Speech Recognition Adaptation Techniques for Computer Assisted Language Learning (Korean)

John Morgan
United States Military Academy
Department of Foreign Languages
745 Brewerton Road
West Point, NY 10996
Email: john.morgan@usma.edu
Phone: 845-938-2966
Fax: 845-938-3585

October 24, 2003

ABSTRACT

This paper describes a Model Merging adaptation method for automatic speech recognition (ASR) for computer assisted language learning (CALL) of Korean. The goal of model merging is to improve the accuracy of an ASR system by incorporating information about the native language of the speaker. In this case Hidden Markov Model (HMM) phone sets are trained for English and Korean, then mixtures from the English system are merged into the Korean system to make a new Korean system that is tolerant of Anglophone learners of Korean. The paper also describes a data-driven method for automatically mapping phones between the two languages.

INTRODUCTION

Speech recognition systems for most applications are typically trained on dozens of hours of speech from native speakers. These kinds of systems are less effective when used by non-native speakers for foreign language learning applications.

Recognizers based on statistical models derived from native speech are designed to tolerate different speech patterns produced by different speakers of a language. However, they are not tolerant enough to accommodate both the speech patterns of native speakers and those of beginning learners of that language.

The acoustic models in the speech recognition systems used here are multiple mixture Gaussian HMMs. An HMM is a finite state machine whose states emit symbols according to a probability distribution. The model merging method combines Gaussian probability distributions from the acoustic models of the two languages to form a new acoustic model for the target language. The model merging method can be applied to the phone, state, or Gaussian levels of the HMM. In this study, merging was done at the phone level only.

In [1] good results were obtained from experiments in which the speakers' source language was Japanese and Spanish and the target language was English. Here we focus

on English as the source and Korean as the target language, but otherwise the steps outlined in the above thesis for acoustic model merging were followed.

MATERIALS

HTK [2] was used for data preparation, model training and adaptation, and decoding. Read speech data from the SANTIAGO Korean Corpus [3] was used as training data for the Korean acoustic models. This corpus will be made available from the Linguistic Data Consortium (LDC). Read speech data from the G3 Corpus [4] was used as training data for the English acoustic models. This corpus will also soon be available from the LDC.

A Beowulf class parallel computing cluster consisting of 7 workstations was used for the data preparation, training and decoding. The machines in the cluster have dual Intel Pentium 1200 MHz processors with 500 megabytes of RAM running Redhat Linux 8.1. The machines were networked via a gigabit switching hub.

METHODS

English acoustic models were trained from native American English speakers in the G3 corpus, the results of a speech data collection effort done in conjunction with Army Research Laboratory and sponsored as a community service project by Cadet Company G3 in 2000. Each monophone in this model set was trained up to 30 Gaussian mixtures. Korean acoustic models were trained on the native speaker data from the SANTIAGO corpus, collected in the fall of 2002 at the Korean Military Academy and at Yonsei University in Seoul, Korea. Each monophone in this model set was trained up to 39 Gaussian mixtures.

In order to merge two model sets a map must be made that decides which pairs of models will be merged. The map was derived automatically from the training data. Two recognition passes were run on the data: one in forced alignment mode using the target native Korean models, and another using a phone loop with the source English models. Then a confusion matrix was produced comparing the results of these two recognition passes. The (i, j) entry in the confusion matrix gives the number of times the phone [pi] was recognized as phone [pj]. The sum of all the entries in one row is the total number of times the phone corresponding to that row was recognized. By dividing the (i, j) entry by the row i^{th} sum the relative number of times phone [pi] was recognized as [pj] is obtained. By performing this calculation for each entry, a relative confusion matrix was produced. All phones that had nonzero entries in a given row of the relative confusion matrix were merged, and their Gaussian weights were multiplied by the relative number of times they were confused.

Each Gaussian component in each state of an HMM has an associated weight, and the sum over all the components in one state of these weights must add to 1. Since the weights from the English phone set add up to 1 and the weights from the Korean phone set also add up to 1, the total sum of the weights after merging was 2. Therefore, before merging components the weights were each multiplied by 0.5. Notice that we arbitrarily chose to assign equal weight to the source and target phone sets. For example, we could have assigned a weight 0.9 to the source and 0.1 to the target. Except for this

modification of the weights, the merging procedure consisted of copying Gaussian components from the two model sets into one set.

The full model sets were not used to do the merging, because otherwise the resulting models would have too many parameters. Hence, the model sets from earlier steps in the mixture incrementing process were used. Specifically, 2 mixtures from the English model set and 7 mixtures from the Korean model set were used for merging.

REFERENCES:

[1] Witt, S.M., "Use of Speech Recognition in Computer-assisted Language Learning", Newnham College, University of Cambridge, 1999.

[2] Young, S. J. and Odell, J. and Ollason, D. and Woodland, P. C., *The HTK Book*. Entropic Cambridge Research Laboratory, 1996. http://htk.eng.cam.ac.uk

[3] SANTIAGO Korean Corpus, COL Stephen A. LaRocca, 2002.

[4] The G3 Corpus. COL Stephen A. LaRocca, Sherri M. Bellinger and John J. Morgan, 2000.