SRI International

Improving Language Recognition with Multilingual Phone Recognition and Speaker Adaptation Transforms

A. Stolcke, M. Akbacak, L. Ferrer, S. Kajarekar
C. Richey, N. Scheffer, E. Shriberg
SRI International, Menlo Park, CA, U.S.A.

Odyssey Workshop, Brno, July 1, 2010

Overview

- Language recognition task
- Standard approaches
- Method, data, baseline
- Phonotactic LM
  - Multilanguage phone recognition
  - MLP features
- MLLR modeling
- Phonotactic SVM modeling
- Future work
- Conclusions

Language Recognition Task

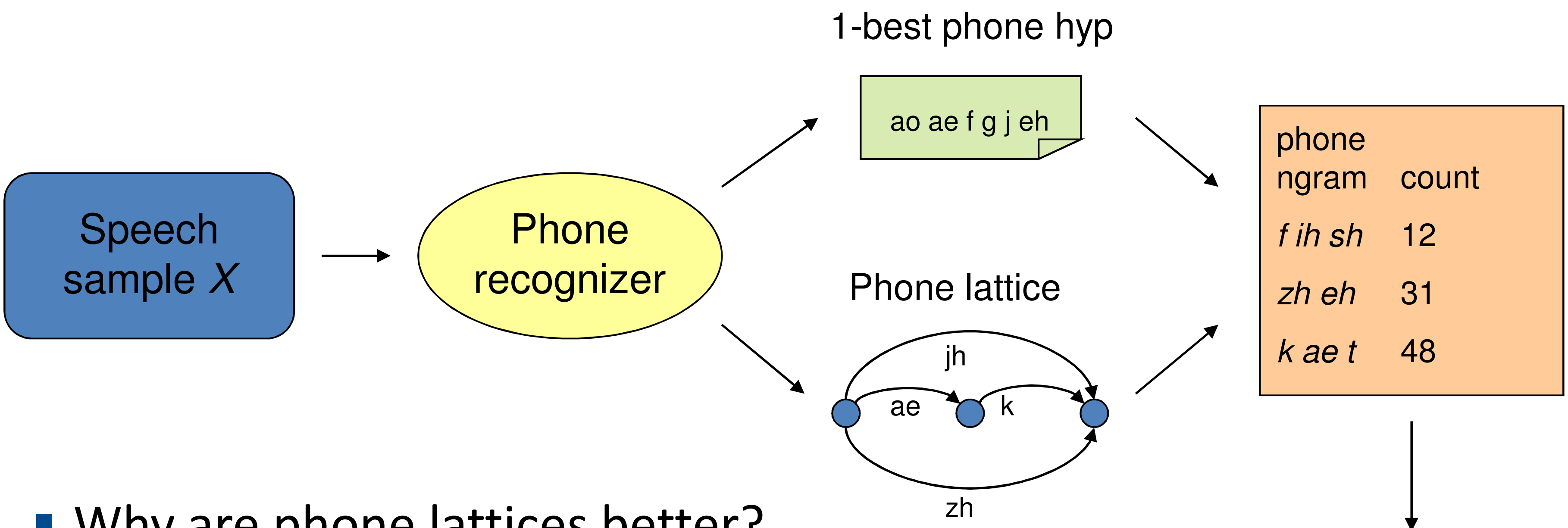
- NIST LRE'05 task
  - Most recent eval set released by LDC at the time of this work
- 7 target languages
- Conversational telephone speech
- Test data: 3662 test segments of ~ 30 seconds each
- Training data (duration after auto-segmentation):
  - English, Mandarin, Spanish: ~ 56 hours each
  - Hindi, Japanese, Korean, Tamil: ~26 hours each
- Metric here: EER averaged over languages (not trials)
- Computed by two-fold cross-validation
  - We didn't have an independent tuning set (see above)

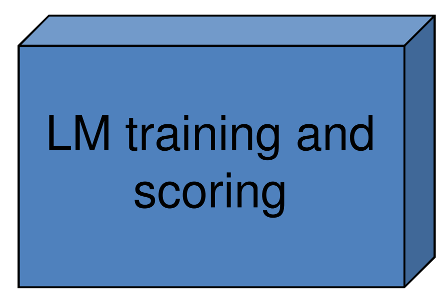
Popular Standard LID Techniques

- Cepstral GMM (similar to SID)
  - Training universal (all languages) background model
  - MAP-adapt to target language training data
  - Form likelihood ratio between target and background models on test data
  - Lately: with JFA for within-language variability compensation
  - AvgEER = 2.87% as implemented by us
- Phone recognition language models (PRLM)
  - Run unconstrained phone recognizer, collect 1-best phone hypotheses
  - Train target and background N-gram LMs; form likelihood ratio
  - Combine two or more language-specific recognizer (***P***PRLM)
- Calibration
  - Map raw model scores to calibrated log likelihood ratios
  - Trained to minimize error metric
  - Here: FoCal multi-class toolkit, based on log linear regression
  - No Gaussian backend used

Phonotactic Language Modeling

Phone N-gram Language Modeling (PRLM)





- Why are phone lattices better?
  - More robust counts
  - Finer granularity in features
- Independently proposed by LIMSI (for LID) and SRI/ICSI (for SID)
- Implemented in SRILM lattice-tool

Parallel Language-specific Phone Recognizers (PPRLM)

- Use standard ASR conversational telephone speech models trained with PLP, VTLN, HLDA, cross-word triphones, MPE (available from other work)

Language	Phoneset	Training data	Gender dependent?
English	47	1400h	yes
Spanish	33	18h	no
Levantine	39	61h	no

- Decoding with "open loop": no phonotactic constraints, all phones equally likely, but using context-dependent triphones
- Phone-loop based CMLLR adaptation (following LIMSI)
  - Note: CMLLR is better than MLLR with 1-best phone hyps

Multilingual Phone Recognizer

- Define a "universal" phone set covering several languages (52 phones)
- Map native word pronunciations to universal phone set
- Train acoustic and phonotactic models on multi-lingual corpus (below)
- Phone recognition accuracy similar to language-dependent recognizers

Language	Native ?	Sources	Duration (h)	Weight
Am. English	yes	Fisher, Swb, CallHome	123	1x
Am. English	no	Fisher	108	1x
Mandarin	yes	CallHome	103	1x
Spanish	yes	CallHome	19	3x
Egyp. Arabic	yes	CallHome	17	3x

- Note 1: Spanish and Arabic data weighted for better balance
- Note 2: Egyptian Arabic used because of available vowelized transcriptions

Phone N-gram LM Scoring

- Standard scoring: background model is trained on **all** languages

$$\text{score}_i = \frac{\log P(X \mid L_i)}{\log P(X \mid \text{all } L)}$$

- Modified scoring: background models is trained on all languages **except** the target language

$$\text{score}_i = \frac{\log P(X \mid L_i)}{\log P(X \mid \text{not } L_i)}$$

Background model	AvgEER	AvgCdet
All languages	3.07	.039
Non-target languages	3.01	.037

Phone N-gram LM Results

Phone set	AvgEER
American English	4.17
Levantine Arabic	4.91
Spanish	5.49
Am.Eng. + Levant. PPRLM	2.99
Am.Eng + Levant. + Span. PPRLM	**2.76**
Multi-lingual PRLM	3.01
Am.Eng. + Levant. + Span. + ML PPRLM	**2.09**

-34%

-24%

- LM building parameters
  - 3gram (not 4gram) LM, no minimum counts (all trigrams)
  - Add-1 smoothing
- Multilingual PRLM better than any language specific PRLM
  - Comparable to PPRLM

Phone Recognition with MLP Features

- MLP features trained for frame-level phone discrimination
  - Training used English phone set
  - Shown to generalize to ASR in other languages even without retraining
- Improves phone recognition accuracy by 2 to 4% absolute (depending on language)
- PLP+MLP models for multilingual PRLM system
  - Similar to BUT LRE'07

Phone set	Multiling. feature	AvgEER
Multilingual	PLP	3.01
Multilingual	**PLP+MLP**	**2.82**
All –language PPRLM	PLP	2.09
All –language PPRLM	**PLP+MLP**	**1.77**

f

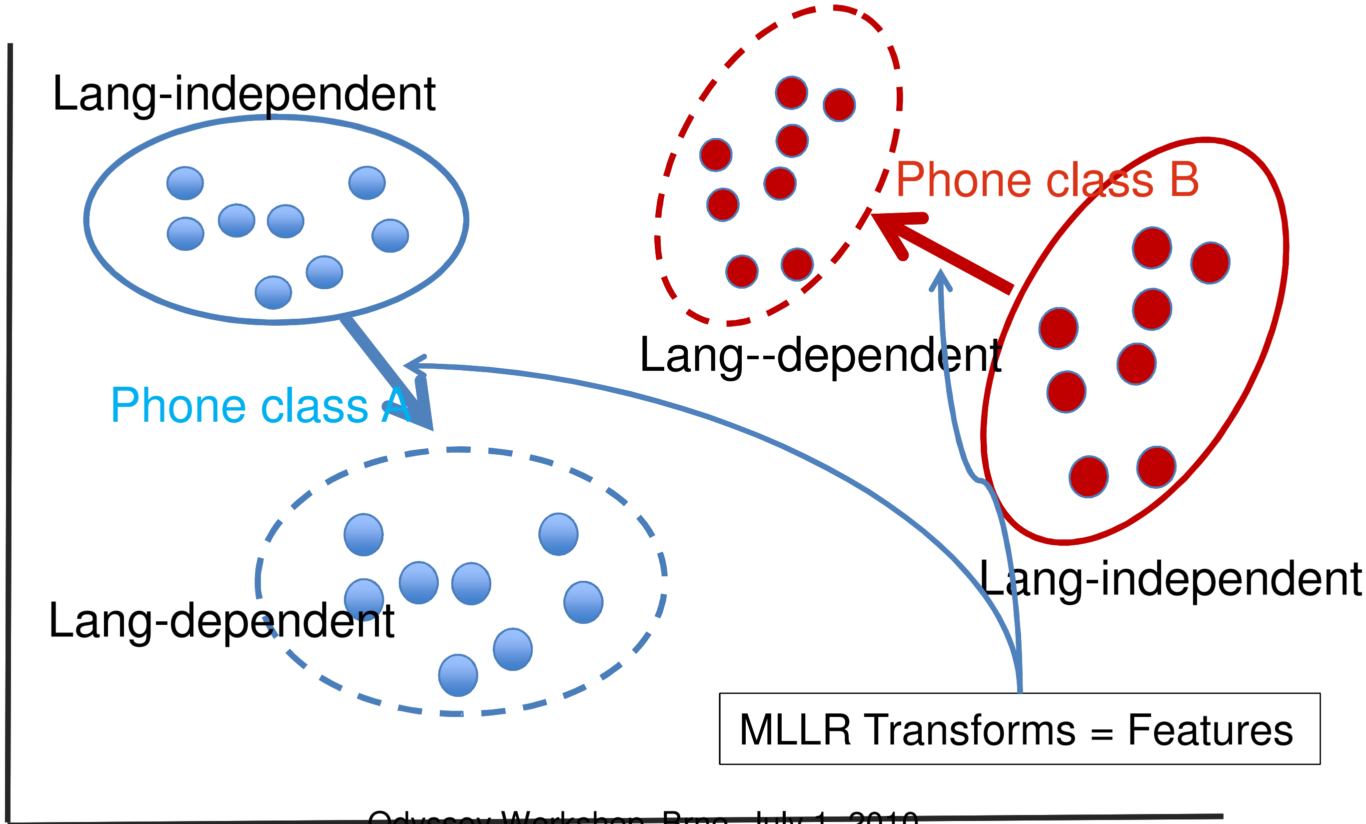
-6%

-15%

MLLR Transform Modeling

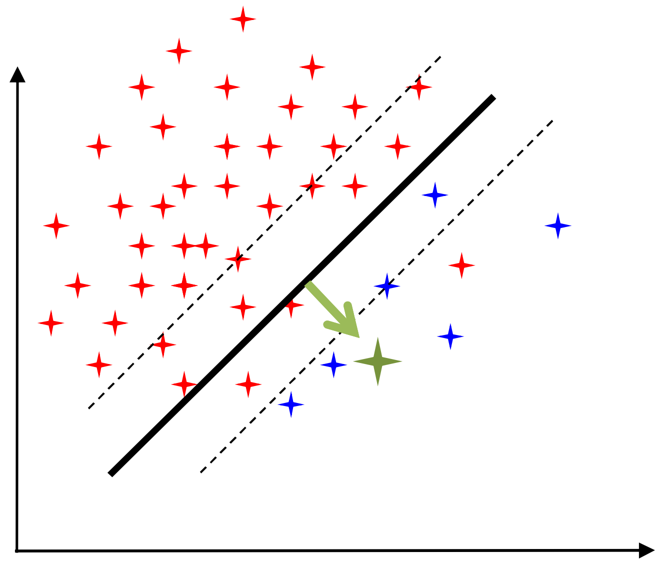
MLLR "Language Adaptation" Transforms

- Estimate transforms mapping language-independent to language-dependent models (using phone-loop MLLR)
- 8 phone classes, 8 transforms
- Transform parameters become language ID features



MLLR Transform Modeling

- 8 x 39 x 40 = 12480 raw feature dimensions
- Rank normalization based on all-language training data
- "Language" models obtained by Support Vector Machine (SVM) training, using linear kernel
- Model = hyperplane separating target from non-target data
- LID score = signed distance from SVM decision boundary



MLLR Transform Results

- Baseline: English gender-dependent acoustic models
- Compare to multi-language, gender-independent models

Acoustic models used	AvgEER
English, female only	12.98
English, male + female	10.25
Multilingual	7.47

- As with phone N-grams, multi-language phone set is much better than language-specific phone set
- Even though there is a nice gain from combining genders
- Try gender-dependent, multilingual phone models?

Improving MLLR Modeling

- In training, split full conversation sides into multiple, 30-second portions (obtain multiple training samples)
- Optimize model size (# of Gaussian): fewer models make for more informative transforms!
- Nuisance attribute projection (NAP): project feature vector to complement of within-speaker and within-language variability space

MLLR system	AvgEER
One transform per train side	7.47
Multiple transforms per train side	5.98
+ Reduce # gaussians (64 → 16)	5.19
+ NAP (12/12480 nuisance dimensions)	4.54
+ MLP features	3.96

MLLR with MLP Features

- Added 25 MLP features to PLP
  - Trained for English phone discrimination
  - Same as for PRLM experiments
- Block-diagonal transform estimation (39x40 + 25x26)
- Feature dimension increased from 12480 to 17680

MLLR feature	AvgEER
PLP	4.54
PLP + MLP	3.96

-13%

Phone N-gram SVMs

Phonotactic SVM Modeling (“PRSVM”)

- For speaker ID, found that SVM models of phone N-grams work better than language models (discriminative training!)
- Try this for language ID, *using multilingual phone recognition*
- TFLLR kernel (Campbell), no ranknorm
- Split training sides into 30sec segments (as for MLLR SVM)

Model	Ngrams	AvgEER
Phone N-gram LM	3g	2.82
Phone N-gram SVM	3g	3.01
Phone N-gram SVM	4g	2.74
Phone N-ngram LM + SVM	3g + 4g	2.42

- Inclusion of 4grams makes SVM better than LM
- LM and SVM modeling somewhat complementary
- Tried NAP, no gain (similar to speaker ID)

Combining Systems

Systems	AvgEER
PRSVM	2.74
Cepstral GMM	2.87
+ Multilingual MLLR-SVM	2.59
+ Multilingual PRLM	1.43
+ Multilingual MLLR-SVM + PRLM	1.19
+ Multilingual MLLR-SVM + PPRLM	1.24
+ Multilingual MLLR-SVM + PRLM + PRSVM	**1.14**

-50%

-20%

- MLLR-SVM gives gains in combination with cepstral system
- PRLM combines better with cepstral systems than PRSVM
- Dual phonotactic modeling (LM+SVM) still gives a small gain
- PPRLM degrades over PRLM in combination
  - Not enough training data for score combination?

Future Work

- Validate experiments on LRE’07 and ‘09 datasets
- Try for dialect ID
  - Hindi vs. Urdu, Indian vs. American English, etc.
- Phone N-gram SVMs for language-specific phone sets
  - Parallel SVM models (“PPRSVM”)
- Retrain MLP features for multilingual phone recognition
- MLP features in language-specific phone recognizers
- Apply detailed linguistic modeling as used in speaker ID:
  - Prosody modeling
  - Constrained cepstral modeling

Conclusions

- Tried various kinds of phone-based systems for LID, inspired by techniques learned in ASR and SID
- First application of MLLR-SVM to language recognition
- Multilingual phone models much better than language-specific models in both PRLM and MLLR systems
- … and still give gains when combined
- Discriminative MLP front end gives gains with both PRLM and MLLR modeling
- MLLR and cepstral GMM combination gives gains
- Phonotactic SVMs allow use of higher-order N-grams than LMs
- Phonotactic LM and SVM over same phone set gives gains