Making A Speech recognizer tolerate Nonnative Speech Through GAUSSEAN MIXTURE MERGING

John Morgan COL. Stephen A. LaRocca
Sherri Bellinger
Center For Technology Enhanced Language Learning
United States Military Academy
Email: john.morgan@usma.edu

June 15, 2004

The Problem

- Importance of speaking practice for language learners in US Army.
- ASR provides self-paced practice to lerners.
- Native trained ASR performs poorly on beginning students.
- ASR has 2 goals in CALL : recognition and evaluation.
- Both goals are not achievable using a single system.
- 2-pass solution.
- Recognition pass uses nonnative tolerant acoustic models.
- Evaluation pass uses native trained models.
- Improve ASR in CALL applications by modeling student phonology .
- What causes errors?
- Since CALL applications use small vocabularies and small language models, we blame recognition errors on confused acoustic models.
- Beginning students use phones from their native language .

Adaptation Techniques

- Modify acoustic models to make them tolerant of typical student pronunciation.
- Use MLLR, MAP, and Model Merging (MM).
- MM inserts model parameters from source language into target language.

ASSERTION 1. *MM via a confusion matrix map makes acoustic models tolerant of student speech .*

ASSERTION 2. *this tolerance can be made specific to different levels of proficiency, that is, we can tailor models to be tolerant of speech from beginner, intermediate, or advanced learners.*

Some Background

- Well-trained acoustic models for English and Arabic.
- Merge English acoustic models into Arabic models.
- perform merging so that resulting models are tolerant of likely learner mistakes.
- Acoustic models are HMMs describing phonetic segments.
- MM problem: decide which phones to merge and in what proportion.
- Witt and Young 1999 obtained good results for Japanese and Spanish source and English target.

The Work

- Train HMM phone sets for English and Arabic.
- Adapt English models to student data.
- Derive mapping between phone sets from student speech .
- Merge English phones into Arabic phones according to mapping.
- Adapt and reestimate parameters.

Speech Corpora

- Use 5 corpora to train acoustic models .
- Train native Arabic acoustic models with Santiago and Tunisian Cadet corpora of read MSA.
- Santiago speakers are mostly Levantine with some from the Gulf region.
- Tunisian Cadet corpus consists of speakers from the Tunisian Military Academy.
- 2 corpora of non-native Arabic for adaptation:
- USMA cadets read short phrases from the TLTS.
- Mmilitary linguists are part of Santiago .
- Train native English monophones on G3 and TIMIT.

Training

- Use HTK version 3.3α for data preparation, training, adaptation, and decoding.
- Follow incremental steps given in chapter 3 of HTK Book .
- Train native Arabic 2-mixture monophones and 10-mixture cross word decision tree clustered triphones.

Making The Map

- Decide which English phones should be merged into an Arabic phone.
- GOAL: Produce an hmm that models student behavior.
- Derive phone map automatically from non-native training data.
- Run Two recognition passes .
- Forced alignment to obtain phone time boundaries with native Arabic triphones.
- strip triphone labels down to monophones.
- Time aligned phone labels serve as reference transcription.
- Run phonetic recognition with native English monophones on same non-native data.
- Produce confusion matrix comparing results of two recognition passes.
- Entry (i, j) is number of times phone p_i was substituted for phone p_j.
- Sum all entries in one row to get total .
- Divide entry (i, j) by row i^th sum to get relative matrix.

Confusion Matrix

	aa_e	ae_e	ah_e	ao_e	aw_e	ax_e	axr_e	ay_e	b_e	ch_e	d_e	dh_e	dx_e	eh_e	er_e	ey_e	
C	31	19	14	1	8	20	3	22	12	0	5	17	8	5	0	5	
D	8	3	10	0	2	9	0	1	19	0	18	11	5	1	0	0	
G	3	4	8	2	10	2	3	2	3	1	0	5	7	3	0	1	
H	22	7	26	1	2	13	6	11	6	0	3	14	3	3	1	0	1
Q	70	113	94	3	44	85	18	33	73	3	42	93	64	62	3	10	3
S	7	2	22	1	0	32	5	3	6	1	6	1	5	4	0	0	1
T	9	5	13	0	4	14	1	4	4	2	4	12	4	4	0	0	
TH	1	5	4	0	1	2	2	1	7	0	3	8	19	3	0	1	
Z	2	0	4	0	1	1	0	0	0	0	2	2	0	1	1	0	
ae	64	75	36	8	32	32	8	27	30	2	18	30	44	16	1	5	2
ah	211	241	359	19	129	275	103	98	151	5	101	158	127	135	6	24	9
aw	10	1	6	3	6	4	0	2	4	0	3	2	3	2	1	1	
ay	3	6	1	0	1	1	0	19	3	0	3	0	0	3	0	12	
b	27	39	40	2	12	29	10	15	101	0	34	21	32	13	0	0	
d	5	11	12	0	2	13	1	8	15	0	31	11	12	10	0	0	
ey	0	1	1	0	0	2	1	4	2	0	2	3	6	1	0	19	
f	18	17	22	2	9	23	5	13	12	0	12	12	4	4	0	0	7
g	0	0	0	0	0	0	0	1	0	0	3	2	0	0	0	0	
h	5	7	11	2	4	3	3	1	10	0	3	19	6	3	0	1	
ih	15	30	42	4	14	67	6	10	27	0	26	41	30	25	0	13	3
iy	13	43	26	4	17	31	8	8	33	0	46	38	41	17	1	13	2
j	1	7	10	1	4	9	4	5	3	1	2	3	1	10	0	7	
k	3	23	11	0	12	11	4	3	4	2	5	11	2	22	0	2	
l	65	76	96	9	35	65	19	34	40	0	26	62	88	46	5	13	3
m	35	24	60	3	28	50	12	14	14	0	12	47	17	15	2	1	
n	30	63	67	2	17	83	17	20	27	1	15	48	27	33	7	2	1
q	13	14	13	2	2	13	5	4	8	1	9	14	9	13	0	1	
r	32	22	38	3	16	30	34	18	12	5	14	23	78	13	4	10	1
s	5	6	13	0	3	16	1	4	3	0	2	14	4	7	1	1	

sh	6	3	5	0	2	16	2	0	2	3	1	6	4	4	0	4	
t	14	12	40	1	4	30	6	1	17	4	14	26	11	12	0	4	
th	1	13	14	0	2	14	1	1	3	1	2	7	1	9	1	0	1
uw	39	35	59	13	14	54	10	16	45	0	28	30	28	17	1	2	1
w	21	13	21	3	13	14	5	17	14	0	3	16	6	4	0	1	1
x	6	4	9	0	3	9	6	4	3	0	2	8	1	3	1	1	1
y	2	12	5	1	6	11	5	1	1	0	1	7	0	4	0	5	
z	6	6	10	0	6	7	3	4	1	0	0	4	7	2	1	0	

Non-nativeness

- Gaussian mixtures form a convex combination.
- Normalize weights prior to merging.

$$\text{MERGED} = (u)\text{Arabic} + (1 - u)\text{English}$$

- Vary u to adjust non-nativeness.

Annealing

- Use 2 mixtures from each model set for merging.
- Run CMLLR with 32 regression classes and MAP.
- Follow up with four passes of BW training.
- Update conbinations of means, variances, transitions, and mixture weights.

Testing

- Separate test set of non-native speech.
- Separate small development data set to adjust language model pruning threshold.
- Use language model containing list of 2800 wordss spoken by informants.

results

- table 2shows 93.64% improvement when CMLLR, MAP and Baum Welch reestimation of all parameters was applied to merged models.
- Table 1shows 79.27% improvement for native models.
- MM outperforms baseline by more than 14%.

Adaptation	Accuracy
none	26.73
mvtw reest	47.28
cmllr map mvtw-reest	47.92

Table 1: Accuracy scores for native system with different adaptation strategies.

Adaptation	Accuracy
merged	18.00
merged mvtw-reest	50.37
merged cmllr map mvtw-reest	51.76
merged global mvtw reest	47.18

Table 2: Accuracy scores for merged model sets with different adaptation strategies.

More Results

- Table 3 shows that knowledge based phone map to adapt English models led to improvement of 97.64 percent over native trained models.
- An extra 4% improvement.

Adaptation	Accuracy
merged knowledge	19.28
merged knowledge mvtw-reest	49.95
merged knowledge cmllr map mvtw-reest	52.83

Table 3: Accuracy scores for merged model sets with knowledge based mapping applied to English models.

More Tables

- Table 4 shows that 86.04% of total improvement from BW training came from updating mean vectors.

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 4: Accuracy scores for merged model sets with different parameter updating.

More Tables

- Table 5 shows that recognition accuracy is sensitive to the weight of the english models.

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

Table 5: Accuracy scores for merged model sets with different weights.

Conclusions

- Fundamental Assumption: HMMs encode speech patterns.
- Insert English patterns into Arabic hmms.
- Estimate phone substitution patterns from corpus of student speech.
- Assertion: confusion matrix captures patterns in many-to-one map.
- Encouraging results for MM.

Future Work

- Confusion matrix only considered phone substitutions.
- Consider deletions and insertions in future work.
- Explore varying weight to make ASR sensitive to levels of proficiency.