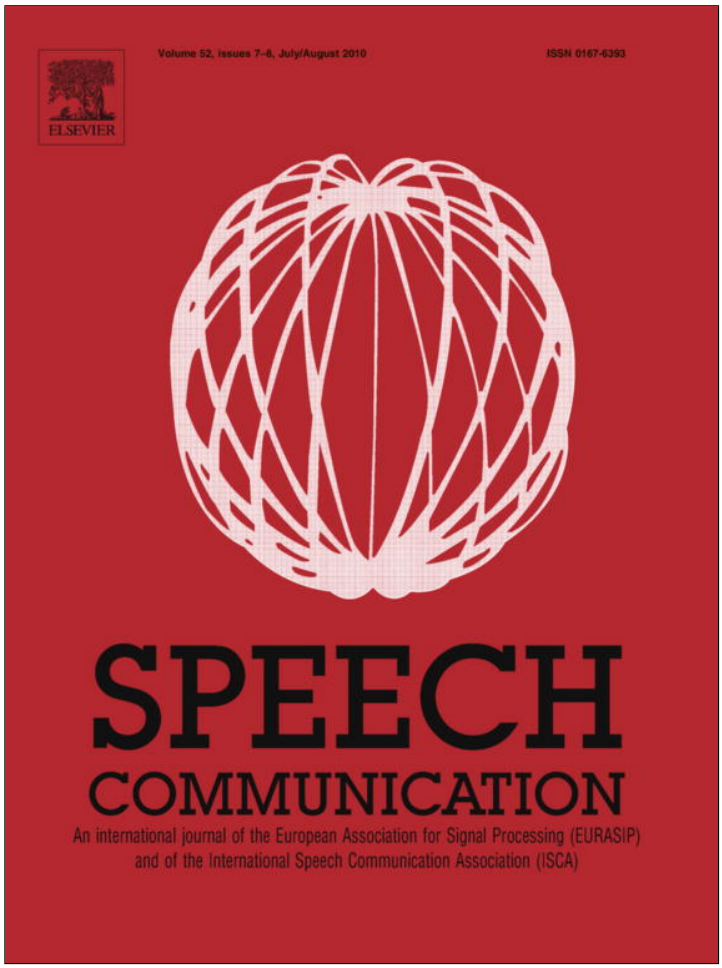

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 652–663
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Unsupervised training and directed manual transcription for LVCSR ☆

Kai Yu *, Mark Gales, Lan Wang [1], Philip C. Woodland

Machine Intelligence Lab, Cambridge University Engineering Department, Cambridge CB2 1PZ, UK

Received 21 February 2009; received in revised form 19 June 2009; accepted 23 February 2010

Abstract

A significant cost in obtaining acoustic training data is the generation of accurate transcriptions. When no transcription is available, *unsupervised training* techniques must be used. Furthermore, the use of discriminative training has become a standard feature of state-of-the-art large vocabulary continuous speech recognition (LVCSR) system. In unsupervised training, unlabelled data are recognised using a seed model and the hypotheses from the recognition system are used as transcriptions for training. In contrast to maximum likelihood training, the performance of discriminative training is more sensitive to the quality of the transcriptions. One approach to deal with this issue is data selection, where only well recognised data are selected for training. More effectively, as the key contribution of this work, an active learning technique, *directed manual transcription*, can be used. Here a relatively small amount of poorly recognised data is manually transcribed to supplement the automatic transcriptions. Experiments show that using the data selection approach for discriminative training yields disappointing performance improvement on the data which is mismatched to the training data type of the seed model. However, using the directed manual transcription approach can yield significant improvements in recognition accuracy on all types of data.

Keywords: Unsupervised training; Discriminative training; Automatic transcription; Data selection

1. Introduction

A recent trend in building large vocabulary speech recognition systems is to use very large acoustic model training sets to improve parameter estimation and hence recognition performance (Evermann et al., 2005). For some tasks, such as automatic transcription of Broadcast data, it is fairly easy to obtain thousands of hours of audio from radio and television shows. However, in order to train acoustic models, word level transcriptions are required. Hence, a major cost of using large amounts of broadcast data for training is the provision of accurate manual transcriptions.

In some cases, approximate manual transcriptions, such as closed captions, are available. These approximate transcriptions can be used with lightly-supervised training (Lamel et al., 2002; Chan and Woodland, 2004). Here a biased language model is created from the approximate transcriptions and used to recognise the audio data. This leads to low error rate semi-automatic transcriptions which yield good performance for hidden Markov model (HMM) parameter estimation using both maximum likelihood (ML) (Lamel et al., 2002) and discriminative criteria (Chan and Woodland, 2004). However, in some cases, for instance broadcast news transcription in Arabic and Mandarin, even approximate transcriptions are not available. Here *unsupervised training* techniques must be used.

☆ This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022. Many thanks go to X.A. Liu for training some of the language models used in the experiments.
* Corresponding author. Tel.: +44 7876570319.
E-mail address: ky219@cam.ac.uk (K. Yu).
[1] Present address: ShenZhen Institute of Advanced Technology, Chinese Academy of Sciences, A-3 NanShan Medical Equipment Industry Park, No. 1019, NanHai Avenue, ShenZhen 518067, PR China. The work was done when the author was at Cambridge University.

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.02.014

In unsupervised training, a seed model is normally used to recognise the untranscribed audio. Then automatically generated transcriptions are used during training. Most previous studies of unsupervised training examined Maximum Likelihood (ML) estimation techniques (Kemp and Waibel, 1999; Lamel et al., 2001; Lamel et al., 2002; Wessel and Ney, 2005; Riccardi and Hakkani-Tur, 2003; Kamm and Meyer, 2002). These studies have found that iterative addition of unsupervised data and confidence score based data selection can yield reduction in word error rate (WER). However, the majority of state-of-the-art speech recognition systems make use of discriminative training approaches, such as Minimum Phone Error (MPE) (Povey and Woodland, 2002). Recently, unsupervised discriminative training has been investigated (Ma et al., 2006; Wang et al., 2007; Yu and Gales, 2007). As discriminative schemes aim to reduce the recognition error of the training data with respect to the (assumed) "correct" transcription, it is not surprising that discriminative training is more sensitive to the accuracy of the transcriptions than ML training. Furthermore, if the transcriptions for untranscribed audio data are automatically generated, then the set of most probable alternative hypotheses required in discriminative training tends to be closer to the "correct" transcription than if manual transcriptions had been used. The discrimination ability of the trained model may then be reduced. These sensitivities may limit the size of performance improvements if the training data has a high error rate with the seed model, such as when the unlabelled data is mismatched to that used in seed model training. Therefore, when testing on mismatched data, the performance gains are often small. For example, Broadcast Conversation (BC) data is normally spontaneous speech, where filled pauses, word fragments, reduced articulation or mispronunciation, and non-speech events such as laughter and coughing may frequently happen. In contrast, Broadcast News (BN) data consists mainly of prepared speech. Thus the two types of speech usually have large acoustic and linguistic differences (Nakamura et al., 2007) and typically BC data has higher error rates than BN data. It has been reported that the associated performance improvement using unsupervised discriminative training on the BC test set is much smaller than for BN data (Wang et al., 2007). This paper investigates an effective approach to improve the performance of unsupervised discriminative training.

For transcription generation in unsupervised training, two strategies may be used. The standard approach is to automatically recognise the audio using a seed model. Data selection can then be applied to remove data that are believed to be poorly transcribed. The retained data are then added to the original training dataset and used to train the acoustic model and optionally the language model (Kemp and Waibel, 1999; Lamel et al., 2002; Wang et al., 2007). An alternative strategy, based on the theory of active learning (Cohn et al., 1994) is proposed in this work. Here, a small amount of data, which is believed to be poorly recognised, is selected automatically. This subset is then manually transcribed to supplement the fully automatic transcriptions (Kamm and Meyer, 2002; Riccardi and Hakkani-Tur, 2003). In this work, this approach is applied to discriminative training. This is referred to as *directed manual transcription* (Yu and Gales, 2007) and is investigated in detail. The underlying assumption is that the inclusion of correctly transcribed high error rate data is likely to be more useful for improving the quality of the acoustic model.

The remainder of this paper is organised as follows. Section 2 describes the unsupervised training procedures. The sensitivity of discriminative training to errorful transcription is discussed in Section 3. In Section 4, two strategies for transcription generation are investigated in the context of a state-of-the-art discriminatively trained system for the recognition of Mandarin broadcast audio.

2. Unsupervised training for LVCSR

This section describes the basic unsupervised training procedure for large vocabulary continuous speech recognition (LVCSR) systems. Seed acoustic models and language models have been trained on some supervised data. The general procedure is to first recognise the untranscribed audio using the seed models. Data selection approaches may then be used to filter out poorly recognised data. The untranscribed audio with the automatic transcriptions are then added to the original training set for acoustic (ML and discriminative) and/or language model training. In this work, unsupervised training for the acoustic models is the focus because the effect of adding unlabelled data for language model training has been shown to have a smaller effect (Yu and Gales, 2007).

The complete procedure requires segmentation of the unlabelled data, automatic transcription generation, data selection and model training. The approach implemented in this work is described in more detail as an example of this overall procedure.

2.1. Automatic transcription generation

The initial stage for unsupervised training is *automatic segmentation* of the untranscribed audio. The procedure described in Sinha et al. (2006) is also used here. First, advertisement removal is run by detecting repeated blocks of audio data, for example jingles or commercials. Acoustic segmentation is then performed based on Gaussian mixture models (GMM) of different sound types. The data is split into wide-band speech and telephone (narrow-band) speech during this process. Segments of music are discarded. Finally gender detection and speaker clustering are used to generate speaker labels for adaptation.

Given the automatic segmentation, automatic transcriptions are then generated using the seed models trained on the original supervised data. In this work, in order to generate transcriptions with low word error rate, a multi-pass recognition system with discriminatively trained acoustic

models and unsupervised adaptation is employed. This system is normally used as the lattice generation stage in the state-of-the-art Cambridge multi-pass speech recognition framework (Evermann and Woodland, 2003), often referred to as a P1-P2 system. The adapted decoding system has two stages:

- P1: Gender-independent models trained with the MPE criterion are used to generate initial transcriptions with a trigram language model and relatively tight beam-widths.
- P2: The 1-best hypotheses from the P1 stage are used to generate adaptation transforms for gender-dependent MPE trained models. Here least square linear regression and diagonal variance transforms are estimated. Using the adapted MPE trained models, lattices are generated using a trigram language model with a wider beam-width. These lattices are then rescored using a 4-gram language model.

In this work, the Viterbi 1-best recognition result of the P2 stage is used as the transcription for unsupervised training, as this is felt to give a better balance between deletions and insertions. However, in order to perform effective data selection in the later stage, confidence scores associated with each word are normally required. Therefore, additional confusion network (CN) decoding (Mangu, 2000) is performed on the P2 lattices. This yields another set of transcriptions with associated confidence scores. Note, these transcriptions are only used in the data selection stage rather than in the training stage. This is because CN decoding tends to increase the deletion rate.

2.2. Data selection

One fundamental issue in unsupervised training is that both the audio and transcriptions are not manually checked to ensure both appropriateness and quality for acoustic model training. This may lead to two problems:

(1) the audio are not guaranteed to be in the target language;
(2) the quality of the automatic transcriptions may be very poor, which may significantly affect the model training.

Both problems occur in the task considered in this work to transcribe Mandarin Chinese broadcasts. In Mandarin broadcasts there are some shows that have significant levels of English content. For carefully selected data, such as that used for the 2003 and 2004 NIST Mandarin broadcast news (BN) transcription evaluations, the percentage of English data is typically in the range of 1–2%. However for some shows this percentage is significantly higher. Rather than relying on the use of confidence scores to remove these large segments of data during data selection, it would be preferable to eliminate the complete broadcast from the training set as they are unlikely to be appropriate for training Mandarin acoustic models. In this work, the adapted decoding system described above is a dual language system that can output both Mandarin and English. Detection of non-Mandarin shows is performed by setting a threshold on the percentage of English words recognised for that show and on the overall show-level confidence score as described below. Though the dual language system was based on English and Mandarin, it is found to detect other non-Mandarin data such as shows containing a large percentage of German speech.

After removing broadcast shows unsuitable for training, further data selection may then be performed in order to filter out the poorly transcribed unlabelled data (Wang et al., 2007; Ma et al., 2006). The selection can be done at segment-level or show-level, and both are based on the confidence scores generated in the P2 stage. Show or segment level confidence scores are calculated by averaging the confidence scores of each word using the same formulae:

$$C_{\mathcal{S}} = \frac{\sum_{\mathcal{W}\in\mathcal{S}} C_{\mathcal{W}} T_{\mathcal{W}}}{\sum_{\mathcal{W}\in\mathcal{S}} T_{\mathcal{W}}}, \tag{1}$$

where $\mathcal{S}$ can be one show or one segment, $C_{\mathcal{S}}$ is the averaged confidence score of $\mathcal{S}$, $C_{\mathcal{W}}$ is the confidence score of word $\mathcal{W}$ within $\mathcal{S}$ calculated by the adapted decoding system, $T_{\mathcal{W}}$ is the duration of word $\mathcal{W}$. A threshold on $C_{\mathcal{S}}$ is set to split the unlabelled data into two parts. Those segments with higher confidence scores than the threshold are retained, while those below the threshold are removed. This approach has previously been adopted in Chan and Woodland (2004), Ma et al. (2006), Wang et al. (2007), where it was shown to improve the performance of both ML and discriminative training on the data with the same genre as the training data of the seed model.

2.3. Directed manual transcription

Though the data selection approach introduced in the previous section can yield improvements on the data of the same type as the training data, the gains for discriminative training are limited when the data is from a different genre to the seed model (Wang et al., 2007; Yu and Gales, 2007). This data tends to have higher error rate. To address this problem, based on the framework of active learning (Cohn et al., 1994), the proposed strategy is to incorporate some supervised data for the poorly recognised genre. This approach can be implemented within the unsupervised training framework described earlier. Rather than discarding the segments whose confidence scores fall below the set threshold $C_{\mathcal{S}}$, those segments are manually transcribed to supplement the automatic transcriptions used in the data selection. This is the *directed manual transcription* approach (Yu and Gales, 2007). Compared to previous research on active learning for speech recognition (Kamm and Meyer, 2002; Riccardi and Hakkani-Tur, 2003), this work concentrates on the performance improvement of *discriminative training*, which is a key issue in state-of-the-art LVCSR systems.

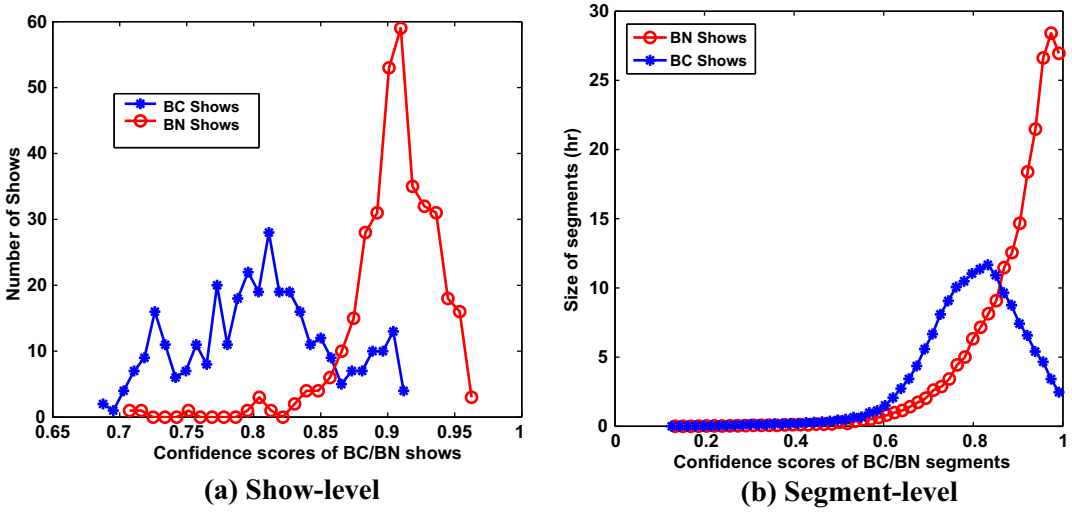

Fig. 1. Confidence score distribution of BC and BN Mandarin data using seed acoustic model trained on BN data.

Directed manual transcription is based on confidence scores. Poorly recognised genre will dominate the data selection for manual transcription. This is because the seed model will normally recognise the unlabelled data of matched type better than data which are mismatched. Consequently the confidence scores are higher for matched data. For example, the Mandarin broadcast task contains two different types of audio: broadcast news (BN) and broadcast conversation (BC). Experiments have shown that the two types of data have significant differences in both the acoustic and transcription statistics (Wang et al., 2007). As the seed models in this work are trained on supervised BN data, the untranscribed BC data will have lower confidence scores.[2]

Fig. 1 shows the confidence score distribution of BN and BC data used in this work. Those confidence scores were generated using the seed model trained on BN dominant data set. Details are described in Section 4. It can be observed that both confidence score distributions have two peaks, corresponding to BN and BC data with the BC data normally having lower confidence scores. As the type of the data is labelled at the show level, the difference between the two types is more distinct in Fig. 1(a). In contrast,the use of segment-level confidence scores yields a smoother distribution and allows finer grained data selection. The initial investigation of directed manual transcription (Yu and Gales, 2007) used show-level confidence scores. In this work, segment-level confidence score selection is used unless explicitly stated.

3. Discriminative training with unlabelled data

After data selection and/or directed manual transcription, transcriptions for a subset of the untranscribed audio are available. These data are then added to the original transcribed training set. The acoustic models are then trained using this combined training set in the standard fashion. The general procedure is to first perform maximum likelihood (ML) training. These initial models are then refined using MPE training.

ML training with untranscribed audio is proved to be fairly insensitive to the transcription quality (Kemp and Waibel, 1999; Lamel et al., 2001; Lamel et al., 2002). However, discriminative MPE training has been shown to be far more sensitive to the transcription quality (Wang et al., 2007). This section will discuss discriminative training in detail and the issues associated with unsupervised discriminative training.

3.1. Discriminative training

Discriminative training criteria explicitly aim to reduce the recognition errors on the training data. Thus these criteria not only take into account the correct word sequence, but also incorrect, confusable hypotheses. One class of discriminative criteria is based on minimizing the Bayes risk (MBR) (Doumpiotis et al., 2003). This can be expressed as

$$\mathcal{M}_{\mathtt{dl}} = \arg\min_{\mathcal{M}} \left\{ \sum_{r} \sum_{\mathcal{H}} P(\mathcal{H}|\mathbf{O}^{(r)}, \mathcal{M}) \mathcal{L}\left(\mathcal{H}, \mathcal{H}^{(r)}_{\mathtt{ref}}\right) \right\}, \tag{2}$$

where $\mathcal{M}$ is the model parameter set, r is the index of utterances, $\mathcal{H}^{(r)}_{\mathtt{ref}}$ is the correct transcription for utterance r, $\mathcal{L}\left(\mathcal{H}, \mathcal{H}^{(r)}_{\mathtt{ref}}\right)$ is a loss function defining the difference between any possible hypothesis $\mathcal{H}$ and the correct transcription $\mathcal{H}^{(r)}_{\mathtt{ref}}$. From Eq. (2), the posterior probability of the hypothesis $P(\mathcal{H}|\mathbf{O}^{(r)}, \mathcal{M})$ is required, which can be expressed as

$$P(\mathcal{H}|\mathbf{O}^{(r)}, \mathcal{M}) = \frac{p(\mathbf{O}^{(r)}|\mathcal{H})P(\mathcal{H})}{\sum_{\breve{\mathcal{H}}} p(\mathbf{O}^{(r)}|\breve{\mathcal{H}})P(\breve{\mathcal{H}})}. \tag{3}$$

[2] As the length of the automatic segments vary a lot, the counts of segment-level confidence scores are based on the accumulated duration.

Due to the use of the above posterior distribution, it is necessary to calculate a number of confusable hypotheses $\breve{\mathcal{H}}$. These confusable competing hypotheses are normally generated using an existing STT system. Lattices are used in this work as a compact representation of multiple hypotheses. Usually word lattices are converted to phone lattices before discriminative training. In this paper, the minimum phone error (MPE) criterion is employed, where $\mathcal{L}(\cdot)$ is defined as the phone error of each hypothesis (Povey and Woodland, 2002). With this loss function, reference transcriptions are used to calculate the average phone accuracy of the confusable hypotheses lattices generated from an STT system. Then, two sets of lattices are used for MPE training. The *numerator lattices* consist of phone arcs with accuracies higher than the average phone accuracy, while the *denominator lattices* consist of the rest low accuracy phone arcs. Details about the MPE lattice definition can be found in Povey (2003).

The MPE criterion can be optimised using the weak-sense auxiliary function described in Povey (2003). The final parameter update formulae are extensions of the standard Baum–Welch algorithm. The new mean of Gaussian component m can be updated as below (Povey and Woodland, 2002):

$$\boldsymbol{\mu}^{(m)} = \frac{\sum_t \gamma_m^{\mathrm{n}}(t)\mathbf{o}_t - \sum_t \gamma_m^{\mathrm{d}}(t)\mathbf{o}_t + D_m\hat{\boldsymbol{\mu}}^{(m)} + \tau^I \boldsymbol{\mu}_{\mathrm{ML}}^{(m)}}{\sum_t \gamma_m^{\mathrm{n}}(t) - \sum_t \gamma_m^{\mathrm{d}}(t) + D_m + \tau^I}, \tag{4}$$

where $\mathbf{o}_t$ is the observation at time t, $\gamma_m^{\mathrm{n}}(t)$ is the posterior occupancy of component m at time t calculated using forward-backward algorithm given the numerator lattices, $\gamma_m^{\mathrm{d}}(t)$ is the occupancy calculated given the denominator lattices, D_m is a smoothing constant to ensure convergence of the weak-sense auxiliary function, $\hat{\boldsymbol{\mu}}^{(m)}$ is the estimated mean vector of the previous iteration, $\boldsymbol{\mu}_{\mathrm{ML}}^{(m)} = \left(\sum_t \gamma_m^{\mathrm{ML}}(t)\mathbf{o}_t\right)/\left(\sum_t \gamma_m^{\mathrm{ML}}(t)\right)$ is the ML estimate of the mean vector used to increase the generalisation ability of the final discriminative estimates, τ^I is a constant to control the weighting of the ML estimate.

In order for discriminative training to operate well, there should be sufficient difference between the reference transcription and the competing hypotheses. To achieve this, an ML model with a weakened language model, normally a heavily pruned bigram or unigram, is normally used in the decoding STT system to generate competing hypotheses (Povey, 2003). The standard rationale for this is that the weakened language model increases the number of confusions in the data, hence improving the generalisation of discriminative training to unseen data.

3.2. Issues in unsupervised discriminative training

For unsupervised discriminative training, the use of a "weakened" language model for generating the denominator lattices is important. Competing confusable hypotheses are normally generated using the same STT system as the system generating the "reference". However, they need to be sufficiently different for discriminative training. If the same language model and acoustic models are used, then the best path of the transcription and the competing hypotheses must, by definition, be the same.[3] This may limit any possible reductions in error rate from discriminative training. However, even if a weakened language model is used, this problem still exists. This effect can be illustrated by comparing the character error rate (CER) for scoring the best path of the competing hypotheses (denominator) against different types of reference transcriptions.

Table 1
% CER of the best path of denominator lattices against two types of reference transcriptions.

Reference type	BC	BN
Manual	42.4	22.1
Auto.	31.9	17.7

Table 1 shows the performance (CER %) of the 1-best hypothesis in the denominator lattices generated using the "weakened" language model and either the manual transcriptions or the automatically-derived transcriptions. The results are quoted on subset1 of the Mandarin training data, see Section 4.1 for details. Note the automatically-derived transcriptions use the multi-pass adaptation framework with 4-gram language models described in Section 2.1. As expected, in Table 1, the CER is lower when using the automatically-derived transcriptions compared to the manual transcriptions. This illustrates the bias from using the same acoustic models discussed in the previous section. To get an idea of the accuracy of the transcriptions generated using the multi-pass adaptation framework, the performance on two types of test data was evaluated. For the BN and BC data, CERs of 24.2% and 11.7% respectively were obtained.

The issues discussed above can also be illustrated by looking at the approximate MPE criterion computed during training. For this experiment the S0 training data was either augmented by subset1 with the manual transcriptions (S1), or the automatically-derived transcriptions (S2). Fig. 2 shows the approximate expected phone error rate, based on the two different types of reference transcriptions, against MPE iteration for these two systems. The MPE criterion for the S2 system is consistently lower than that of the S1 system. Thus using the automatically-derived transcriptions yields an artificially low expected phone error rate compared to the accurate manual transcriptions. Thus the MPE trained system may not be able "correct" errors due to the incorrect transcriptions.

The above results have not used any data selection approaches. Using data selection can reduce the error rate of the transcriptions in the selected data. Thus it would be expected that the differences between the equivalent S1 and

[3] In this case, it is still possible to discriminatively update the system as the *expected* loss is used for MPE training. However, this is only of theoretical interest.

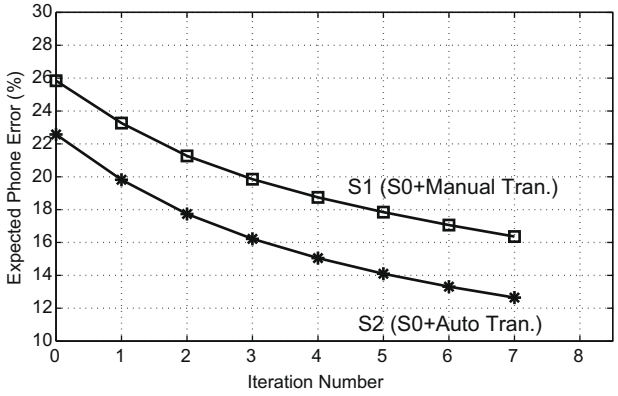

Fig. 2. Expected phone error on training data for MPE training.

S2 systems above would be decreased. However, a side-effect of selecting data that the seed system recognises well is that the data most useful for improving the models, the data on which the system performs poorly, will be exactly the data removed during data selection. This has resulted in limited reductions in error rate using data selection with discriminative training (Wang et al., 2007). This limitation will be discussed in more detail in the next section.

4. Experimental investigation of unsupervised discriminative training

This section discusses the recognition performance using unsupervised discriminative training. A Mandarin broadcast transcription task is used. The seed model will be trained primarily on BN-style data. This allows the performance for both relatively low error rate, matched BN data, and high error rate, mismatched BC data, to be investigated. The data summary and experimental setup is described in Section 4.1. Then, the two unsupervised training strategies are discussed in detail in the following sections.

4.1. Data summary and experimental setup

As discussed in Section 2, the starting point for unsupervised training is seed acoustic models trained on supervised data, here referred to as S0. The training data set for S0[4] is shown in Table 2. The dominant type of the S0 data set is BN-style, while BC-style data comprises only about 10% of the whole data set. Therefore, the seed model trained on S0 is expected to yield worse performance on BC-style untranscribed data than on BN-style data.

The primary dataset used for unsupervised training experiment includes broadcast news, broadcast conversation and some non-Mandarin shows. For these data, manual transcriptions as well as manual segmentations are available. This allows a contrast to the use of unsupervised training with standard supervised training. This data set was used in both supervised and unsupervised training. Non-Mandarin shows were removed before supervised training. This data set is referred to as subset1. S1 is the model trained on the combination of the S0 training data and subset1 data set with manual transcriptions and segmentations. The same audio was also used in unsupervised training as if the manual transcriptions were not available. In unsupervised training, automatic non-Mandarin show detection was first performed. The untranscribed audio was decoded using the system described in Section 2.1 with the seed model S0. Two thresholds were then used in the show level selection. The first was a show-level confidence score of 55%, the second was a threshold of 20% for the percentage of English. Four shows were detected as non-Mandarin and these were manually checked. Three of the shows contained large amounts of English interviews and the other show contained only songs. The amount of data removed using this show selection approach depends significantly on the care taken in selecting the sources and the shows recorded. In previous work on the BN data released under the DARPA EARS program, a far larger percentage of shows were detected as English (Sinha et al., 2006). After automatic non-Mandarin show detection, the remaining audio was used to train models with unsupervised training techniques. This is the S2 system, which uses the same broadcast audio data as S1 but with some unlabelled data. It is worth noting that, since automatic segmentation and automatic removal of non-Mandarin shows were required, the actual amount of data used for unsupervised training of the S2 system (before confidence score based data selection) is slightly less than the data for the S1 system.

In addition to subset1, a second dataset, subset2, with no manual transcription was also used in some experiments. Systems trained on the combination of subset1, subset2 and the S0 training data set are referred to as S3. This was used to investigate whether increasing the amount of unlabelled data will continuously improve the system performance. In the above systems, BN-style data is always dominant. It is also useful to know the performance of different types of data when the training data has a balanced combination of BC and BN data. To get such an unsupervised training data set, the BN part of subset1 and the BC part of subset2 were added to the S0 data set, forming the

Table 2
Training data summary and different acoustic models.

Sys.	Data	Trans. type	Size (h)		
			BN	BC	ALL
S0	Baseline	Manual	155.6	19.7	186.4
S1	+Subset1		363.2	150.8	514.0
S2	+Subset1	Auto.	352.3	152.2	504.5
S3	+Subset1/2		654.0	303.2	957.2
S4	+Subset1(BN)/2(BC)		352.3	303.2	655.5

[4] This training data set also includes 11.1 h of English data including 10 h of randomly selected TDT4 English data and 1 h of English data contained in the Mandarin data set.

S4 system in Table 2. Here, the amount of BN and BC training data are approximately the same.

Two test sets were used to evaluate different systems: bnmdev06 and bcmdev05. bnmdev06 comprises 3.6 h of data taken from a range of BN sources. It includes some of the standard existing test sets described in Sinha et al. (2006), dev04f (0.5 h), eval03m (0.6 h) and eval04 (1.0 h). In addition a more recent set of 4 shows (1.6 h) taken from July to October 2006 were also used. The evaluation data for BC, bcmdev05, comprises 2.5 h of data taken from 5 BC shows during March 2005.

In all acoustic systems, the basic acoustic features were 13 Cepstral coefficients (including energy) and their derivatives, derived from Mel-Frequency Perceptual Linear Prediction (MF-PLP) analysis (Woodland et al., 1997) and segment level Cepstral Mean Normalisation (CMN) (Woodland et al., 1994). The static coefficients were appended with 1st, 2nd and 3rd order derivatives to form a 52-dimensional feature vector and then projected using a Heteroscedastic Linear Discriminant Analysis (HLDA) (Kumar, 1997) transform to 39-dimensions. As Mandarin is a tonal language, smoothed log pitch frequency features were also extracted along with the 1st and 2nd order derivatives (Gales et al., 2005) and appended to the other features. State-clustered triphone HMMs, with 6K distinct states and an average of 36 Gaussian components per state were used for all systems. The same decision tree and HLDA transform was used for all systems in this paper.

The baseline language model used in the experiments was trained using various sources including the LDC Chinese giga-word release and web download data. In addition, manual transcriptions associated with the acoustic data for the S0 baseline were also used for language model training. All text was processed using a simple character-to-word segmenter based on longest-first match. The multi-character word-list for this consists of about 51K words. Any Chinese character that was not present in the multi-character word-list was processed as an individual word. The total word-list, including single-character Mandarin words and the 10K most frequent English words, was 68K words. Three separate language model (LM) components were built and interpolated to construct the baseline language model. The first component, broadcast news, was trained using about 1074M words of text. This component was interpolated with a general English LM at a ratio of 9:1 for the interpolation weights. The use of English component is because, during the recognition of unlabelled audio, English needs to be output for non-Mandarin show detection. The second is a broadcast conversation component, trained only on the transcriptions for the 19 h of BC data, 0.24M words. Finally, the third component was built using web-data from Phoenix TV (PHX).[5] This training set consists of 64M words, which was checked to ensure there is no overlap with any of the test or unlabelled data. This data was found to be suitable for both BN and BC transcription. Word-based trigram and 4-gram LMs were then trained for each component and interpolated and merged to form the final language model.

[5] Thanks to SRI and the GALE Nightingale team for making this data available.

Table 3
Unadapted decoding performance using HMMs estimated with supervised and unsupervised training.

System	S0 + (h)		bnmdev06		bcmdev05	
	Man.	Auto.	ML	MPE	ML	MPE
S0	0	0	15.1	13.6	29.3	25.4
S1	327.6	0	12.8	10.5	26.1	21.8
S2	0	318.1	13.8	12.0	27.7	24.8
S3	0	770.8	13.3	11.6	27.7	24.6
S4	0	469.1	13.8	12.2	27.3	24.7

The baseline (S0) acoustic and language models were used to recognise the untranscribed audio. An adapted decoding framework was used, i.e. the P1-P2 system described in Section 2.1. All experiments on the test datasets bnmdev06 and bcmdev05 used unadapted Viterbi decoding with the trigram baseline LM unless explicitly stated. The initial performance comparison between supervised and unsupervised training is shown in Table 3:

The second row of Table 3 is the performance of supervised training, which can be treated as the upper bound of unsupervised training performance. For ML training, the CER reduction of supervised training compared to S0 is 2.3% on bnmdev06 and 3.2% on bcmdev05. For MPE training, the corresponding CER reductions are 3.1% and 3.6%, respectively. The CER reductions of both ML and MPE training on bcmdev05 are better than those on bnmdev06. This is because the relative amount of BC data was significantly increased in S1. However, a different trend appears in unsupervised training with automatic transcriptions (S2). With ML training, S2 led to a larger CER reduction over S0 on bcmdev05 (1.6%) than on bnmdev06 (1.3%). In contrast, for MPE training, the CER reduction on bcmdev05 is only 0.6%, which is less than half of that on bnmdev06 (1.6%). The relative gain from supervised training can also be calculated to show the effectiveness of unsupervised training. For the S2 MPE system, the proportion of the CER reduction from supervised MPE training is much bigger on BN (57%), while on BC, it is only 17%. This demonstrates that the higher error rate of BC data can significantly affect the performance of unsupervised discriminative training. If the type of the test data is well matched to the well recognised unlabelled data, the performance gain is closer to that from supervised training, otherwise, it may be greatly limited.

One option to improve the performance on BC-style test data is to add more unlabelled data. This was investigated by adding subset2 data into unsupervised training pool to build the S3 system. Table 3 shows that this only yielded small additional MPE gains on bcmdev05 (0.2%). It can be seen that performance of unsupervised training with

770.8 h data is disappointingly compared to supervised training with less than half of the data on bcmdev05. This is because, due to the use of the same adapted decoding system, the automatic transcriptions generated for the additional subset2 data are still of low quality for the BC portion. To examine whether the large quantities of BN data has taken some possible gains away from the BC data, a system was built by adding just the BC data from subset2, which led to an approximately balanced training data set. This is the S4 system, whose performance is shown in the last row of Table 3. For bnmdev06, the CER reductions of the S4 system over the S2 system are smaller than those of the S3 system. This is because the proportion of the BN-style training data in the S4 system is reduced compared to the S3 system. However, even with a large increase in the absolute and relative amount of the BC-style data, the performance gain of discriminative training on bcmdev05 is small. The S4 system yielded a significant[6] gain of 0.4% on bcmdev05 for ML training compared to the S2 system, but an insignificant gain of 0.1% for MPE training. Unsupervised discriminative training on the additional data is clearly not effective for the BC-style test data. Therefore, simply adding more unlabelled data will not be further discussed. Instead, data processing strategies on the automatic transcriptions will be used to get improvements. Note that all experiments in the following sections will be based on the subset1 data set.

4.2. Data selection

To obtain larger improvements from discriminative training, data selection approaches can be used to filter out poorly recognised data (Wang et al., 2007; Yu and Gales, 2007). As indicated in Section 2.2, data selection may be performed based on the confidence scores. In this work, unless explicitly stated, the segment-level confidence score was used, as preliminary experiments yielded better performance than show-level data selection.

When performing confidence score based data selection, the selection threshold is normally empirically set. The impact of the thresholds on unsupervised training data selection performance was initially investigated. Four thresholds 0.7, 0.77, 0.8 and 0.86 were used. Fig. 3 shows the relative reduction in CER of unsupervised training with automatic transcriptions from supervised training. The x-axis shows the percentage of data retained after confidence score based data selection. The y-axis shows the percentage of the CER reduction from unsupervised training relative to supervised training, G_c. For a confidence score threshold c, G_c is calculated using

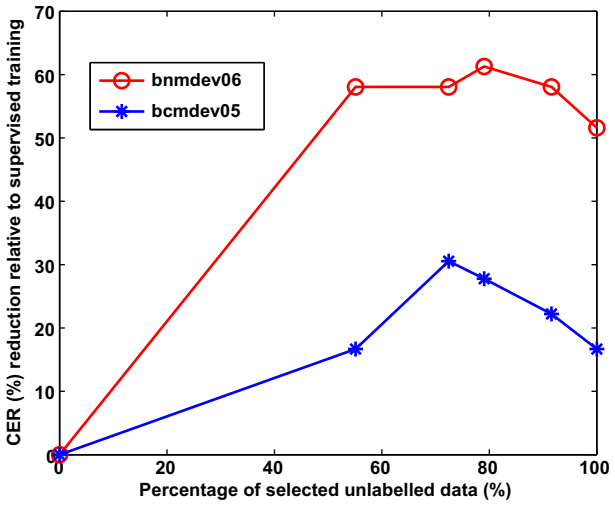

Fig. 3. Reduction in CER (%) from unsupervised MPE training with automatic transcriptions relative to that from supervised training.

$$G_c = 100\frac{\mathcal{E}_c - \mathcal{E}_S 0}{\mathcal{E}_S 1 - \mathcal{E}_S 0}, \tag{5}$$

where $\mathcal{E}_c$ is the CER of the model trained on unlabelled data selected using confidence score threshold c. $\mathcal{E}_S 1$ is the CER of supervised training and $\mathcal{E}_S 0$ is the performance of the baseline S0 model, which are shown in Table 3. This gives an idea of how close the unsupervised training performance gains are to the "ideal" supervised training gains.

From Fig. 3, for both BN-style and BC-style test sets, adding unlabelled data yielded reduction in CER over the baseline S0. However, the improvements did not always increase as more data were selected. This is expected because if too little data are selected, they will contribute less to model training; if too much data are selected, the quality of some of the selected data is poor. Either may limit the reduction in CER. A trade-off threshold should be empirically selected to balance these two issues. Furthermore, it can be observed that the relative reduction in CER on bnmdev06 is always greater (about double) than that on bcmdev05. As indicated in Section 2.2, there is always a large amount of BN data selected when confidence score based selection is used. Hence, the small reduction in CER with MPE training on bcmdev05 is also expected. The figure illustrates that the performance improvement on the data of mismatched genre is smaller for discriminative training.

Fig. 3 shows that using an appropriate confidence score can yield reduction in error rate for discriminative training. However, using confidence scores reduces the quantity of the training data. To investigate this issue, confidence scores were used to select data from the subset2 data set. The threshold used is the one corresponding to the best improvement on bcmdev05 in Fig. 3, 0.77. Experiments show that with the additional subset2 data, the MPE trained model can obtain a 0.5% absolute reduction in CER on bnmdev06 while no gain was obtained on bcmdev05. This is because simple data selection discards the

[6] Wherever the term "*significant*" is used in CER comparison, a pairwise significance test was done using NIST provided scoring toolkit sctk-1.2. The significance difference was reported using the Matched-Pair Sentence-Segment Word Error (MAPSSWE) test (Gillick and Cox, 1989) implemented at NIST (Pallett et al., 1990) at a significance level of 5%, or 95% confidence.

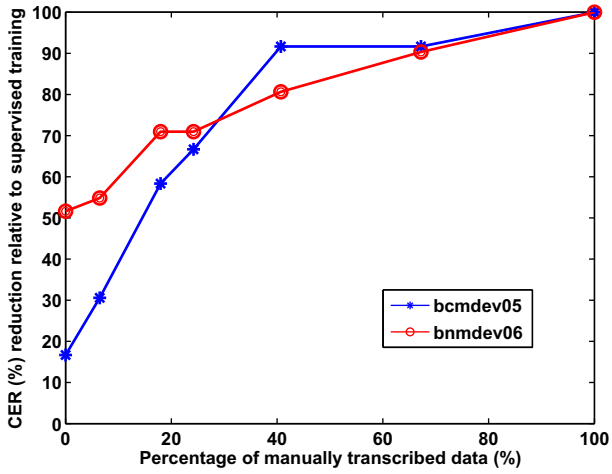

Fig. 4. Reduction in CER (%) from unsupervised MPE training with directed manual transcriptions relative to that from supervised training.

poorly recognised data, which is just the data expected to help most in improving the performance on the BC-style test data. This experiment shows that even with data selection, the use of only automatic transcriptions cannot yield substantial improvements on the test data of mismatched type. Hence, the alternative strategy of using directed manual transcription has been investigated.

4.3. Directed manual transcription

As discussed in Section 2.3, in order to improve the recognition performance on mismatched data, the data with low average confidence scores can be selected for manual transcription rather than discarded. This allows poorly transcribed data to be used. The selected data along with their manual transcriptions are then used together with the other automatically transcribed data for acoustic model training. Though the *directed manual transcription* (Yu and Gales, 2007) method requires some manual effort, it is still much less costly than manual transcription of the complete dataset. The aim is to find a reasonable trade-off between the increase in cost and the recognition performance. To find an appropriate operating point, different confidence score thresholds were used to select different amount of data to be manually transcribed.

Experiments show that BC data was always predominantly selected for all thresholds. This is consistent with Fig. 1(b) as BC data normally has lower confidence scores. Therefore, confidence score based selection implicitly performs a BC/BN segment selection. This implicit selection may contribute to the increased gain on the BC test set.

Fig. 4 gives the proportion of CER reduction from supervised MPE training when adding varying amount of manual transcriptions. The x-axis is the proportion of the unlabelled data that is manually transcribed. It can be seen that the reductions in CERs increase as more data is added at the cost of producing the additional manual transcriptions. The CER reduction on BC is larger than on BN when transcribing a small amount of data. This is because the confidence score based data selection tends to select BC data due to the poor performance. At the threshold of 0.77, transcribing 18% of the data can yield 58% of the complete supervised training improvement for BC data. This is believed to be a good trade-off between the cost and gain and hence is used for further experiments.

Given the confidence score threshold, there are several data selection strategies that could be used. Table 4 gives a comparison of unsupervised MPE training with different strategies. Only performance of MPE training is shown here as ML training is more robust.

Table 4 shows the results of a range of combinations of automatic and manual transcriptions of subset1. Using complete manual transcriptions (S1) can yield a 3.1% absolute reduction in CER on BN and a 3.6% absolute improvement on BC, compared to the baseline S0 system. Using automatic transcriptions for all the data, S2 yielded a 1.6% reduction in CER on BN and 0.6% on BC. The relative CER reductions from supervised training are shown in the last two columns. S2 yielded 52% relative for BN and only 17% on BC. The fourth row shows the performance using confidence score based data selection, which reduced the CER on both test sets. However, the relative improvement on BC is still poor. The sixth row shows the performance of directed manual transcriptions together with automatic transcriptions, which is referred to as S2c. By transcribing about 18% of the unlabelled data, the S2c system obtained a significantly larger reduction in CER on BC (2.1%). This is comparable to the improvement on BN (2.2%). The relative CER reductions compared to the ideal S1 system are also greatly improved. As a contrast, the fifth row shows the performance when only manual transcriptions were used. Compared to the S2c system, purely adding the data with manual transcriptions yielded significantly poorer performance on `bnmdev06` while better performance on `bcmdev05`. This is expected because the data with manual transcriptions are mostly BC data, while the data with automatic transcriptions are mostly BN data, hence, the model set is heavily tuned to the BC data. It is also clear that adding the selected automatic transcriptions is beneficial for the BN data, while for the BC data, it is just slightly better than the use of only manual transcriptions. This is because there was significantly less BC data in the selected automatic transcriptions due to the high error rate. Compared to the S1 system, 56% of the supervised training gain was obtained by only using the manually transcribed data. This illustrates that the poorly recognised data helps most on the mismatched type of data.

The above experiments showed the advantage of adding manual transcriptions for lower confidence score segments during unsupervised training. It is also interesting to know whether this selection approach is preferable to random selection.

Table 5 gives the comparison between confidence score based directed manual selection and random selection.

Table 4
Unadapted decoding performance (%) of unsupervised MPE systems trained on subset1 data with different amount of manually or automatically transcribed data.

Sys.	S0 + (h)		CER (%)		CER reduction rel. to supv. train (%)	
	Man.	Auto.	bnmdev06	bcmdev05	bnmdev06	bcmdev05
S0	0	0	13.6	25.4	–	–
S1	327.6	0	10.5	21.8	100	100
S2	0	318.1	12.0	24.8	52	17
–	0	251.3	11.7	24.4	62	28
–	58.9	0	12.7	23.4	29	56
S2c	58.9	251.3	11.4	23.3	71	58

Table 5
Unadapted decoding performance (%) of MPE trained system trained on subset1 data with directed manual selection and random selection.

Sys.	S0 + (h)		CER (%)	
	Man.	Auto.	bnmdev06	bcmdev05
S2c	58.9	251.3	11.4	23.3
Random (ALL)	58.9	248.7	11.7	23.9
Random (BC Only)	58.9	254.6	11.5	23.6

Table 6
Unadapted single-pass CER (%) of S2c MPE system. Both acoustic model and language model were rebuilt with directed manual transcription.

Update	bnmdev06	bcmdev05
AM	11.4	23.3
AM + LM	11.4	23.3

The second row in Table 5 is the random selection from all unlabelled data, which has similar amount of data as the S2c system in Table 4. This is equivalent to the automated data selection approach, but does not require any recognition on the selected data to be manually transcribed. It can be observed that using confidence score based selection significantly outperforms random selection on both test sets for MPE training. In particular, the improvement on bcmdev05 over random selection is larger than on bnmdev06 since directed selection favours BC data. The third row is the random selection only on the BC part of the unlabelled data, which ensures only BC data is selected. In this experiment, correct BC labels are assumed to be known in advance. In practice, this approach requires a BC/BN classification stage, which may lead to more errors. Even if the ideal BC labels were used, the confidence score based selection still obtained better results on both test sets for MPE training, though the improvements are smaller. This further demonstrates that the essential issue in unsupervised discriminative training is the high error rate transcriptions. Though the confidence score based selection approach is better than random selection, from Fig. 4, the curves have some portions with constant reduction in CER or small increase. This implies that there is room to further improve the confidence score based approach.

Table 7
Adapted decoding performance of unsupervised training and supervised training systems with adaptation.

Sys.	S0 + (h)		CER (%)	
	Man.	Auto.	bnmdev06	bcmdev05
S0	0	0	11.7	24.2
S1	327.6	0	9.4	20.4
S2	0	318.1	10.8	23.3
S2c	58.9	251.3	10.3	22.0

4.4. System refinement

The previous section has described the basic systems using directed manual transcriptions. This section will discuss some refinements.

One possibility is to use the additional transcriptions for language model estimation. The basic procedure is to build separate language model components for the manually transcribed data. Then the two newly estimated components are interpolated with the three language model components described in Section 4.1. Table 6 shows the performance of the S2c system, where unlabelled data with 58.9 h of directed manual transcriptions were incorporated for both acoustic and language model building.

From Table 6, rebuilding the language model did not give additional improvements on any of the test sets. Therefore, further system refinement discussions will only concentrate on acoustic model training.

4.4.1. Speaker adaptation

All previous experiments are unadapted single-pass decoding experiments. This section will investigate how discriminative training on untranscribed audio performs after adaptation. The two-pass adapted system described in Section 2.1 was used to test the adaptation performance.

From Table 7, after adaptation, the supervised training S1 system yielded a 2.3% absolute reduction in CER on bnmdev06 and 3.8% on bcmdev05 compared to the baseline S0 system. Using complete automatic transcription, the S2 system obtained 39% of the supervised training improvement on bnmdev06, while only 24% on bcmdev05. With 18% data manually transcribed, the S2c system yielded 60% of the CER reduction on bnmdev06 and 59% on bcmdev05. The relative improvement on the BC

Table 8
Comparison between show-level and segment-level directed data selections.

Conf. Level	S0 + (h)		Unadapted decoding		Adapted system	
	Man.	Auto.	bnmdev06	bcmdev05	bnmdev06	bcmdev05
Show	58.9	257.9	11.7	23.6	10.5	22.1
Segment	58.9	251.3	11.4	23.3	10.3	22.0

data is even larger than the relative improvement in unadapted single-pass decoding. This further illustrates the effectiveness of using directed manual transcription.

4.4.2. Level of data selection

The previous experiments on directed manual transcription are all based on segment-level confidence scores because the segment based selection is felt to be finer than show-level selection. However, in practice, it is not always convenient to manually transcribe non-contiguous segments. Manual transcriptions are often produced at the show level. It is therefore interesting to contrast the two levels of data selection.

Table 8 gives the comparison between show-level and segment-level selection of directed manual transcription. For the show level selection, a threshold of 0.80 was used to yield similar quantity of data for manually transcribed data as the S2c system (the second row). With a similar amount of manual transcriptions (58.9 hours), segment level selection outperforms show level selection on both test sets for MPE training with and without adaptation. Comparing show level selection to the baseline S0 performance in Table 7 and 4 shows that, for unadapted decoding performance, the proportion of the supervised training improvement are 61% for BN and 50% for BC; for adapted performance, the proportions are 52% for BN and 55% for BC. Those relative gains are still good and just slightly smaller than the segment level selection. This illustrates that show level selection can also be effective for discriminative training with directed manual transcription.

5. Conclusions

Sensitivity to transcription errors is an important issue in unsupervised discriminative training for LVCSR. A standard approach to deal with this issue is to only use automatic transcriptions of unlabelled data for discriminative training. With this approach, the performance of unsupervised discriminative training can be poor if the initial recognition system to generate the automatic transcription has too high error rate for particular data types. In this work, an alternative approach, discriminative training with directed manual transcription, is discussed in detail to address the problem. In this approach, a small amount of poorly transcribed data are manually transcribed to supplement the automatic transcription. The performance of unsupervised and directed manual transcription based MPE training were evaluated on a Mandarin transcription task, where both Broadcast Conversation (BC) and Broadcast News (BN) data were used. Experiments show that incorporating directed manual transcription can effectively improve the discriminatively trained system on the BC data compared to the traditional unsupervised approach in both unadapted and adapted decoding. With more data manually transcribed, the MPE improvements on both BC and BN became larger. A reasonable trade-off between the CER improvements and increased manual transcription cost can be obtained by performing confidence score based data selection with the confidence score distribution of the unlabelled data. It was shown that confidence score based data selection outperforms random data selection. Though segment-level confidence score yielded better performance, show-level confidence scores can also lead to reasonable improvements and is easier to use in practice.

References

Chan, H., Woodland, P., 2004. Improving broadcast news transcription by lightly supervised discriminative training. In: ICASSP, Montreal.
Cohn, D., Atlas, L., Ladner, R., 1994. Improving generalization with active learning. Machine Learning 15, 201–221.
Doumpiotis, V., Tsakalidis, S., Byrne, W., 2003. Lattice segmentation and minimum Bayes risk discriminative training. In: Proc. EuroSpeech.
Evermann, G., Woodland, P.C., 2003. Design of fast LVCSR systems. In: Proc. ASRU, St. Thomas.
Evermann, G., Chan, H.Y., Gales, M.J.F., Jia, B., Mrva, D., Woodland, P.C., Yu, K., 2005. Training LVCSR systems on thousands of hours of data. In: Proc. ICASSP, Philadelphia.
Gales, M.J.F., Jia, B., Liu, X., Sim, K.C., Woodland, P.C., Yu, K., 2005. Development of the CUHTK2004 Mandarin conversational telephone speech transcription system. In: Proc. ICASSP, Philadelphia.
Gillick, L., Cox, S.J., 1989. Some statistical issues in the comparison of speech recognition. In: Proc. ICASSP, Glasgow.
Kamm, T.M., Meyer, G.G.L., 2002. Selective sampling of training data for speech recognition. In: Proc. Human Language Technology, San Diego.
Kemp, T., Waibel, A., 1999. Unsupervised training of a speech recognizer: recent experiments. In: Proc. EuroSpeech, Budapest.
Kumar, N., 1997. Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition. Ph.D. Thesis, Johns Hopkins University.
Lamel, L., Gauvian, J.L., Adda, G., 2001. Unsupervised acoustic model training. In: Proc.ICASSP, Salt Lake City.
Lamel, L., Gauvian, J.L., Adda, G., 2002. Lightly supervised and unsupervised acoustic model training. Computer Speech and Language 16, 115–129.
Ma, J., Matsoukas, S., Kimball, O., Schwartz, R., 2006. Unsupervised training on large amount of broadcast news data. In: Proc. ICASSP, Toulouse.
Mangu, L., 2000. Finding consensus in speech recognition. Ph.D. Thesis, Johns Hopkins University.
Nakamura, M., Iwanoa, K., Furui, S., 2007. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. Computer Speech and Language 22 (2), 171–184.

Pallett, D.S., Fisher, W.M., Fiscus, J.G., 1990. Tools for the analysis of benchmark speech recognition tests. In: Proc. ICASSP.

Povey, D., 2003. Discriminative training for large vocabulary speech recognition. Ph.D. Thesis, Cambridge University.

Povey, D., Woodland, P.C., 2002. Minimum phone error and I – smoothing for improved discriminative training. In: Proc.ICASSP, Orlando.

Riccardi, G., Hakkani-Tur, D., 2003. Active and unsupervised learning for automatic speech recognition. In: Proc. EUROSPEECH, Geneva.

Sinha, R., Gales, M.J.F., Kim, D.Y., Liu, X.A., Sim, K.C., Woodland, P., 2006. The CU-HTK Mandarin broadcast news transcription system. In: Proc. ICASSP, Toulouse.

Wang, L., Gales, M.J.F., Woodland, P.C., 2007. Unsupervised training for Mandarin broadcast news and conversation transcription. In: Proc. ICASSP, Honolulu.

Wessel, F., Ney, H., 2005. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing 13 (1), 23–31.

Woodland, P.C., Odell, J.J., Valtchev, V., Young, S.J., 1995. The development of the 1994 HTK large vocabulary speech recognition system. In: ARPA Workshop on Spoken Language Systems Technology, pp.104–109.

Woodland, P.C., Gales, M.J.F., Pye, D., Young, S.J., 1997. The development of the1996 HTK broadcast news transcription system. In: DARPA Speech Recognition Workshop, pp.73–78.

Yu, K., Gales, M.J.F., Woodland, P.C., 2007. Unsupervised training with directed manual transcription for recognising Mandarin broadcast audio. In: Proc. INTERSPEECH, Antwerp.