PERVASIVE UNSUPERVISED ADAPTATION FOR LECTURE SPEECH TRANSCRIPTION

Daniel Willett[†]*, Thomas Niesler*[*]*, Erik McDermott, Yasuhiro Minami, Shigeru Katagiri*

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
[*]Electrical Engineering Department, University of Stellenbosch, South Africa
{willett,mcd,minami,katagiri}@cslab.kecl.ntt.co.jp, trn@dsp.sun.ac.za

ABSTRACT

Unsupervised adaptation has evolved as a popular approach for tuning the acoustic models of speaker-independent speech recognition systems to specific speakers, speaker groups or channel conditions while making use of only untranscribed data. This study focuses on procedures for unsupervised adaptation of other probabilistic models that are involved in state-of-the-art speech recognizers and on the joint adaptation of multiple knowledge sources. In particular, we outline and evaluate approaches for adapting both the language model and the pronunciation model (lexicon) without supervision. Initial experiments on off-line lecture speech transcription achieved small but promising word error rate improvements with each approach applied separately. The experimental results on the joint application of acoustic, language and pronunciation model adaptation indicate that the individually achievable performance improvements are additive.

1. INTRODUCTION

In order to decode a given utterance X into a word sequence W according to the MAP decision rule

$$W^* = \arg\max_W P(W|X) \quad (1)$$

state-of-the-art speech recognizers combine several knowledge sources, represented as probabilistic models, such as language model (p_{lm}), pronunciation model (p_{pm}) and acoustic model (p_{am}). The posterior probability above (with Viterbi approximation selecting the best pronunciation) is then expressed by

$$W^* \approx \arg\max_W (p_{lm}(W) \max_Q p_{pm}(Q|W) p_{am}(X|Q)) \quad (2)$$

The probabilistic models are most commonly set up to optimize a Maximum Likelihood criterion. This means that the model parameters are chosen in a way that the models yield maximum training data likelihood. For a general purpose recognizer, the training data is chosen to represent a broad variety of speaking styles, speakers, channel conditions and topics. The resulting models offer a stable performance over various speakers, speaking styles, topics and channel conditions. It is well known, however, that more specific models, such as speaker-dependent acoustic models, topic-specific language models or pronunciation models, that explicitly represent a certain speaking style, outperform more general models in terms of recognition accuracy whenever these specific conditions are met. Various adaptation techniques have been proposed to specialize the probabilistic models to specific speakers, topics and speaking styles in supervised procedures that use labeled data that represents these specific conditions. Supervised adaptation has successfully been applied not only for acoustic models but for all the three types of probabilistic models as mentioned above [3, 1, 2]. Again, Maximum Likelihood (ML) is usually chosen as the optimization criterion.

[†]now with Temic Speech Processing, Ulm, Germany

2. UNSUPERVISED ADAPTATION

In many speech recognition applications though, there is no prior knowledge about the speaker, speaking style or the topic at hand. However, even in the scenario in which supervised adaptation as described above cannot be applied, speaker, speaking style, topic and channel can be assumed to stay constant over some period of time (sentence, utterance, speech). This enables the use of unsupervised adaptation in the hope of improving recognition accuracy on this particular part of the acoustic observation. The length of this constant period largely depends on the application and the type of modeling component. In the task that is being targeted here for example, an utterance of at least 10 minutes can be assumed to be spoken by a single speaker about a single topic with only moderately varying speaking style and channel condition.

The idea of unsupervised adaptation is to exploit these slowly varying or constant characteristics to tune the models for improved recognition performance on the portion of the data during which a given speech characteristic is assumed not to change. Usually, this is performed in a two pass procedure in which adaptation is performed with respect to an automatically derived transcription of the data. This is illustrated in Figure 1. First, the unadapted baseline models are applied to automatically transcribe the data and then the models are adapted with respect to this transcription. Possibly, this process is iterated a number of times in order to exploit the adapted models for a better transcription in a succeeding iteration. Despite the errors in the automatically generated transcription, unsupervised adaptation has successfully been applied to the acoustic models in several studies [11, 9, 6]. Most commonly, the Maximum Likelihood Linear Regression (MLLR) adaptation approach [3] is applied. Small performance improvements are obtained from making use of word-based confidence measures for excluding unreliably recognized data [6].

The target application of this study is automatic lecture speech transcription. In lecture speeches, a single person speaks about a specific topic in a specific style. In order to investigate whether this amount of data on a specific topic and in a specific speaking style allows successful unsupervised adaptation of language model [5] and pronunciation model [10] we performed a series of exper-

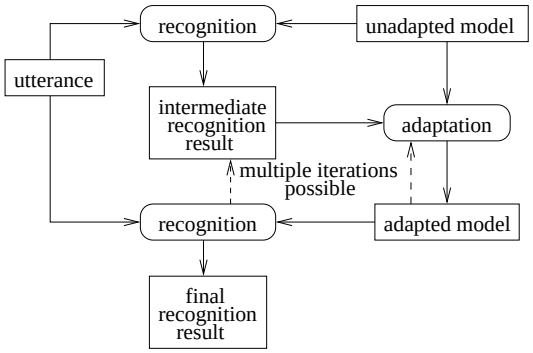

Fig. 1. Principle of unsupervised model adaptation

iments. The following sections summarize these approaches.

2.1. Language Model

The question of how to set up a useful special purpose language model from only a small special purpose text corpus while integrating additional knowledge from a large general purpose text corpus is of major practical relevance. Therefore, supervised language model adaptation has been the subject of several studies [4, 2]. Generally, the idea is to set up a language model that has the broad knowledge about word dependencies from the large corpus, but still favors those words and word combinations seen on the small, more relevant corpus. This is achieved by some process of adaptation of the general purpose language model on the small relevant corpus. The approaches that we chose for evaluating the usefulness of such procedures in unsupervised language model adaptation are text selection, inspired by [4], and Minimum Discriminant Estimation (MDE), a model interpolation procedure proposed in [2]. Furthermore, we also combined the two approaches into a joint adaptation procedure using both text selection and MDE-adaptation. This work was presented in greater detail in [5].

2.1.1. Text Selection and Linear Interpolation

The idea of adaptation by text selection is to select a subset of the language model training data, that is similar in topic and style to the text to adapt to. For choosing similar texts, Various similarity measures can be used to compare texts. We used a $term\ frequency\ inverse\ document\ frequency$ (tf-idf) based measure. On the selected similar texts, a new language model is established. The final adapted language model is obtained by a linear combination of this specialized language model and the broader baseline language model, which is assumed to contain better likelihood estimates for infrequent and less topic-specific words:

$$p_{adapt}(w|h) = \alpha * p_{spec}(w|h) + (1-\alpha) * p_{base}(w|h) \quad (3)$$

Here, w is a word and h an arbitrary context (word history). In this linear combination, the factor α is determined in an EM procedure with the optimization goal of minimizing the perplexity of the adaptation text. In our case of unsupervised adaptation, this adaptation text is the first pass recognizer output.

2.1.2. Minimum Discriminant Estimation

The idea behind Minimum Discriminant Estimation (MDE) adaptation [2] is the assumption that the uni-gram p_{uni} of the adaptation text is a rather good model for the real uni-gram of the text to be recognized, while for context-dependent word likelihoods, it is better to rely on the model p_{base} trained on more, but less relevant data. Hence, the resulting adapted language model p_{adapt} should have the uni-gram p_{uni} of the adaptation text as its marginal distribution according to

$$\sum_h p_{adapt}(h)p_{adapt}(w|h) = p_{uni}(w) \quad (4)$$

and among those models fulfilling this equation, MDE chooses the one which is the closest to the baseline language model p_{base}:

$$p_{adapt}(w|h) = \underset{p(\cdot|\cdot)}{\operatorname{argmin}} \sum_h p(h) D(p(\cdot|h)||p_{base}(\cdot|h)) \quad (5)$$

The effect of MDE adaptation in the unsupervised adaptation framework is a shift of the language model towards words that are actually found in the first pass output. This exploits the observation that seeing a word uttered at some place within the speech increases the likelihood of an additional appearance.

2.2. Pronunciation Rules

The common practice in state-of-the-art speech recognizers (for Japanese) is to apply a rather simple pronunciation model with a single fixed pronunciation (HMM sequence) per word.

Approaches introducing more complex pronunciation models have been proposed, but their success has been rather limited. This is particularly true for speaker- and speaking-style independent speech recognition, where pronunciation rules seem to improve the acoustic models as much as they decrease discrimination among words because of multiple pronunciation variants. For an overview and further references on pronunciation adaptation consult [8]. In speaker-dependent Japanese speech recognition, pronunciation variants have successfully been made use of in [1].

An approach and first experiments on unsupervised adaptation of the pronunciation rules when specializing the recognizer for a specific speaker were presented in [10]. The idea is to refrain from using multiple pronunciations in the general speaker- and speaking style-independent recognizer, as this was found to worsen recognition performance, but instead to introduce pronunciation rules trained on the recognizer output of the first pass in a second recognizer pass. Figure 2 illustrates this procedure. Using the first pass output, a couple of pronunciation variant rules get expanded and the best matching model sequence is recalculated. Based on this sequence, Maximum Likelihood estimates of the (context-dependent) likelihood of each rule application can be obtained. Those estimates are then used to weight the pronunciation rules in a full second recognition pass. Refer to [10] for further details. The pronunciation variant rules used in this study are very broad rules for skipping short vowels, lengthening short vowels, shortening long vowels, voicing of unvoiced consonants and the unvoiced articulation of voiced consonants.

2.3. Acoustic Model

Unsupervised acoustic model adaptation has been utilized in this study for completeness and for evaluating whether the performance improvement is additive to the gain achieved by adapting language model and pronunciation model. We used HTK's [12] implementation of the MLLR framework with 250 regression classes while applying linear transformations on the Gaussians'

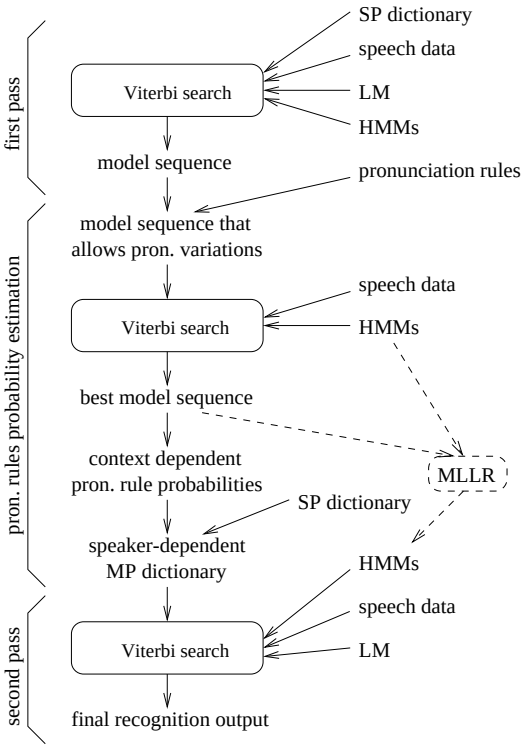

Fig. 2. Unsupervised pronunciation model adaptation

mean vectors only. When performed in conjunction with pronunciation adaptation, we applied acoustic model adaptation based on the realigned model sequence as outlined by the dashed lines in Figure 2.

3. EXPERIMENTS AND RESULTS

The task we are working on is the automatic transcription of recorded Japanese lecture speeches. The speech data and its transcription is provided within the Spontaneous Speech Program [7]. On this task we assume speaker, topic and speaking style not to change over an entire lecture.

A series of experiments was performed to investigate the usefulness of unsupervised adaptation of both language and pronunciation model on this task. We set aside two separate test sets of 7 speeches each, one for development and tuning of the adaptation procedures, one for evaluating them. From the available 158 speeches, this left 144 for language model estimation and for training the context-dependent acoustic models. It should be noted that this is very little data for the purpose of language model estimation. The transcriptions of the 144 speeches comprise to about 413K words. In all experiments we made use of the same closed vocabulary which contains all distinct words of training and test sets (approximately 13K words). The acoustic models are tree-clustered gender-dependent cross-word triphones of 2000 physical states (per gender). These models have originally been set up on read speech. For this study, their structure was kept and their parameters were simply retrained using the 144 training speeches.

3.1. Language Model Adaptation

We evaluated the effect of the two procedures, text selection and MDE, separately on the development set (see Table 1). They both yield some improvement. The best result of 2.3% (1.2%) of absolute word error rate reduction on the development (evaluation) set was achieved when combining the two procedures and applying them in multiple iterations.

configuration	dev-set	eval-set
baseline (unadapted)	33.5%	36.9%
adapted by text selection	32.8%	
adapted by MDE	31.7%	
adapted by text sel. & MDE	31.2%	35.7%

Table 1. Unsupervised language model adaptation

As described above, the text database used for language model training and adaptation comprises only the transcription of 144 training speeches. Thus, it is of very limited size with little variation in topic and speaking style compared to broader text corpora as all lectures were recorded at speech and language related conferences. In this respect, we see the achieved improvements as an encouraging result, and expect greater gains on larger, more diverse text corpora.

3.2. Pronunciation Adaptation

In initial experiments on pronunciation modeling, we first evaluated the effect of introducing pronunciation variants and training of context-dependent variant probabilities in the speaker-independent case. With the very general pronunciation variations like dropping of a short vowel or voicing of an unvoiced consonant and training context-dependent ML-weightings for these cases, we measured a severe decrease in performance compared to the simple single pronunciation case (see Table 2). In the speaker-independent (SI) case, new pronunciation variants seem to introduce more additional confusability than they help discriminating words. In the

configuration	dev-set	eval-set
baseline (single pron)	33.5%	36.9%
SI trained pronunciation rules	36.3%	40.1%
SD unsuperv. adapted prons.	33.2%	36.5%

Table 2. Pronunciation model adaptation

case where the pronunciation variants are only introduced when specializing the recognizer to a specific speaker (without supervision), however, we gain a small, but significant word error rate improvement. This approach is what we refer to as unsupervised pronunciation adaptation.

3.3. Acoustic Model Adaptation

Table 3 lists the word error rate improvements that we gain from performing two iterations of unsupervised acoustic model adaptation (MLLR on the 1-best output, 250 Gaussian mean transforms).

configuration	dev-set	eval-set
baseline	33.5%	36.9%
unsupervised acoustic model adaptation	28.6%	31.8%

Table 3. Acoustic model adaptation (MLLR)

3.4. Pervasive model adaptation

The goal of this research is a unified unsupervised adaptation approach that combines acoustic model, language model and pronunciation model adaptation. In first experiments on this kind of pervasive adaptation we integrated pronunciation model adaptation and acoustic model adaptation into a joint adaptation procedure as indicated in Figure 2. Acoustic model adaptation is performed with reference to the regenerated model sequence that allows for pronunciation variations. For the sake of simplicity, language model adaptation is performed in a separate procedure. Hence, the joint adaptation is accomplished in two separate adaptation steps. In our experiments, language model adaptation is performed first, and acoustic and pronunciation model adaptation afterwards taking into account the improved transcription corrected through language model adaptation.

Initial experiments on combining acoustic model adaptation (MLLR) and pronunciation adaptation yield an absolute improvement of 5.3% on the development set. Concluding from the achieved absolute error rate improvements of 4.9% and 0.3% when applying the adaptation approaches separately, it appears that the individual unsupervised adaptation approaches do not compensate, but that the achieved improvements seem to sum up. The evaluation set shows a similar tendency. Table 4 summarizes the experimental results on combining the various adaptation approaches. The additive tendency can also be observed when performing acoustic model adaptation or combined acoustic and pronunciation model adaptation after language model adaptation.

configuration	dev-set	eval-set
baseline (no adaptation)	33.5%	36.9%
unsupervised acoustic model adaptation (as in Table 3)	28.6%	31.8%
integrated unsupervised acoustic and pronunciation model adaptation	28.2%	31.5%
unsupervised language model adaptation (as in Table 1)	31.2%	35.7%
unsupervised l.m. adaptation first, acoustic model adaptation afterwards	26.9%	30.8%
unsupervised l.m. adaptation first, then integrated pron. + ac. model adaptation	26.6%	30.6%

Table 4. Combined unsupervised adaptation

Overall, the pervasive adaptation approach gives us an absolute improvement of 2.0% (1.2%) on dev-set (eval-set) over pure acoustic model adaptation. This can be regarded as a major improvement. Further improvement should be achievable by a tighter integration of language model adaptation and acoustic and pronunciation model adaptation, which could simply be a second language model adaptation step after acoustic and pronunciation model adaptation. Also, for more diverse language model training data and other languages with more pronunciation diversity than Japanese, even more significant improvements might be achievable. This, however, remains to be evaluated.

4. CONCLUSION

We have described our approach to pervasive unsupervised adaptation that includes language model and pronunciation model adaptation and combines these with unsupervised acoustic model adaptation. In experiments on lecture speech transcription, we achieved small but significant performance improvements with isolated language model adaptation and pronunciation model adaptation procedures. First experiments that combine the three types of unsupervised adaptation showed a promising performance, in which the individual improvement add up rather than compensate.

Acknowledgment. We thank the Japanese Science and Technology Agency Program "Spontaneous Speech" for providing speech data and transcriptions.

5. REFERENCES

[1] T. Imai, A. Ando, E. Miyasaka, "A New Method for Automatic Generation of Speaker-Dependent Phonological Rules", ICASSP, Detroit, 1995, pp. 864–867.

[2] R. Kneser, J. Peters. D. Klakow, "Language model adaptation using dynamic marginals", Eurospeech, Rhodos, 1997, pp. 1971–1974.

[3] C. J. Leggetter, P. C. Woodland, "Flexible Speaker Adaptation using Maximum Likelihood Linear Regression", ARPA Spoken Language Technology Workshop, 1995, pp. 104-109.

[4] M. Mahajan, D. Beeferman, X.D. Huang, "Improved topic-dependent language modelling using information retrieval techniques", ICASSP, Phoenix, 1999, pp. 541–544.

[5] T. Niesler, D. Willett, "Unsupervised Language Model Adaptation for Lecture Speech Transcription", ICSLP, Denver, 2002, pp. 1413–1416.

[6] M. Pitz, F. Wessel, H. Ney, "Improved MLLR Speaker Adaptation using Confidence Measures for Conversational Speech Recognition", ICSLP, Beijing, 2000, pp. 548–551.

[7] T. Shinozaki, C. Hori, S. Furui, "Toward Automatic Transcription of Spontaneous Speech", Eurospeech, Aalborg, 2001, pp. 491–494.

[8] H. Strik, "Pronunciation adaptation at the lexical level", ISCA Adaptation Workshop, Sophia-Antipolis, France, 2001, pp. 123–131.

[9] F. Wallhoff, D. Willett, G. Rigoll, "Frame Discriminative and Confidence-Driven Adaptation for LVCSR", ICASSP, Istanbul, 2000, pp. 1835–1838.

[10] D. Willett, E. McDermott, S. Katagiri, "Unsupervised Pronunciation Adaptation for Off-Line Transcription of Japanese Lecture Speeches", ISCA PMLA Workshop, Estes Park, 2002.

[11] P. C. Woodland, D. Pye, M. J. F. Gales, "Iterative Unsupervised Adaptation using Maximum Likelihood Linear Regression", ICSLP, Philadelphia, 1996, pp. 1133–1136.

[12] S. Young et al., "The HTK Book, Ver. 3.1", Cambridge University Engineering Department, 2001.