Synergy between Speaker Recognition and Speech Recognition

Andreas Stolcke

Reporting on joint work with:

L. Ferrer, S. Kajarekar, E. Shriberg, K. Sonmez, G. Tur

Speech Technology & Research Laboratory

SRI International

Speech Recognition & Speaker Recognition

- Opposing goals:
  - Invariance to speaker differences (ASR)
  - Invariance to what was said (speaker recognition)
- (Largely) separate research communities
- ASR (and ASU) have always relied on speaker modeling
  - Speech/nonspeech segmentation
  - Diarization / speaker tracking
  - To help in feature normalization & model adaptation
- What can ASR do for speaker modeling ?

ASR & Speaker Recognition (2)

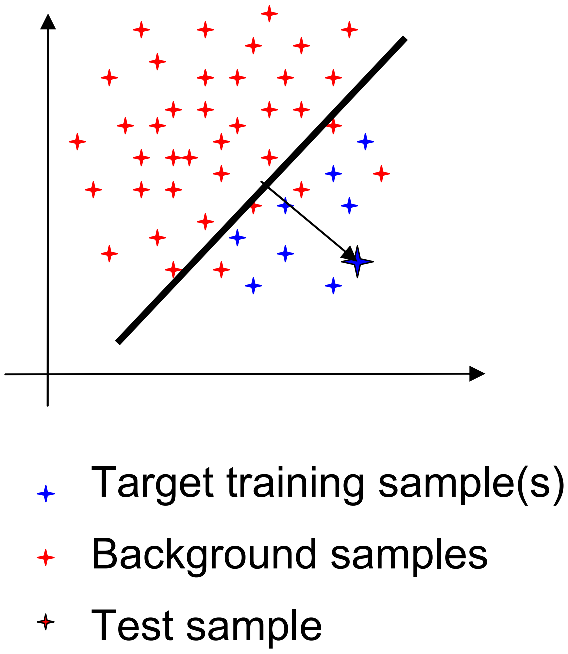
- Recent years have seen increasing ASR use in state-of-the-art speaker recognition
  - NIST speaker recognition evaluation
  - Telephone data (mostly)
  - Long (conversation-length) data samples
- Goal here:
  - Overview of what’s been done
  - Incite interest among ASR researchers
  - Point out challenges

“Generative” Speaker Verification

- GMM-UBM (Reynolds et al.)
- Models cepstral features
  - Feature normalization / mapping
- Training:
  - Train “background model” on a large population
  - “Speaker model” obtained via MAP-adaptation to enrollment data
- Testing:
  - Log-likelihood ratio between speaker and background model
  - Threshold for decision (accept / reject)

"Discriminative" Speaker Verification

- Mostly based on SVMs (Campbell et al.)
- Each conversation side = one point in feature space
- SVM trained to separate target from background samples
- Score = distance from test sample to decision hyperplane
- Linear kernel functions work well for most features tried to date
- Crucial step: how to map variable-length speech sample into fixed-length vector



How Can ASR Help?

- Phonetic / text conditioning
- Modeling “speaking style”
  - Pronunciation
  - Lexico-grammatical choice
  - Prosodic patterns
- ASR “by-product” features
  - MLLR-SVM features
- Challenges

Phonetic Conditioning

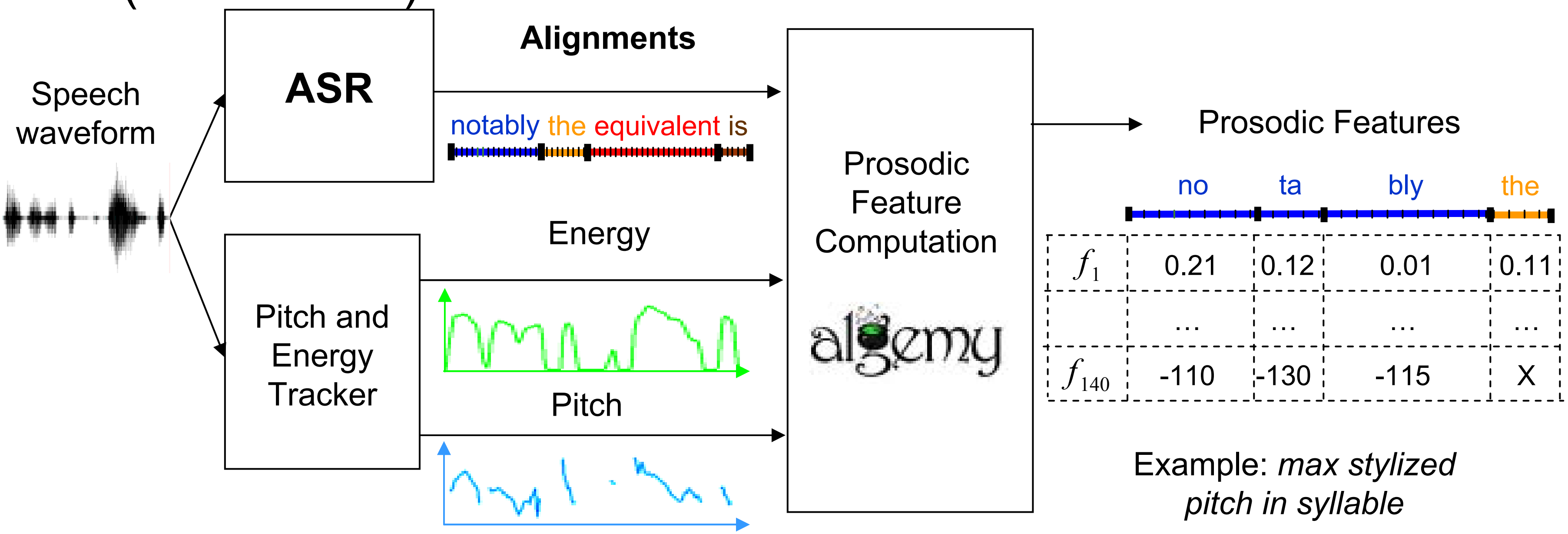
- Condition cepstral features on phone or word identities
- Removes within-speaker variability due to phonetic content
  - More like text-dependent speaker verification
- Possibly focuses features on regions of greater inter-speaker variability
  - E.g., discourse markers
- Explored by MIT-LL, Dragon, ICSI, et al.

Speaking Style Modeling

- Pronunciation modeling (many variants)
  - Phone N-gram frequency SVM vectors (Campbell)
  - Greatly enhanced by lattice decoding (Hatch et al.)
- Lexical & grammatical choice (Doddington)
  - Word N-gram frequency vectors
  - Distinguish “slow” from “fast” pronunciations for frequent words (Tur et al.)
- Prosodic modeling (Adami, Shriberg et al.)
  - Syllable-based energy, duation, and pitch features
  - Enhanced by lexical constraints

Prosodic Speaker Modeling

- SNERFs: Syllable-level Non-Uniform Extraction Region Features
- Compute a set of (140) duration, pitch and energy features on each syllable
- Transformations to fixed-length vectors using Fisher score (Ferrer et al.)

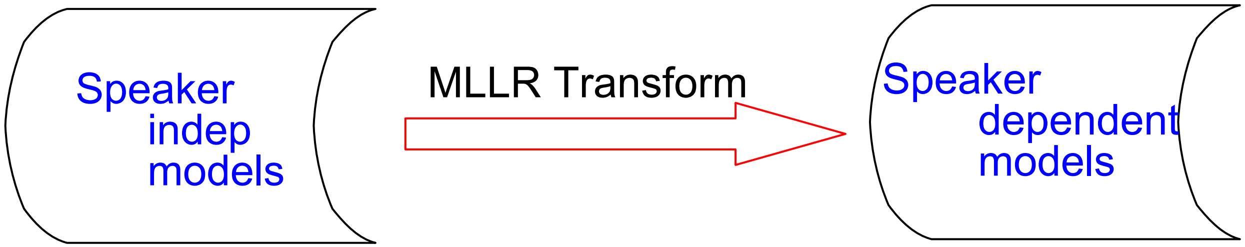


Recognizer-internal Features

- Idea: speech recognizer by-products encode speaker-specific information
  - Results of recognizer modeling inter-speaker variability
- Examples:
  - Sub-word unit duration modeling (Ferrer et al., Eurospeech ’03)
  - Speaker adaptation (MLLR) transforms (Eurospeech ’05, IEEE Trans. ASLP ’07)

MLLR Speaker Adaptation

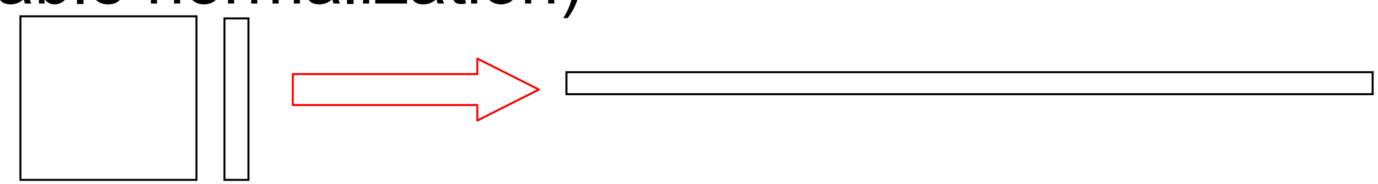
- Speech recognizer adapts speaker-independent model to best fit test speaker



- Adaptation transform estimated by Maximum Likelihood Linear Regression (MLLR)
  - Maximizes likelihood of test data under recognition hypothesis
- Transform rotates and shifts Gaussian means (= matrix + vector)

MLLR-SVM Speaker Recognition

- Idea: MLLR transform encapsulates what makes target speaker different from the "average speaker"
- Transforms are based on detailed, sequential speech models (unlike std. cepstral speaker models)
- Use transform coefficients as feature vector (after suitable normalization)



- Refinements:
  - Combine transforms for different phone classes
  - Combine transforms relative to different recognition models
- Model feature vectors with support vector machines (SVMs)

Results (on NIST SRE'06)

Use ASR

System	%EER
Cepstral GMM	4.75
Cepstral Polynomial SVM	5.07
Gaussian Supervector SVM	4.15
MLLR SVM	4.00
State/word duration GMM	16.03 / 22.24
Word + duration N-gram SVM	23.46
Prosodic SVM	10.41
Combination	**2.59**

- Note: MLLR and prosodic SVM best 2-system combination

Challenges

- Novel recognizer-based features
  - Non-linear adaptation transforms ?
- Need fast, accurate ASR for variable, “unexpected” conditions
  - Noisy environments
  - Variable channels
  - Nonnative speakers
- Mapping of ASR-based features across languages
- How to compare English to non-English ASR features (bilingual speakers) ?