An Active Approach to Speaker and Task Adaptation based on Automatic Analysis of Vocabulary Confusability

Qiang HUO and Wei LI

Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong
(Email: qhuo@cs.hku.hk, pwli@cs.hku.hk)

Abstract

Speaker and task adaptation can be made more efficient if an automatic speech recognition system can actively elicit particularly useful adaptation data from a new speaker for a given speech recognition task. This paper presents such an active approach based on an automatic analysis of how difficult the given task vocabulary is. Comparative experiments are designed and conducted for a simple application scenario of searching an item from a long list via voice. The experimental results demonstrate that the proposed active adaptation strategy performs much better than traditional passive adaptation strategies.

Index Terms: speech recognition, active learning, speaker adaptation, task adaptation, hidden Markov model.

10.21437/Interspeech.2007-127

1. Introduction

As reviewed in [7], over the years, many good techniques have been developed to adapt a set of pre-trained speaker-independent (SI) and/or task-independent (TI) hidden Markov models (HMMs) in an automatic speech recognition (ASR) system by using certain amount of adaptation data from a new speaker/task for achieving a high recognition performance. Obviously, the less adaptation sentences the speaker has to read the better. For supervised speaker/task adaptation, it is possible to improve adaptation efficiency via a good design of adaptation sentences. Traditional approaches were designed, disregarding the confusability of the task vocabulary, to select a minimum set of phonetically rich sentences from a large text corpus to achieve a good coverage of basic speech units adopted in the HMM-based ASR system being adapted (e.g., [9, 1]). In this paper, we propose an *active* approach to address the above problem based on an automatic analysis of how difficult the given task vocabulary is.

The idea of *active learning* has been studied extensively in the field of machine learning in the past two decades (e.g., references in [2]), and applied to different applications, including several spoken language processing applications (e.g., [2, 4, 6]). Given large amount of unlabeled training data, the previous attempts of active learning aim mainly at reducing (ideally minimizing) manual annotation efforts by intelligently *selecting* only a subset of training data for manual labeling that are most useful for learning purposes of building up the application at hand. Our proposed approach extends the concept of *active learning* to the scenario of supervised speaker and task adaptation, where the system takes the initiative of *eliciting* small (ideally minimum) amount of adaptation data from the user for achieving high (ideally maximum) performance improvement by adapting the HMMs using the elicited adaptation data. It is noted that different from the traditional active learning approaches, the transcriptions of the adaptation sentences are known in this case, which can be derived from adaptation scripts prompted to the user.

In the following, we first describe our proposed approach in Section 2. Then we present experimental results in Section 3. Finally, we conclude the paper in Section 4.

2. Active Adaptation Approach

2.1. Application Scenario and Overall Formulation

For ease of discussions, let's assume that we are dealing with an ASR application scenario (aka, Voice Lookup), in which a user can say a tag name in an inventory to obtain the detailed information of the item associated with the above tag (e.g., name dialing, song/music retrieval via voice, etc.). The set of the above tag names forms a *task vocabulary*. Let's further assume that our ASR system is equipped initially with a set of SI and TI acoustic models corresponding to a set of basic speech units (e.g., triphones), each being a left-to-right (LR) continuous density HMM (CDHMM) with a mixture of Gaussians as an observation probability density function (PDF) for each state.

Given the above set of CDHMMs and a task vocabulary, a specific pronunciation of a vocabulary word can be represented as a sequence of state-dependent Gaussian mixture models (GMMs). Therefore, the problem of measuring the confusability of two vocabulary words can be cast as a problem of measuring the dissimilarity of two sequences of GMMs with possibly different lengths. Such a problem can be solved efficiently by using the technique we developed in [3]. Given the confusability scores for all pairs of vocabulary words, an appropriate procedure can be used to identify subsets of confusable words which may cause recognition errors, from which a probability distribution of the relevant basic speech units (referred to as *coverage units* hereinafter) can be measured. Then an efficient algorithm will be used to select a set of adaptation sentences from a back-end text corpus (task vocabulary itself in our case) to achieve a good coverage of the *coverage units* identified in the above sets of confusable words. Finally, the user will be prompted to speak the above adaptation sentences one by one, and an appropriate adaptation algorithm will be used to derive a set of speaker- and task-adapted (SA and TA) CDHMMs accordingly.

Fig. 1 illustrates an overall architecture of the above active adaptation approach. In the following two subsections, we elaborate on the two critical components, namely *automatic analysis of task vocabulary* and *generation of adaptation script*, respectively.

This research was supported by a grant from HKU's Seed Funding Programme for Basic Research. Dr. Donglai Zhu's contribution in training the triphone models used in this study is also gratefully acknowledged.

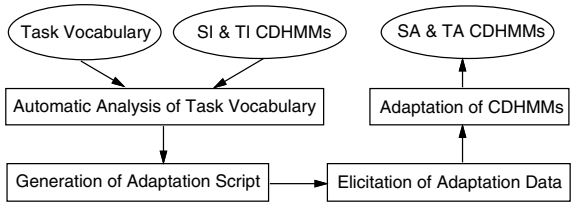

Figure 1: Overall architecture of the active approach to speaker and task adaptation.

2.2. Automatic Analysis of Task Vocabulary

Let's use $V = \{v_i | i = 1, 2, \cdots, N_v\}$ to denote the set of words in the task vocabulary, V_c to denote the subset of confusable words in V, V_n to denote the difference of V and V_c, and $U = \{u_i | i = 1, 2, \cdots, N_u\}$ to denote the set of *coverage units* which can be either *triphones* or *tied HMM states* in this study. Given a specific pronunciation of a word v_i, it can be represented as a sequence of *coverage units*, and modeled by an LR-CDHMM λ_i. Let's further define a symmetric dissimilarity score between words v_i and v_j as follows:

$$SIM(v_i, v_j) = [D_{HMM}(\lambda_i, \lambda_j) + D_{HMM}(\lambda_j, \lambda_i)]/2, \quad (1)$$

where $D_{HMM}(\lambda_i, \lambda_j)$ is an asymmetric dissimilarity measure between λ_i and λ_j as defined in [3]. The automatic analysis of task vocabulary is then performed according to the following procedure:

Step 1: Initialization
Initialize V_c as an empty set. Set $C(u_i) = 0$ for $i = 1, 2, \cdots, N_u$.

Step 2: Identification of Confusable Words
For each pair of vocabulary words (v_i, v_j), calculate $SIM(v_i, v_j)$. If $SIM(v_i, v_j) < TH_c$, do

- Update $V_c = V_c \cup \{v_i, v_j\}$;
- Derive, by using dynamic programming, an alignment between two strings of *coverage units* corresponding to v_i and v_j respectively. For each pair of symbols on the optimal "matching path" with different *coverage units*, say (u_k, u_l), do
  - $C(u_k) = C(u_k) + 1$;
  - $C(u_l) = C(u_l) + 1$.

Step 3: Termination

- Calculate the target probability distribution of coverage units as follows:

$$P_t(u_i) = \frac{C(u_i)}{\sum_{j=1}^{N_u} C(u_j)} ; \quad (2)$$

- Set $V_n = V - V_c$.

Consequently, we obtain the set of confusable words, V_c, and a probability distribution of coverage units, $\{P_t(\cdot)\}$, which are used to form the objective functions for our adaptation script generation procedure described in the next subsection. It is noted that a key technical issue in the above procedure is how to set the threshold TH_c. In our experiments, it is set empirically. How to set this control parameter automatically for a given task is an important topic for future research.

2.3. Generation of Adaptation Script

Given a set of words, S, let's use $U(S)$ to denote the set of *coverage units* in S, and $\{P(S, u) | u \in U(S) \cap U(V_c)\}$ to denote the probability distribution of the *coverage units* in both S and V_c. We can then define the following two *coverage scores* for S [10, 5, 1]:

$$CS_1(S) = \sum_{u \in U(S)} P_t(u) \quad (3)$$

$$CS_2(S) = -\sum_{u \in U(S) \cap U(V_c)} P(S, u) \log \frac{P(S, u)}{P_t(u)} . \quad (4)$$

Given V_c, V_n, and $\{P_t(\cdot)\}$, an appropriate procedure can be designed to generate a set of adaptation words, S_a, with a total number of syllables N_{syl}, which satisfies the following conditions:

$$S_a = \begin{cases} \underset{S}{\operatorname{argmax}}\, CS_1(S) & \text{if not all } \textit{coverage units} \text{ with} \\ & \text{nonzero probability in } \{P_t(\cdot)\} \\ & \text{are covered by } S; \\ \underset{S}{\operatorname{argmax}}\, CS_2(S) & \text{otherwise.} \end{cases}$$

Among many possibilities, the following procedure is used to achieve the above objective approximately in this study:

Step 1: Initialization
Set S_a as empty set.

Step 2: Incremental Generation of Initial Adaptation Script
If the total number of syllables in S_a is less than N_{syl} and not all the words in V have been examined, repeat the following actions:

- Evaluate the coverage score of S_a as
  - $CS(S_a) = CS_1(S_a)$, if not all *coverage units* with nonzero probability in $\{P_t(\cdot)\}$ are covered by S_a;
  - $CS(S_a) = CS_2(S_a)$, otherwise.
- Select randomly an unexamined candidate word v from
  - V_c, if it has not been exhausted;
  - V_n, otherwise;

  and set $S_{temp} = S_a \cup \{v\}$.
- Evaluate the coverage score of S_{temp} as
  - $CS(S_{temp}) = CS_1(S_{temp})$, if not all *coverage units* with nonzero probability in $\{P_t(\cdot)\}$ are covered by S_a;
  - $CS(S_{temp}) = CS_2(S_{temp})$, otherwise.
- Update $S_a = S_{temp}$, if $CS(S_{temp}) > CS(S_a)$; skip, otherwise.

Step 3: Refining the Adaptation Script
Let's use $S_a(V_c)$ to denote the set of words in S_a which are selected from V_c, and $S_a(V_n)$ to denote the set of words in S_a which are selected from V_n, respectively. If not all the words which are in V but not in S_a have been examined, repeat the following actions:

- If the total number of syllables in S_a is less than N_{syl}, go to **Step 2**.
- If $S_a(V_c) = V_c$, do

- Select randomly an unexamined candidate word v from V_n which is not in S_a.
- identify the word t in $S_a(V_n)$ such that

$$t = \arg \max_{w \in S_a(V_n)} CS_2(S_{sub}(w, v)) \ ,$$

where $S_{sub}(w, v)$ represents the result of substituting w in S_a by v, i.e., $S_{sub}(w, v) = (S_a - \{w\}) \cup \{v\}$.
- Update $S_a = S_{sub}(t, v)$, if $CS_2(S_{sub}(t, v)) > CS_2(S_a)$; skip, otherwise.

- If $S_a(V_c) \neq V_c$, do
  - Evaluate the coverage score of S_a as
    - $CS(S_a) = CS_1(S_a)$, if not all *coverage units* with nonzero probability in $\{P_t(\cdot)\}$ are covered by S_a;
    - $CS(S_a) = CS_2(S_a)$, otherwise.
  - Select randomly an unexamined candidate word v which is not in S_a from
    - V_c, if it has not been exhausted;
    - V_n, otherwise.
  - Identify the word t in S_a such that

$$t = \arg \max_{w \in S_a} CS(S_{sub}(w, v))$$

where
    - $CS(S_{sub}(w, v)) = CS_1(S_{sub}(w, v))$, if not all *coverage units* with nonzero probability in $\{P_t(\cdot)\}$ are covered by S_a;
    - $CS(S_{sub}(w, v)) = CS_2(S_{sub}(w, v))$, otherwise.
  - Update $S_a = S_{sub}(t, v)$, if $CS(S_{sub}(t, v)) > CS(S_a)$; skip, otherwise.

By using the above procedure, we can generate a set of adaptation words, S_a, with about N_{syl} syllables. The user will then be prompted to speak the above adaptation words one by one, and the elicited data will be used by an appropriate adaptation algorithm to derive a set of SA and TA CDHMMs accordingly.

3. Experiments and Results

3.1. Experimental Setup

In order to verify the effectiveness of the proposed active adaptation approach, we choose a simple Putonghua (Mandarin Chinese) 'Voice Lookup' application as our task. The basic speech units of our ASR system are triphones considering both the within-syllable and cross-syllable contextual dependencies. The context-independent (CI) phone set consists of 36 phones plus silence. Each triphone is modeled by a three-emitting-state LR-CDHMM without state skipping. Each state has 8 Gaussian mixture components with each component having a diagonal covariance matrix. A special three-state CDHMM is also used for silence modeling. The 39-dimensional feature vector used in this study consists of 12 MFCC's and log-scaled energy normalized by the peak of the individual sentence, plus their first and second order derivatives. Sentence-based cepstral mean subtraction is applied for acoustic normalization both in training and testing. HTK toolkit [11] is used to train a set of speaker-independent (SI) decision-tree-based tied-state triphone CDHMMs with 3001 tied states in total. The training data consists of 153796 sentences (about 180 hours speech) from 216 male and 114 female speakers extracted from four task-independent (TI) Putonghua corpora, namely HKU96, HKU99, 863, and MSRA, respectively.

Three task-dependent (TD) vocabularies with a fixed size of 202 words are designed and used for our experiments. Each vocabulary consists of two parts. The first part consists of a number of Chinese names (i.e., 2- or 3-syllable words), which are divided into a number of confusable-word groups. Each confusable-word group may have 2 to 10 confusable words, which are designed manually. Words within the same confusable-word group have the same number of syllables, and differ by only one syllable at the same position, namely, the second syllable for 2-syllable words, and the second or third syllable for 3-syllable words. Those syllables which differentiate different words are selected from different subsets of confusable syllables derived from a syllable recognition confusion matrix as detailed in [3]. The second part of each TD vocabulary consists of a number of Chinese words, which are extracted automatically from a large Chinese-word lexicon such that they are most distinct from those Chinese names in the first part of the vocabulary according to the dissimilarity scores of the corresponding word CDHMMs. It is noted that the *matching-based approximation* method [3] is used for measuring the dissimilarity of two CDHMMs in this study. The three 202-word TD vocabularies with 52, 102, 152 confusable words are designed accordingly and hereinafter referred to as *TD-V1*, *TD-V2*, and *TD-V3*, respectively. The ASR task is then the recognition of isolated words in each TD vocabulary. The relevant recognition engine in HTK is used for all the recognition experiments.

Given the above TD vocabularies, 5 male and 5 female native Putonghua speakers are recruited to read each of the vocabulary words 5 to 6 times. Among them, one utterance per word is reserved for designing adaptation experiments, and the rest of utterances are used as testing data. Consequently, there are in total 9898 testing utterances for each task. All speech recordings are made in an office room with a single close-talking microphone. Speech is digitized at 16-bit accuracy and with a sampling rate of 16KHz.

3.2. Adaptation Experiments and Results

Given the above set of SI&TI CDHMMs and for each TD vocabulary, several sets of adaptation scripts with different number of syllables can be generated to achieve a good coverage of *coverage units* by using the approach described in Section 2. Two types of adaptation scripts are generated by using *triphones* and *tied states of CDHMMs* as *coverage units* respectively. A fixed threshold, $TH_c = 9$, is used in all experiments. To simulate the effect of a traditional approach for the generation of adaptation script, corresponding sets of adaptation scripts are also generated to achieve a good coverage of triphones as distributed in each TD vocabulary by using the approach described in [1].

Starting from the SI&TI CDHMM-based baseline system, a series of supervised batch-mode speaker/task adaptation experiments are conducted for each speaker and each TD vocabulary by using the standard MLLR adaptation algorithm [8] and different sets of adaptation data. In all the adaptation experiments, a common regression tree with 32 leaf nodes derived from the SI&TI CDHMMs is used, and a specific regression class is al-

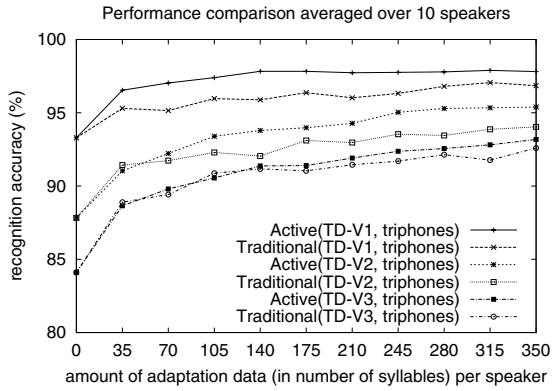

Figure 2: Performance (word accuracy in %) comparison averaged over 10 speakers with respect to the amount of adaptation data per speaker (in number of syllables): active *vs.* traditional approaches.

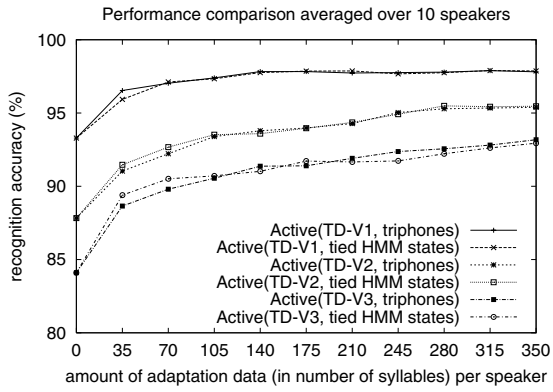

Figure 3: Performance (word accuracy in %) comparison averaged over 10 speakers with respect to the amount of adaptation data per speaker (in number of syllables): the effect of coverage units (triphones *vs.* tied HMM states) on active approach.

ways used for the silence model only. Block diagonal transformation matrices are used and the threshold for the *occupation counts* [8, 11] is set to be 400.

Fig. 2 shows the performance comparison between the proposed active approach and the traditional approach when both use triphones as *coverage units*. It is observed that the proposed active approach requires much less adaptation data to achieve the same performance of the traditional approach. For example, in *TD-V1* task, the proposed active approach achieves a word recognition accuracy of 97.03% with 70 syllables of adaptation data while the traditional approach requires 315 syllables of adaptation data in order to achieve a word recognition accuracy of 97.06%. Understandably, the performance difference between two approaches becomes smaller when most vocabulary words are confusable, e.g., in *TD-V3* task. Fig. 3 shows the effect of using triphones or tied HMM states as *coverage units* on the proposed active approach. The results are mixed, therefore no definite conclusion can be drawn at this stage.

4. Conclusion and Discussions

In this paper, a new active approach has been proposed to improve the efficiency for speaker and task adaptation. It is safe to conclude that the above approach has passed the proof-of-concept test based on the above experimental results and has a good potential. Ongoing and future works include 1) improving the way of identifying the subsets of most confusable words that may cause recognition errors without running recognition experiments; 2) studying more efficient procedures for generation of adaptation script; 3) developing new adaptation algorithms that can make better use of the relevant information identified from the vocabulary confusability analysis; 4) investigating the effectiveness of incremental active adaptation approaches by accounting for both confusable words and frequently used query words; and 5) exploring the possibility of combining the concept of *active learning* with that of *reinforcement learning* to develop a more efficient and user-friendly technique for speaker and task adaptation. We will report those results elsewhere when they become available.

5. References

[1] X. Cui and A. Alwan, "Efficient adaptation text design based on the Kullback-Leibler measure," *Proc. ICASSP*, 2002, pp.I-613-616.

[2] D. Hakkani-Tur, G. Riccardi, and G. Tur, "An active approach to spoken language processing," *ACM Trans. on Speech and Language Processing*, Vol. 3, No. 3, pp.1-31, 2006.

[3] Q. Huo and W. Li, "A DTW-based dissimilarity measure for left-to-right hidden Markov models and its application to word confusability analysis," *Proc. Interspeech - ICSLP*, 2006, pp.2338-2341.

[4] T. M. Kamm and G. G. L. Meyer, "Robustness aspects of active learning for acoustic modeling," *Proc. ICSLP*, 2004, pp.1095-1098.

[5] H. Kawai, S. Yamamoto, N. Higuchi, and T. Shimizu, "A design method of speech corpus for text-to-speech synthesis taking account of prosody," *Proc. ICSLP*, 2000, pp.420-425.

[6] H.-K. Kuo and V. Goel, "Active learning with minimum expected error for spoken language understanding," *Proc. Interspeech - Eurospeech*, 2005, pp.437-440.

[7] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proceedings of the IEEE*, Vol. 88, pp.1241-1269, 2000.

[8] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, Vol. 9, pp.171-185, 1995.

[9] J.-L. Shen, H.-M. Wang, R.-Y. Lyu, and L.-S. Lee, "Automatic selection of phonetically distributed sentence sets for speaker adaptation with application to large vocabulary Mandarin speech recognition," *Computer Speech and Language*, Vol. 13, pp.79-97, 1999.

[10] J. P. H. van Santen and A. L. Buchsbaum, "Methods for optimal text selection," *Proc. Eurospeech*, 1997, pp.553-556.

[11] S. J. Young, *et al.*, *The HTK Book* (revised for HTK Version 3.2), 2002.