

lisheng@ar.media.kyoto-u.ac.jp

Abstract

We present a novel data selection method for lightly supervised training of 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 without tuning any threshold parameters. 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 and confidence measure scores.

Index Terms: speech recognition, acoustic model, lightly supervised training, lecture transcription

1. Introduction

Automatic transcription of lectures has been investigated for almost a decade in many institutions world-wide [1, 2, 3, 4, 5, 6, 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. In this work, we address effective acoustic model training, based on DNN (Deep Neural Network), 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.

To address this issue, a scheme of lightly supervised training, which does not require faithful transcripts but exploits available verbatim texts, has been explored for broadcast news [10, 11, 12] and parliamentary meetings [13]. In the case of TV programs, closed caption texts are used as a source for the scheme. A typical method 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 the 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.

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 accept either the ASR result or 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 Section 2. Next, our proposed method of classifier design for lightly supervised training is formulated in Section 3. Then, the implementation of the method on the lecture transcription task is explained and experimental results are presented in Section 4. The paper is concluded in Section 5.

2. Corpus and baseline ASR system

2.1. Corpus of Spoken Lectures

Since studies on Chinese lecture speech recognition are limited [8, 9], and no large-scale lecture corpus for this study is available, we have designed and constructed a corpus of Chinese spoken lectures from a popular academic lecture program "Lecture Room". We call all of the data both annotated (faithful transcript) and un-annotated (caption text only) as the Corpus of Chinese Lecture Room (CCLR).

As listed in Table 1, we selected 58 annotated lectures as the training set (CCLR-TRN), 19 annotated lectures as the test set (CCLR-TST) and 12 annotated lectures as the development set (CCLR-DEV). The 126 un-annotated lectures are used for lightly supervised training (CCLR-LSV) and they have caption text only.

Table 1 *Organization of CCLR (Corpus of Chinese Lecture Room).*

	#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-DEV	12	7.2	0.06M	0.10M	faithful
CCLR-LSV	126	62.0	0.54M	0.81M	caption

2.2. Baseline ASR system and performance

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

The baseline system uses PLP features for GMM system, consisting of 13 cepstral coefficients (including C0), plus their first and second derivatives, leading to a 39-dimensional feature vector. For each speaker, cepstral mean normalization (CMN) and cepstral variance normalization (CVN) are applied to the features. We adopt 113 phonemes (consonants and 5-tone vowels) as the basic HMM unit. We build GMM (Gaussian Mixture Model)-HMM and then DNN (Deep Neural Network)-HMM systems.

The DNN system uses 40-dimensional filterbank features plus their first and second derivatives, and has 1320 nodes as input (5 frames on each side of the current frame), 3000 nodes as output and 7 hidden layers with 1024 nodes per layer. Training of DNN consists of the unsupervised pre-training step and the supervised fine-tuning step. They are implemented with Kaldi toolkit (Karel's setup) [22]. For decoding, we use Julius 4.3 (DNN-std version) [16] using the state transition probabilities of the GMM-HMM.

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

A word trigram language model (LM) was built for decoding with Julius. We complemented the small sized text of CCLR-TRN with lecture texts collected from the web, whose size is 1.07M words. Then, this lecture corpus was interpolated with other three corpora (Hub4 of 0.34M, TDT4 of 4.75M and GALE of 1.03M) distributed through LDC. The interpolated weights were determined to get a lowest perplexity on CCLR-DEV.

This baseline system achieved an average Character Error Rate (CER) of 36.66% with the GMM-HMM model, and 30.2% with the DNN-HMM model for CCLR-TST.

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 0.9 and 0.1.

Then, the ASR hypothesis is aligned with the corresponding caption text. 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.

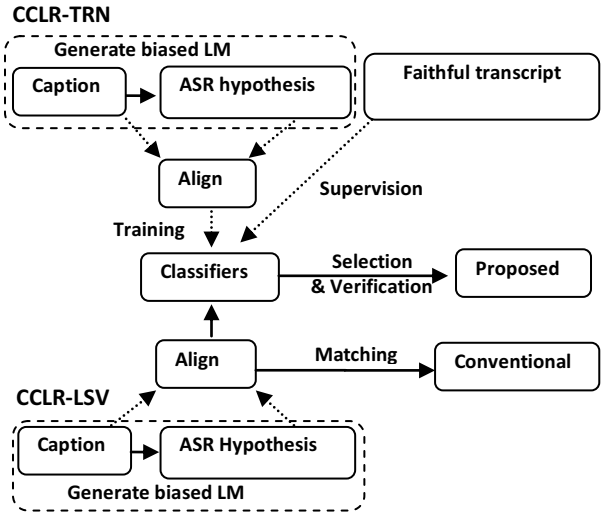

Fig. 1 *Framework of proposed data selection for lightly supervised training.*

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 2.

- ***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*, *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.

The insertion and deletion cases are regarded as a null token.

Table 2 *Category of alignment patterns.*

	Caption		ASR hypothesis		Reference (Faithful)	Percent
C1	发表	√	发表	√	发表	75.7%
C2	沦亡	x	沦亡	x	论文	2.9%
C3	雪山	x	学说	x	学术	3.9%
C4	雪辉	x	学会	√	学会	13.5%
C5	法人	√	发热	x	法人	4.0%

(x means mismatching with reference, √ means matching)

Note that the conventional method 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 Table 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 because of the complex decision and the data imbalance. Therefore, we adopt a cascaded approach.

3.3. Cascaded classifiers for 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. 2. The binary classifiers are focused on specific classification problems, so they are easily optimized. This design also mitigates the data imbalance problem. In Fig. 2, one classifier is used for selection of the word hypothesis with highest credibility either from the 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.

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

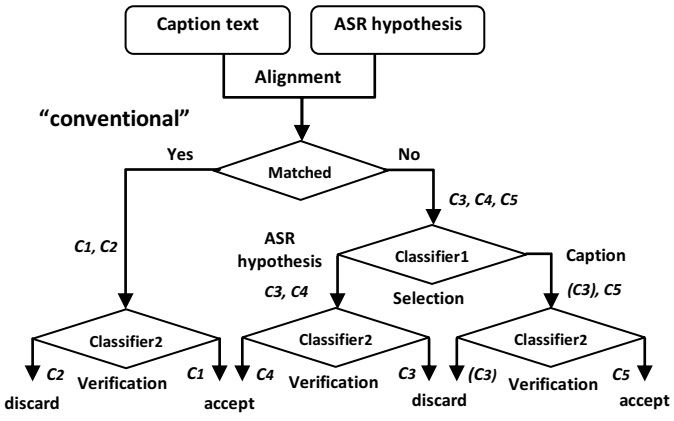

Fig. 2 *Cascaded classification scheme for data selection.*

4. Experimental evaluations

4.1. Classifier implementation and performance

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. When training the classifiers and conducting data selection, we need to convert the alignment patterns into a feature vector. A list of candidate features is shown in Table 3. These features include both acoustic and linguistic information sources. 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.

Table 3 *Feature set for classification.*

	Features	Definition
Text-based	LEX	Lexical entry (ID) of the current word
	POS	Part-of-Speech tag
	LM	Language model probability of the current word
	TF-IDF	Product of the tf-value (the word frequency in the current lecture text) and the log idf-value (inverse document frequency which is computed from the entire lecture text archive)
Speech-based	CMS	Posterior probability of the ASR hypothesis word by Julius decoder [23]
	DUR	Number of frames of the current word

The proposed method is applied to CCLR-LSV to make an enhanced acoustic model, which are tested on CCLR-TST. We first conducted speech segmentation to the utterance unit based on the BIC (Bayesian Information Criterion) method [19] and speech clustering to remove non-speech segments and speech from other than the main lecturer in CCLR-LSV.

In our implementation, we used Wapiti CRF classifier [20] 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*.

Classification accuracy with various feature sets is compared by 5-fold cross validation on CCLR-TRN, as shown in Table 4. 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. 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.

Table 4 *Classification accuracy by 5-fold cross validation on CCLR-TRN.*

Feature	CRF-2		CRF-1	
	C1	*C2*	*C3* + *C4*	*C5*
LEX	0.880	0.698	0.825	0.718
POS	0.889	0.743	0.808	0.688
LM	0.878	0.680	0.794	0.695
TF-IDF	0.828	0.581	0.799	0.656
LEX+TF-IDF +POS+LM	0.895	0.740	0.831	0.736
CMS	0.876	0.702	0.786	0.699
DUR	0.885	0.717	0.797	0.692
CMS+DUR	0.887	0.723	0.808	0.696
All Features	**0.903**	**0.766**	**0.848**	**0.763**

4.2. Utterance selection for model training

For utterance selection for acoustic model training, the phone acceptance (PA) rate is defined for every utterance of CCLR-LSV by distributing the "accept" and "reject"

classification results to all phones. We can set the lower bound of PA as a threshold for selecting utterances.

However, 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. 3. Note that adding more data by relaxing the PA threshold only degraded 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 in the lightly supervised acoustic model training.

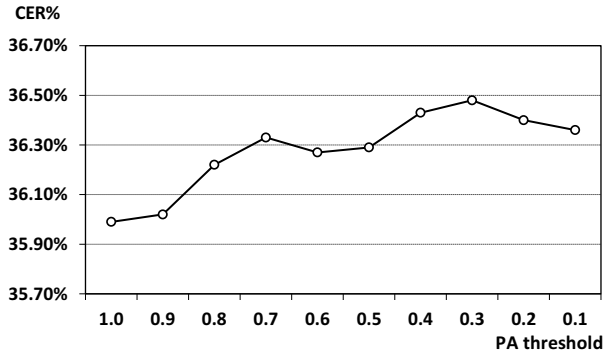

Fig. 3 *ASR performance (GMM-HMM on CCLR-DEV) for different PA threshold values.*

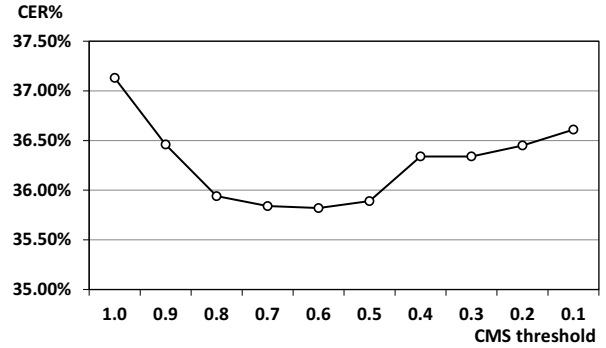

Fig. 4 *ASR performance (GMM-HMM on CCLR-DEV) for different CMS threshold value.*

4.3. ASR performance with enhanced model training

Next, we conduct lightly supervised training of the acoustic model after classification on CCLR-LSV and utterance selection. We use the same setting with the baseline system described in Section 2 for acoustic model training as well as the lexicon and language model. ASR performance of the DNN 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 Section 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.
- **CMS filtering:** For the ASR hypothesis of CCLR-LSV, the word-level confidence measure score (CMS) computed by the baseline ASR system is distributed to all phones in each word, and is averaged over the utterance unit for data selection. We train a series of GMM-HMM models (MLE) by adding the selected utterances, with different threshold values on CMS, from CCLR-LSV to CCLR-TRN. We observed an optimum point at CMS $\geqslant 0.6$ where we can get the best ASR performance on CCLR-DEV, as shown in Fig. 4.
- **Conventional matching:** the conventional lightly supervised training which selects the data based on simple matching of the ASR hypothesis and the caption text (upper part of Fig. 2).

ASR performance in CER% is listed for DNN models in Table 5. The results show that our proposed lightly supervised training method outperforms all other methods. The percentage of 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%). However, without any selection, ASR performance is degraded due to inclusion of erroneous segments. This result demonstrates that the classifiers work effectively for CCLR-LSV. Compared with the CMS filtering, the proposed method selects usable data more effectively, as confirmed in Table 5.

Table 5 *ASR performance (CER%) of DNN model by lightly supervised training.*

	Amount of data (hours)			ASR performance
	CCLR -TRN	CCLR -LSV	Total	CER%
Baseline	35.2	0	35.2	30.2
No selection	35.2	62.0	97.2	27.5
CMS filtering	35.2	46.3	81.5	27.7
Conventional matching	35.2	26.5	61.7	29.0
Proposed (PA=100%)	35.2	48.9	84.1	27.2

Another advantage of our method is it can select usable data effectively without tuning threshold parameters. Comparing Fig.3 and 4, it is apparently difficult to find the optimal point in the CMS threshold, which depends on the ASR system and the training data.

5. Conclusions

We have proposed a new data selection scheme for lightly supervised training of 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 methods.

References

[1] K.Maekawa, Corpus of Spontaneous Japanese: Its Design and Evaluation. In 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. In 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. In 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. In 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. In Proc. ICASSP, pp.4929-4932, 2008.

[7] M.Paul, M.Federico, and S.Stucker, Overview of the IWSLT 2010 Evaluation Campaign. In 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. In Proc. INTERSPEECH, pp. 2781-2784, 2007.

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

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

[11] L.Nguyen and B.Xiang. Light Supervision in Acoustic Model Training. In Proc. ICASSP, Vol. 1, pp. I-185, 2004.

[12] H.Chan and P.Woodland. Improving Broadcast News Transcription by Lightly Supervised Discriminative Training. In 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. In 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. In Proc. INTERSPEECH, 2013.

[15] J.Driesen and S.Renals. Lightly supervised automatic subtitling of weather forecasts. IEEE-ASRU, 2013.

[16] A.Lee and T.Kawahara. Recent development of open-source speech recognition engine Julius. In 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. In Proc. ICML, 2001.

[18] H.Lin, and J.Bilmes, How to select a good training-data subset for transcription: submodular active selection for sequences, In Proc. INTERSPEECH, pp.2859-2862, 2009.

[19] M.Mimura, T.Kawahara, Fast Speaker Normalization and Adaptation Based on BIC for Meeting Speech Recognition, in Proc. APSIPA, 2011.

[20] T. Lavergne, O. Cappé, and F. Yvon. Practical Very Large Scale CRFs. In Proc. 48th Annual Meeting Association for Computational Linguistics (ACL), pages 504-513, July 2010.

[21] N. Sokolovska, T. Lavergne, O. Cappé, and F. Yvon. Efficient Learning of Sparse Conditional Random Fields for Supervised Sequence Labeling. IEEE J. Sel. Topics Signal Process, 4(6):953-964, December 2010.

[22] 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.

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