Making a Speech Recognizer Tolerate Non-native Speech through Gaussian Mixture Merging

John J. Morgan
Center for Technology Enhanced
Language Learning
United States Military Academy
West Point, NY 10996
John.morgan@usma.edu

Abstract

Practicing the spoken language is important to language learners in critical parts of the United States military. Automatic speech recognition (ASR) is a technology that promises to provide self-paced practice opportunities to language learners. We propose to improve ASR in computer assisted language learning (CALL) applications by modeling the speech behavior of language learners. ASR systems trained on native data perform poorly when used to recognize beginning language learners. The Model Merging adaptation method via a confusion matrix map makes our Arabic speech recognizers more tolerant of Anglophone students.

Hidden Markov Model (HMM) phone sets are trained for English and Arabic, and then English phones are merged into the Arabic phones to make a new Arabic system. A data-driven procedure is presented for automatically mapping phones between two HMM sets.

Accuracy improvements were observed when model merging was combined with other adaptation techniques. The positive results indicate that the speech patterns of non-native speakers are carried over to the new system by the mapping of phones and their weighting.

1. Introduction

Foreign language learning, especially mastery of its spoken component, has become important recently to the United States Army because of commitments in countries all over the world. These commitments increasingly involve teams that work closely with local inhabitants and speaking the language is often essential to accomplish their missions. Automatic speech recognition (ASR) technology provides army linguists with ways of practicing their speaking skills independently and at their own pace. ASR in language learning, however, poses some special problems for software developers.

1.1 The Problem

A speech recognizer has two main goals in CALL applications: recognition and pronunciation evaluation. Recognition is simply rendering the learners intended utterance in textual format; pronunciation evaluation, on the other hand, strives to assign a score to the utterance by evaluating it against native standards. However, these two goals are not both achievable using a single ASR system. One method of dealing with this dilemma is to run two passes with two different recognizers, one in recognition mode and another in evaluation mode. The recognizer used for the recognition mode is optimized to tolerate non-native speech behaviors while the recognizer in the second pass is trained strictly on native speech. We will only discuss the recognition mode of this problem, see for example Neumeyer et al (1998), for work on evaluation techniques.

ASR systems trained on native speakers do not perform well when recognizing beginning language learners. Chase (1997) lists some sources of errors in ASR systems, including the following:

1. out-of-vocabulary (OOV) word spoken,
2. search error,
3. homophone substitution,
4. language model overwhelming correct acoustics,
5. transcript/pronunciation problems, or
6. confused acoustic models

Since CALL applications use small vocabularies and small language models, we blame recognition errors on confused acoustic models.

1.2 Adaptation Techniques

To improve recognition in CALL applications, acoustic models of phones are modified to make them tolerant of the typical pronunciations of a language learner. Two well-known ways of modifying the acoustic models are re-training and adaptation. Both techniques use speech from language learners. Re-training the models entails boot-strapping from the native-trained models and continuing the same training process on the non-native speech.

Adaptation techniques can also use well-trained native acoustic models, but they use the learner speech in different ways. We will work with adaptation techniques that modify the parameters of acoustic models, including Constrained Maximum Likelihood Linear Regression (CMLLR), and Maximum A Posteriori (MAP). In addition to these methods we will employ a technique called Model Merging (MM) that inserts model parameters from the source language into the acoustic models of the target language.

1.3 The Hypothesis

We claim that model merging via a confusion matrix map incorporates information about the typical pronunciation mistakes made by learners into the acoustic models and thereby makes the resulting ASR system more tolerant of student speech.

1.4 Some Background

We assume that we have well-trained acoustic models for both the learner's source language and the target language. We also assume a small corpus of student speech is available. The central idea of model merging is that in order to account for the speech behavior of a typical student we must merge the acoustic models of the source language into the target language models. We perform this merging in such a way that the resulting models are more tolerant of the mistakes learners are likely to make when attempting to speak the target language.

Recall that our goal in using these models at this point is not to evaluate the student's pronunciation, but instead to recover the words they intended to utter. The evaluation of their speech is left to another step using acoustic models trained on native speech.

We also assume that each acoustic model in each language is a Hidden Markov Model (HMM) that describes a phonetic segment, or phone, and that the inventory of phones in each of the two languages is different. The model merging problem is then to decide which phones should be merged. Witt (1999) obtained good results for experiments in which the speakers' source language was Japanese or Spanish and the target language was English. Here we focus on English as the source and Arabic as the target language.

1.5 The Work

HMM phone sets are trained for English and Arabic. Using a hand-made map, the non-native training data is re-labeled with English phones. Using these re-labeled transcriptions, the English models are adapted to the non-native training data. Next, a mapping between the two phone sets is derived using a method that was first suggested to the authors by J. M. Huerta (see Beyerlein et al, 1999). The English phones are inserted into the Arabic phones according to this mapping. Finally, CMLLR, MAP adaptation and Baum Welch (BW) embedded parameter reestimation training are applied using the same student speech.

2. Methods and Materials

2.1 Speech Corpora

Five corpora were used to train the acoustic models in this study. Two corpora, the Santiago and the Tunisian Cadet corpora of read Modern Standard Arabic (MSA) speech (United States Military Academy, 2002, 2003), were used as training data for the native Arabic acoustic models. The Santiago corpus is available from the Linguistic Data Consortium (LDC) and the Tunisian Cadet corpus will be published at a future date.

The non-native Arabic speech used in the adaptation process came from two corpora. Cadets at the United States Military Academy (USMA) at West Point read words and short phrases extracted from the Tactical Language Training System (TLTS), a prototype language learning system for Arabic funded by the US Government. Speech data from military linguists is also included in the Santiago corpus and was used in the adaptation process.

Two corpora of read English speech were used in training the native English acoustic models. The G3 Gophers Corpus of American English (United States Military Academy, 2002) was collected from volunteers at USMA. This corpus is as yet unpublished. The TIMIT corpus (Linguistic Data Consortium, 1993) was also used.

2.2 Tools and Hardware

The Hidden Markov Model toolkit (HTK) was used for data preparation, model training and adaptation, and decoding. For the adaptation portion we used alpha release version 3.3. A Beowulf class supercomputer consisting of a cluster of 7 workstations was used for the data preparation, training and decoding. The machines are Intel Pentium 1200 MHz dual processors with 500 MB of RAM running Redhat Linux 8.1. The machines are networked via a gigabyte hub.

2.3 Training the Models

English acoustic models were trained from American English speakers in the G3 and TIMIT corpora by following the incremental steps given in chapter 3 of Young et al (1996).

In order to model more closely the speech behavior of Anglophones speaking Arabic, the English model set was adapted to the student training data. To accomplish this, the phone level transcriptions of the training data had to be relabeled with English phones. The simple map, shown in Table 1 below, was used to re-transcribe the non-native training data in terms of phones from the English model set.[1] Consequently, MLLR, MAP, and BW training were applied to the English models using the new transcriptions.

AR	EN	AR	EN	AR	EN
C	-	D	d	G	r
H	hh	Q	-	S	s
T	t	TH	th	Z	th
ae	ae	ah	ah	aw	aw
ay	ay	b	b	d	d
ey	ey	f	f	g	g
h	hh	ih	ih	iy	iy
j	zh	k	k	l	l
m	m	n	n	q	k
r	r	s	s	sh	sh
t	t	th	th	uw	uw
w	w	x	hh	y	y
z	z				

Table 1: Map from Arabic to English phones

Two sets of native Arabic acoustic models, one monophone and one triphone, were trained on the two native speaker corpora. Although both corpora consist of read MSA speech, the speakers in the Santiago corpus were mostly of Levantine origin with some speakers coming from the Gulf region. The Tunisian Cadet corpus consists entirely of speakers from the Tunisian Military Academy.

The monophone system was trained in exactly the same way as the English system described above. The second native Arabic system consists of cross-word decision tree clustered triphones. Young et al (1996) describes the development of a word internal triphone system, and the steps we followed for a cross-word system were similar. The questions required for the tree-based clustering were derived from the feature geometry described in chapter 9 of Kenstowicz (1994).

2.4 Making the Map

Since the phonetic inventories of Arabic and English are different and it is not obvious what source language phones a student will substitute for when attempting to pronounce an Arabic word, an algorithm must be designed that when given a phone in Arabic decides which English phones should be merged into it. The goal here is to produce an HMM that models student disfluencies.

We derived a phone map automatically from the non-native training data. Two recognition passes were run on the data. The first pass was run in forced alignment mode to obtain phone time boundaries using the native Arabic triphones. The triphones were later stripped down to monophones. These time-aligned phone labels served as our reference transcription. The adapted English system was run on the same non-native data under a phone loop, where the recognizer chooses from a list of the English phones. 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 p_i was recognized as phone p_j. 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 p_i was recognized as p_j was obtained. By performing this calculation for each entry, a relative confusion matrix was produced. All phones that had non-zero 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.

[1] This map was designed by one of the authors, who is not a native Arabic speaker, therefore we do not claim that this map is the ideal model for typical learner substitutions.

The gaussian mixtures in each state of an HMM form a convex combination, that is, each mixture has an associated weight, and the sum over all the mixtures in one state of these weights must add to 1. Since the weights of both the English and the Arabic phone sets add up to 1, the total sum of the weights after merging was 2. Therefore, the weights had to be normalized prior to merging. Prior to merging mixtures the weights in the source phone set were each multiplied by an extra factor *u*, and the weights in the target phone set were multiplied by 1-*u*. By varying the value of *u* the weight of the native English models were adjusted to the desired amount of "non-nativeness".

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

After the model merging process, three types of adaptation were applied to the acoustic models. First we ran global MLLR, followed CMLLR with 32 regression classes, and finally MAP. After adaptation we followed up with four passes of BW training. In this last step we updated different combinations of the four parameters: means, variances, transitions, and mixture weights.

2.5 Testing

We tested the models on a separate set of non-native speech data and also separated a small development data set to adjust the pruning threshold for the language model. We used a word loop language model incorporating a list of the 2800 vocabulary words spoken by the informants.

3. Results

In Tables 2 through 5 below, mvtw reest stands for means, variances, transitions, and weights Baum Welch reestimation, and, as previously defined, CMLLR for constrained maximum likelihood linear regression, and MAP for maximum a posteriori. As shown in Table 3, a 93.64 percent improvement was observed when CMLLR, MAP and Baum Welch reestimation of all parameters were applied to the merged models. Table 2 indicates that when the same adaptation strategy was applied to the native models, an improvement of 79.27 percent was observed. Hence, the model merging strategy outperforms the baseline adaptation strategy by more than 14 percent.

As can be seen in Table 4, the use of a hand-made phone map to adapt the English models before producing the confusion matrix led to an improvement of 97.64 percent over the native trained models. Thus the adaptation of the English models provided an extra 4 percent improvement. The data presented in Table 5 indicates that 86.04 percent of the total improvement from BW training came from updating the mean vectors, and in Table 6 we see that recognition accuracy is sensitive to the weighting applied to the English acoustic models.

Adaptation	Accuracy
none	26.73
mvtw reest	47.28
CMLLR MAP mvtw reest	47.92

Table 2: Accuracy scores for native system with different adaptation strategies

Adaptation	Accuracy
merged models	18.00
merged mvtw reest	50.37
merged CMLLR MAP mvtw reest	51.76
merged global mvtw reest	47.18

Table 3: Accuracy scores for merged model sets with different adaptation strategies

Adaptation	Accuracy
merged (hand-made map)	19.28
merged (hand-made map) mvtw reest	49.95
merged (knowledge map) CMLLR MAP mvtw reest	51.76

Table 4: Accuracy scores for merged model sets with hand-made mapping applied to English models

Adaptation	Accuracy
merged t-reest	19.38
merged v-reest	19.38
merged w-reest	30.67
merged m-reest	47.07
merged mv-reest	47.92
merged mw-reest	49.09
merged mvtw-reest	50.37

Table 5: Accuracy scores for merged model sets with different parameters

English weight	Accuracy
0.2	19.60
0.4	18.00
0.5	18.32
0.7	18.10
0.9	17.25

Table 6: Accuracy scores for merged model sets with different weights

4. Discussion

We made the assumption that HMM sets trained on native speech data incorporate speech patterns specific to native speakers. We also assumed that English-specific speech behavior influences the way Anglophone learners produce Arabic sounds. We intended to model the speech behavior of Anglophone students of Arabic, with the goal of producing a learner-tolerant ASR system. To this end we obtained a sample of phone substitutions from an actual corpus of speech data. We asserted that a confusion matrix made from these substitution counts models the speech behaviors produced by students speaking Arabic in a many-to-one map. Since we see improvements over traditional adaptation techniques we believe that model merging via a confusion matrix map is indeed incorporating English speech patterns into the Arabic HMMs.

The map we produced from the confusion matrix in our experiments only considered phone substitutions. Perhaps more precise models of phonological differences can be made by also using data collected on phone deletions and insertions.

We believe that a lot more work needs to be done in the steps taken to adapt the English models to Anglophone Arabic students' speech. In this project we used a hand-made map to re-label the transcriptions of the student data. Further work should focus on a narrower labeling of the student speech with the English phones. Finally, future work should also explore varying the weights between the English and Arabic model sets to make an ASR system sensitive to different levels of spoken language proficiency.

5. Acknowledgements

All of the speech corpora used in this project (with the exception of the TIMIT Corpus) were collected by staff and faculty of the Center for Technology Enhanced Language Learning, the Department of Foreign Languages and members of the West Point community during 1997 – 2004.

6. References

Beyerlein, P., Byrne, W., Huerta, J.M., Khudanpur, S., Marthi, B., Morgan, J., Peterek, N., Picone, J. & Wang, W. (1999). "Towards language independent acoustic modeling". Presented at the IEEE Workshop on Automatic Speech Recognition and Understanding", Boulder, CO, December 1999.

Chase, L. (1997). "Blames assignment for errors made by large vocabulary speech recognizers." *Proc. of Europeech, 1997.*

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., & Zue, V. (1993). "TIMIT Acoustic-Phonetic Continuous Speech Corpus." U. Penn, Linguistic Data Consortium.

Harris, G. & Morgan, J. (2004). "West Point Tactical Language Corpus of Arabic Student Speech. Unpublished speech corpus.

Huang, C., Chen, T., & Chang, E. (2004) Accent Issues in Large Vocabulary Continuous Speech Recognition. *International Journal of Speech Technology*, 7, 141-153.

Kenstowicz, M. (1994). *Phonology in Generative Grammar.* Cambridge, MA: Blackwell.

LaRocca, S., Morgan, J. & Bellinger, S. (2002). "The G3 Gopher Corpus of Spoken American English". Unpublished speech corpus.

LaRocca, S. & Chouairi, R. (2002). "West Point Arabic Speech Corpus." U. Penn: Linguistic Data Consortium.

LaRocca, S., & Morgan, J.(2003). "The Tunisian Military Academy Arabic Speech Corpus". Unpublished speech corpus.

Neumeyer, L., Franco, H., Digalakis, V., & Weintraub, M. (1998). "Automatic scoring of pronunciation quality". *Speech Communication,* 30 (2000), 83-93.

Witt, S. (1999). "Use of speech recognition in computer-assisted language learning", Ph.D. Thesis, Newnham College, University of Cambridge

Young, S., Odell, J., Ollason, D., & Woodland, P.C. (1996). *The HTK Book.* Univerisity of Cambridge: Entropic.