Adaptation and Environmental Robustness

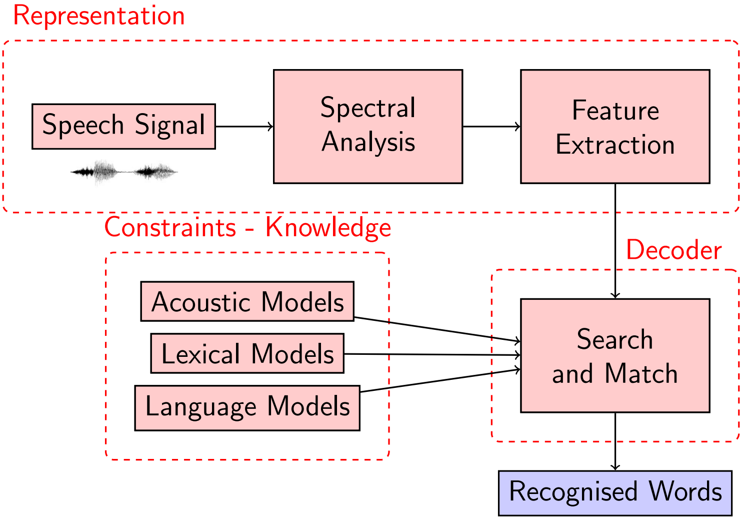
DT2118 Speech and Speaker Recognition

Giampiero Salvi

KTH/CSC/TMH giampi@kth.se

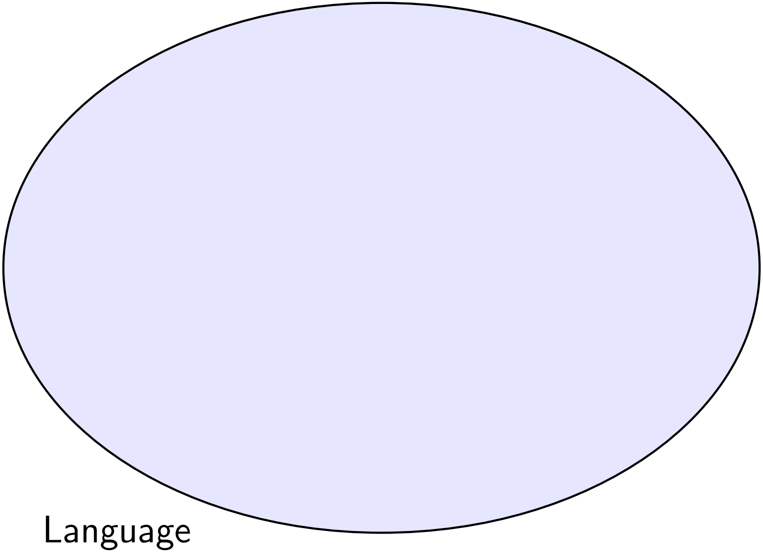
VT 2015

Components of ASR System

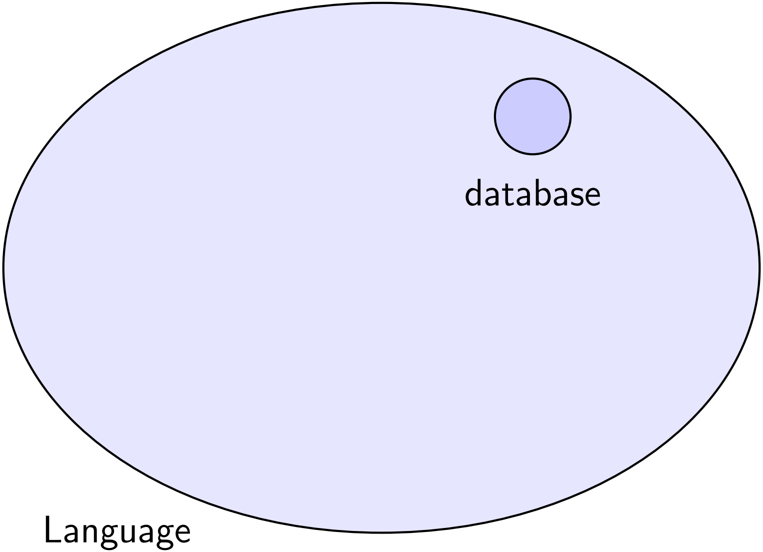


ASR seldom works out of the box!

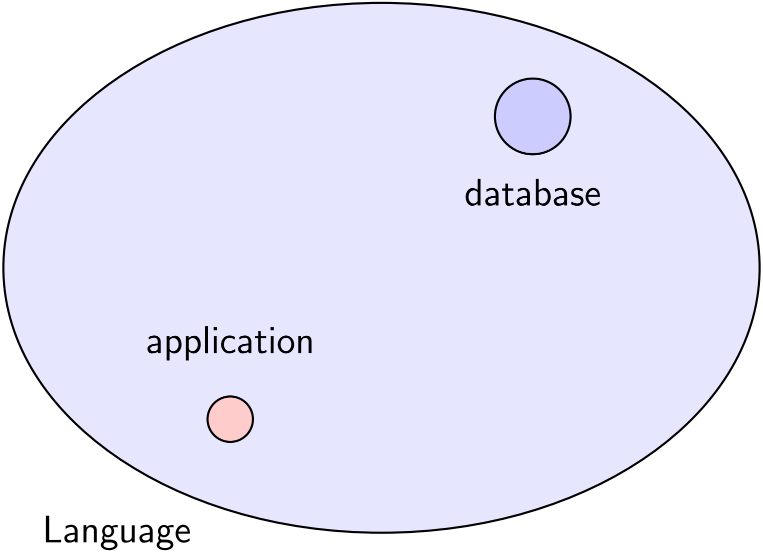
Why is it so hard?



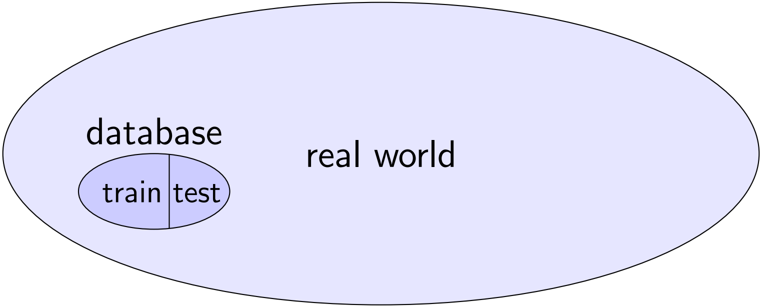
Why is it so hard?



Why is it so hard?



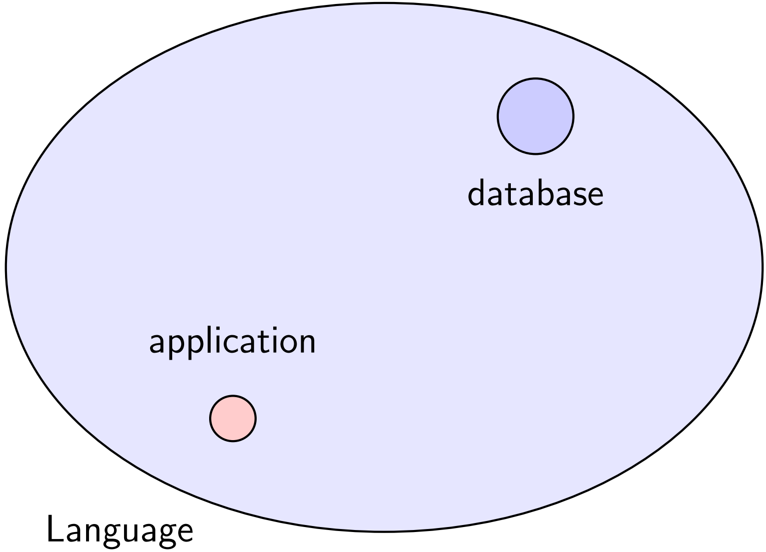
Misleading Training/Test Set



- mismatch between speakers
- unknown words or grammatical constructs
- environmental mismatch

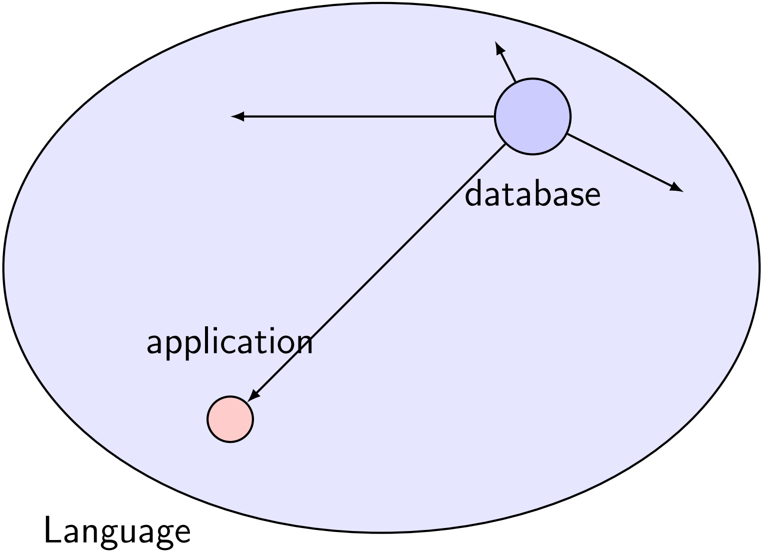
How do we cope with variability?

Ideally: models that generalise



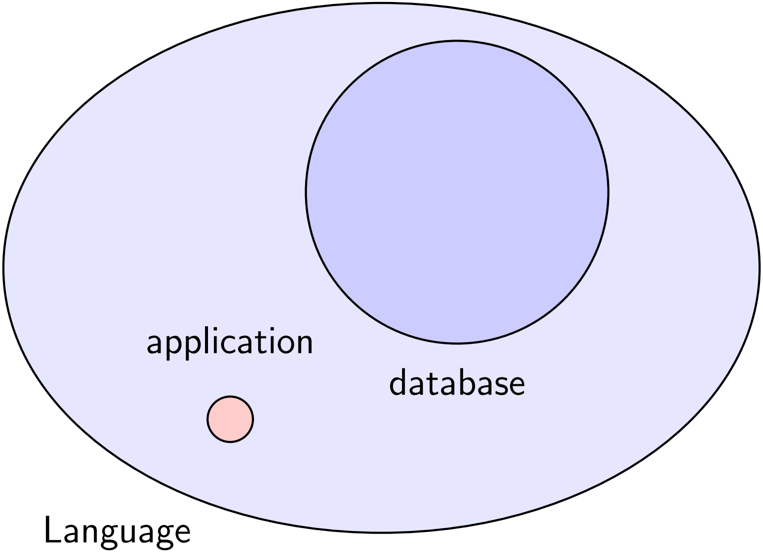
How do we cope with variability?

Ideally: models that generalise



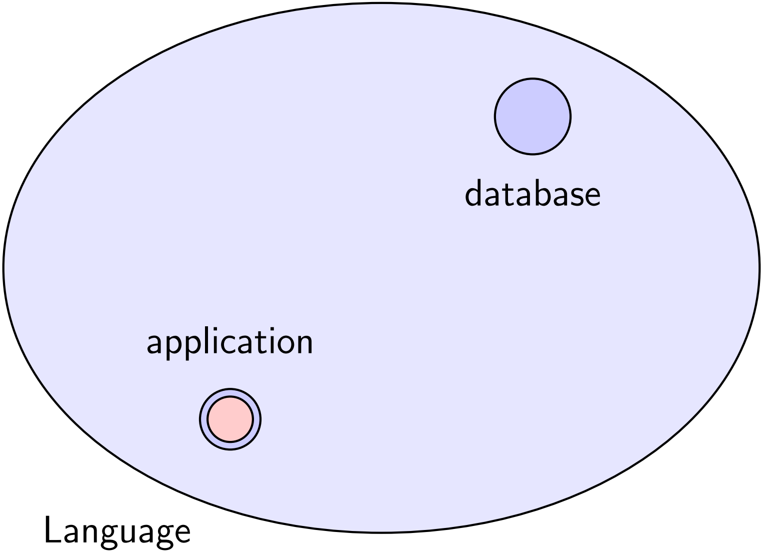
How do we cope with variability?

Large companies use insane quantities of data



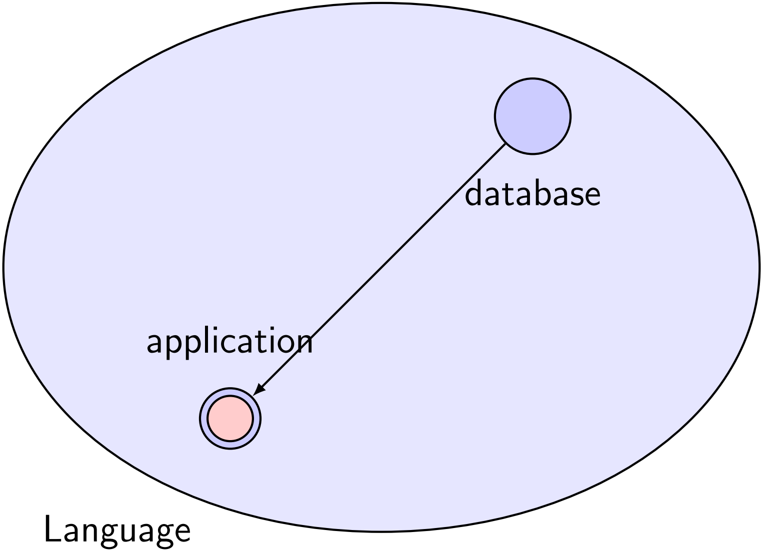
How do we cope with variability?

Adaptation



How do we cope with variability?

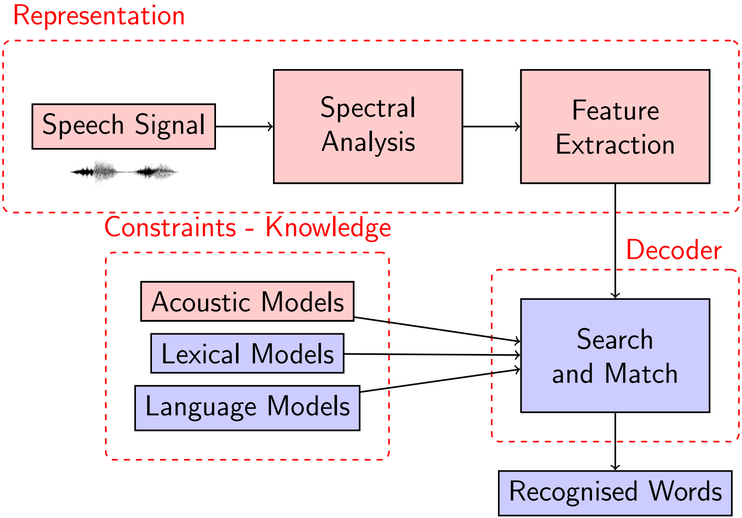
Adaptation



Adaptation

- adapt the acoustic features
- adapt the models (acoustic, language)

Components of ASR System



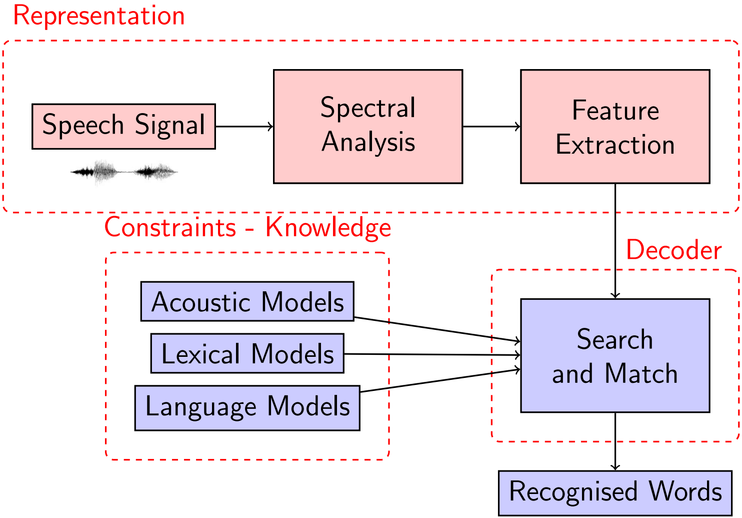
Outline

Outline

Adaptation and Speaker Characteristics

- anatomy, age, gender, dialect
- speaking style
- speaker adjustment to environment
- speaker adjustment to listener

Components of ASR System



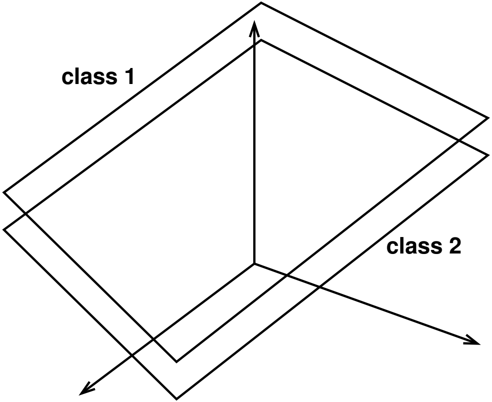
Feature Transformation

- general (PCA, LDA)
- explicit speaker modelling (VTLN)
- Speaker Specific (Statistical)

Principal Component Analysis (PCA)

- Aka Karhunen-Loewe transform
- Most used for dimensionality reduction
- new basis: ordered by data spread
- we can discard dimensions with small variation
- uncorrelated components

Problem with PCA



Linear Discriminant Analysis (LDA)

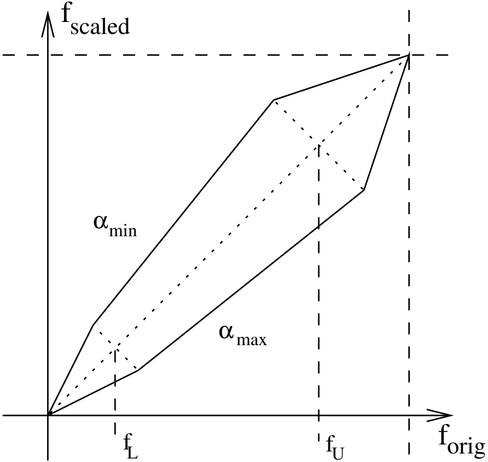
- supervised
- maximise ratio between:
    1. between class scatter matrix S_B
    2. within class scatter matrices S_W
- example: $J = \mathrm{tr}(S_W^{-1} S_B)$

Explicit Speaker Modelling

- focus on anatomy
- leave out all idiosyncrasies
- most salient parameter: Vocal Tract Length
- not correlated with body height
- possibly not correlated with formants [1]

[1] H. Hatano, T. Kitamura, H. Takemoto, P. Mokhtari, K. Honda, and S. Masaki. "Correlation between vocal tract length, body height, formant frequencies, and pitch frequency for the five Japanese vowels uttered by fifteen male speakers". In: *Proc. of Interspeech*. 2012

Vocal Tract Length Normalisation (VTLN)



VTLN factor

- vary factor α between α_{min} and α_{max} with regular steps
- run recogniser N times
- choose results with highest likelihood
- with adults α ranges between 0.8 and 1.25
- children to adults it ranges between 1.0 and 1.7

VTLN properties

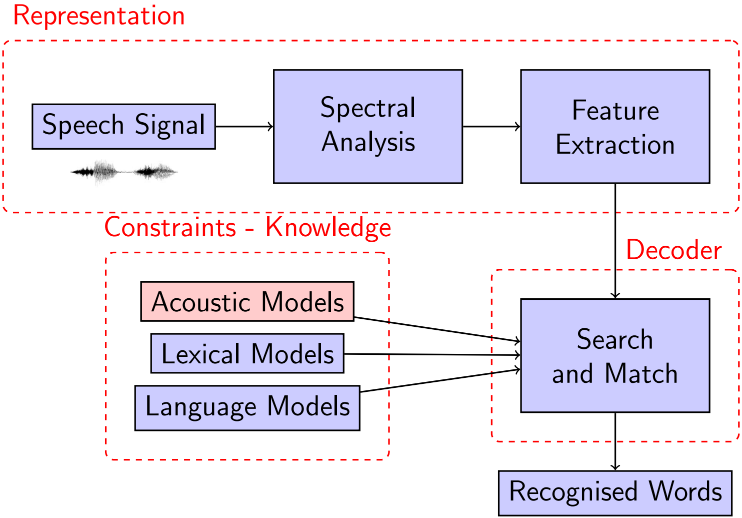
Advantages:

- no adaptation data needed
- simple transformation (one parameter)
- good improvements for children

Disadvantages:

- need to run recogniser N times
- phoneme dependent transforms (more parameters to tune)
- not powerful

Components of ASR System



Model Adaptation

- objective: adjust model parameters to new observations
- if plenty of data: retrain with Baum-Welch
- supervised vs unsupervised
- example: enrolment for dictation systems
- more often little data, no transcriptions
- use results from recogniser: risky

MAP Adaptation

- Maximum a Posteriori
- model parameters are stochastic variables
- define meaningful prior

$$\hat{\mu}_{ik} = \frac{\tau_{ik}\mu_{nw_{ik}} + \sum_{t=1}^{T} \zeta_t(i,k)x_t}{\tau_{ik} + \sum_{t=1}^{T} \zeta_t(i,k)}$$

MAP Problems

- need good prior
- all model parameters potentially updated
- need adaptation data to cover all phonetic classes

Maximum Likelihood Linear Regression (MLLR)

- constrained transformations
- reduce parameters to re-estimate
- linear regression:

$$\hat{\mu}_{ik} = A_c \mu_{ik} + b_c$$

- estimate A_c and b_c maximising likelihood
- one transform per regression class (example: for each phoneme)

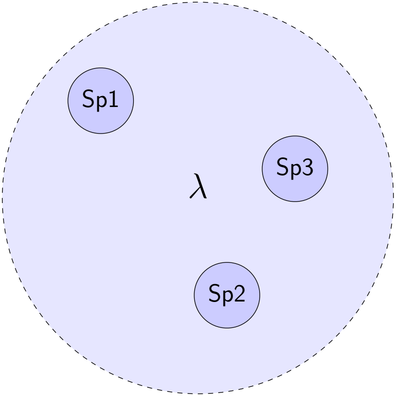
MLLR

if not enough data:

- use broader classes to be transformed
- for example one transform for fricatives, one for front vowels...

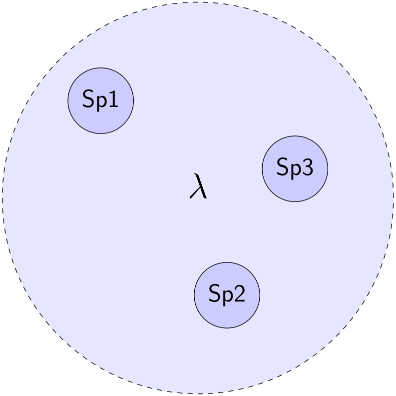
Speaker-Adaptive Training (SAT)

Conventional training

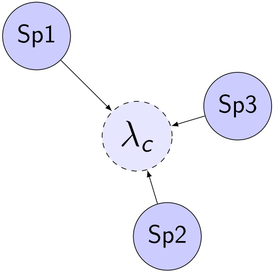


Speaker-Adaptive Training (SAT)

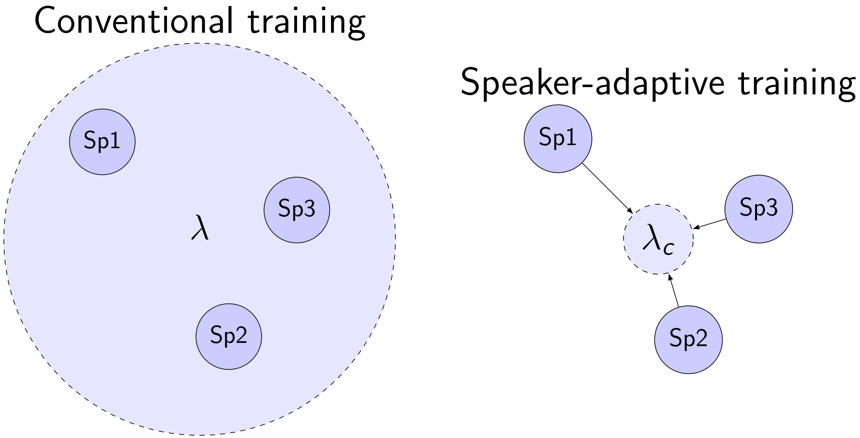
Conventional training



Speaker-adaptive training



Speaker-Adaptive Training (SAT)



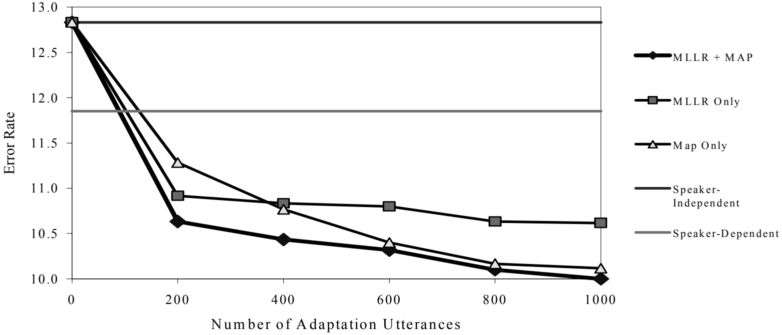
needs adaptation during recognition

Effects of Adaptation

Models	Relative Error Reduction (%)
CHMM	baseline
MLLR on mean only	12
MLLR on mean and variance	2
MLLR SAT	8

Here one regression class per phoneme was used (group all triphones with the same middle phoneme)

Combined MLLR and MAP

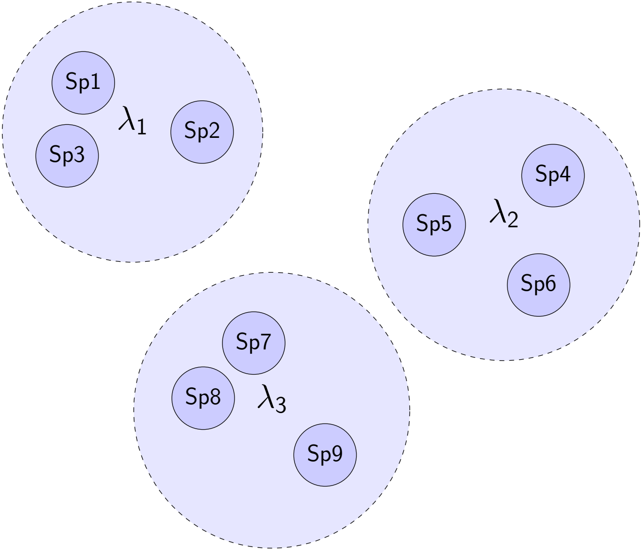


(from Huang, Acero and Hon)

Speaker Clustering

- MAP and MLLR require adaptation data
- not always available

Speaker Clustering



Speaker Clustering Variants

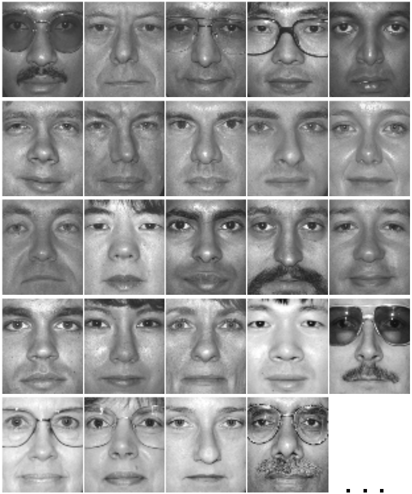
build models for each group

- at recognition time find best model
- can be integrated in search algorithm (pruning)
- combine with MLLR

use speaker dependent (SD) models and:

- represent each new speaker as linear combination of SD models
- eigenvoices

Eigenfaces

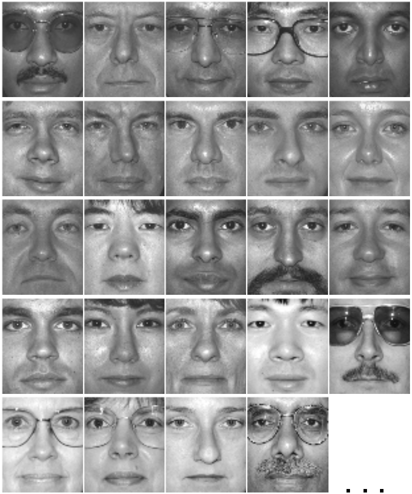


. . .

Faces from the FERET database

$$64 \times 73 \text{ pixels}$$
$$= 4672 \text{ dimensions!}$$

Eigenfaces



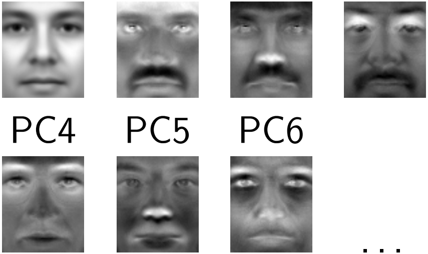
…

Faces from the FERET database

Eigenfaces

mean PC1 PC2 PC3

PC4 PC5 PC6



…

from 4672 dimensions to a small basis

Eigenvoices

- each voice (face) represented by model parameters
- thousand of dimensions
- subtract mean and run PCA
- during recognition find new speaker in eigenvoice space
- very little adaptation data required

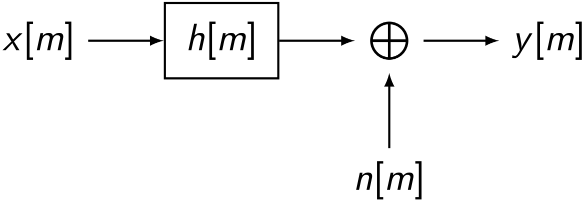
Outline

The Acoustical Environment

- additive noise
- reverberation (room)
- channel distortion (microphone, telephone line, codec)

A Model of the Environment

- A model of combined noise and reverberation effects



Additive Noise

- Stationary vs non-stationary
- White vs coloured (pink noise low frequency emphasis)

Additive Noise: Sources

Environment:

- air conditioner
- PC, keyboard
- cars
- other speakers (cocktail party effect)

The speaker:

- breath and puff noise
- lip smack
- mic and wire contacts

Additive Noise: Lombard Effect

- The speaker may change his voice when speaking in noise
- Reported recognition experiments are mainly performed in simulated noise
- do not capture this effect

Reverberation

- sound reflections from walls and objects in a room are added to the direct sound
- recognition systems are very sensitive to this effect
- strong sounds mask succeeding weak sounds
- reverberation radius: the distance from the sound source where the direct and the far sound fields are equal in amplitude

Typical office:

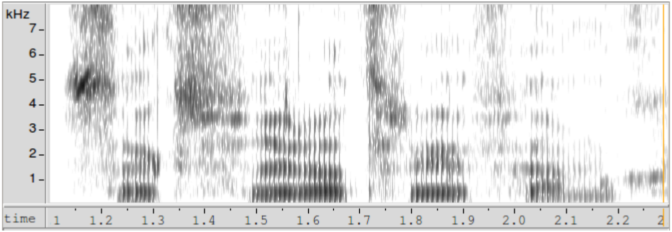
- reverberation time up to 100 ms
- reverberation radius 0.5 m

Acoustical Transducers

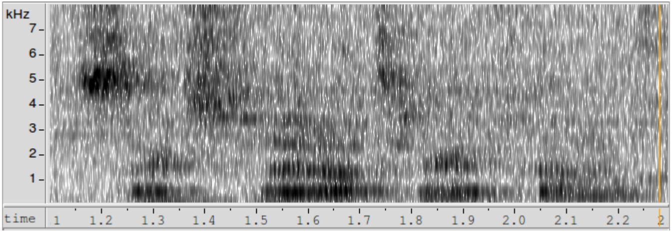
- Close-talk microphones
  - background noise is attenuated
  - sensitive to speaker non-speech sounds
  - positioning is critical
    - mouth corner recommended
    - plosive bursts may saturate the mic signal if right in front
- Far field microphones
  - pick up more background noise
  - positioning less critical
- Most popular type: condenser microphone
- Multimicrophones - Microphone Arrays
  - Adjustable directivity

Near and far distance microphones

Headset



2 m distance



Environment Compensation Pre-processing

Compensate with signal processing (feature extraction)

- Spectral Subtraction
- Cepstral Mean Normalisation (CMN)
- Real-time Cepstral Normalisation
- RASTA

Adaptive Echo Cancellation

- also used in voice over IP
- adjust parameters of a FIR filter online
- The Least Mean Squares (LMS) Algorithm

Multi-microphone Speech Enhancement

- Microphone Arrays (beam forming)
- Blind Source Separation

Spectral Subtraction

Assumption 1, noise additive:

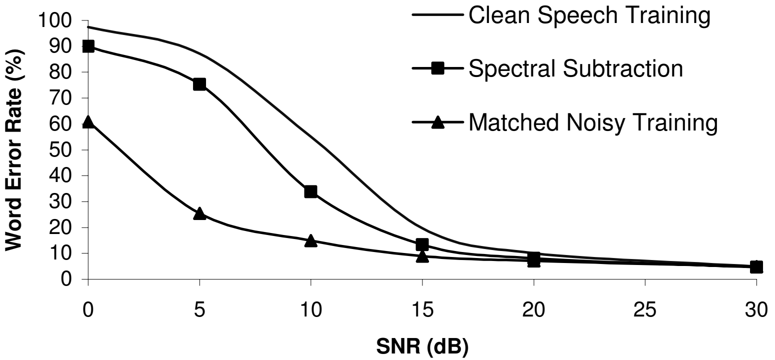
$$y[m] = x[m] + n[m]$$

Assumption 2, signal and noise decorrelated: In frequency domain

$$|Y(f)|^2 \approx |X(f)|^2 + |N(f)|^2$$

- estimate $|N(f)|^2$ in silent segments
- subtract $|N(f)|^2$ from $|Y(f)|^2$

Spectral Subtraction



(from Huang, Acero and Hon)

Cepstral Mean Normalisation (CMN)

- Subtract the average cepstrum over the utterance from each frame
- Compensates for different frequency characteristics

CMN Problem: Phonetic Information

The average cepstrum contains both channel and phonetic information

- The compensation will be different for different utterances, especially for short utterances ($<$ 2–4 sec)
- Still provides robustness against filtering operations
- For telephone recordings, 30% relative error reduction
- Some compensation also for differences in voice source spectra

Real Time CMN

Problem: Need whole utterance to computer average

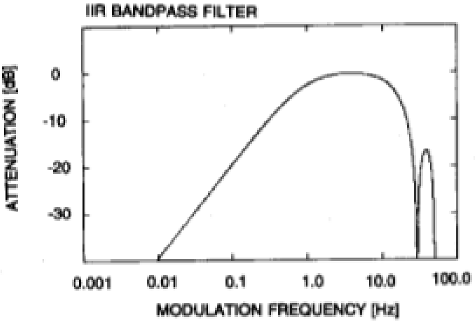
- Not suitable for live recognition
- use high-pass filter with about 5 sec time constant

$$\bar{x}_t = \alpha x_t + (1 - \alpha)\bar{x}_{t-1}$$

- other filters are also popular
- need good initialisation

RASTA: RelAtive SpecTrAl

- Hearing-inspired bandpass filtering of filterbank amplitude envelopes
- Removes long-term bias in the signal but leaves syllable rate modulation mainly unchanged



[2]

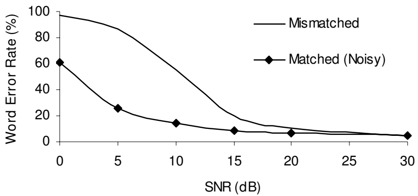
[2] H. Hermansky and N. Morgan. "RASTA Processing of Speech". In: *IEEE Trans. Speech Audio Process.* 2.4 (Oct. 1994), pp. 578–589

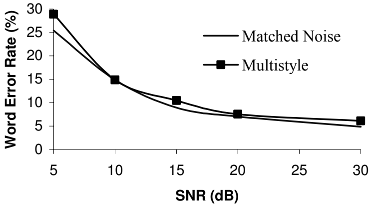
Environmental Model Adaptation

- Retraining on Corrupted Speech
- Model Adaptation
- Parallel Model Combination
- Retraining on Compensated Features

Retraining on Corrupted Speech

- If the distortion is known, then models can be trained by distorting the training data in this way (noise added, filtering)
- Several distortions can be used in parallel (multi-style training)
- Ignores the effect of the distortion on the speaker





Model Adaptation

- Same methods possible as for speaker adaptation (MAP and MLLR)
- MAP requires large adaptation data - impractical
- MLLR needs ca 1 min

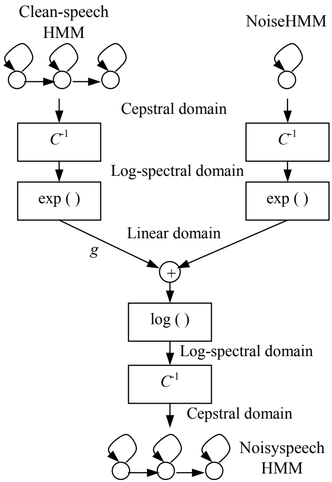
MLLR for Noise Adaptation

one regression class and only bias

- Combined speech recognition and MLLR estimation of the distortion
- Slightly better than CMN, especially for short utterances
- Slower than CMN since two-stage procedure and model adaptation as part of recognition

Parallel Model Combination

- Gaussian distribution converts into Non-Gaussian distribution
- No problem, a Gaussian mixture can model this
- Non-stationary noise can be modelled by having more than one state at the cost of multiplying the total number of states

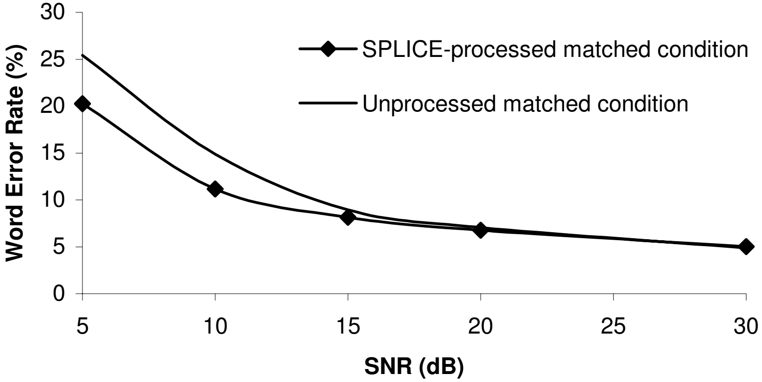


Environments

- Office - 200 speakers
  - at least 4 different rooms (close and far wall)
  - close talk, hands-free, medium distance (0.75 m), far distance (2 m)
- Public Place - 200 speakers
  - at least 2 locations: hall > 100 m^2 and outdoors
- Entertainment - 75 speakers
  - at least 3 different living rooms with radio on/off,
- Car - 75 speakers
  - middle or upper class car (VW Golf, Opel Astra, Mercedes A Class, Ford Mondeo, Mercedes C Class, Audi A6)
  - motor on/off, city 30-70, road 60-100, highway 90-130 km/h
- Children
  - 50 speakers (children's room)

Retraining on Compensated Features

- The algorithms for removing noise from noisy speech are not perfect
- Retraining can compensate for this



Modelling Non-stationary Noise

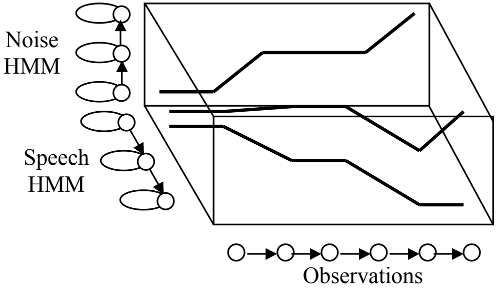
- speaker noise (clearing voice, breathing, lip smack)
- door slams, keyboard, other speakers
- can be between words or overlap with them

Approach 1: Explicit Noise Modelling

- Include non-speech labels in the training data
- Perform training
- Update the transcription with optional noise between words
- Retrain
- Problem when speech and noise overlap in time

Approach 2: Speech/noise decomposition

- During recognition
- 3-dimensional Viterbi
- Computationally complex



Outline

Confidence Measures

- errors are unavoidable
- in a larger system essential to diagnose errors
- dialogue system may be able to correct them

Confidence Measures

If accurate, $P(\text{words}|\text{sounds})$ best confidence measure
Problem: in

$$P(\text{words}|\text{sounds}) = \frac{P(\text{sounds}|\text{words})P(\text{words})}{P(\text{sounds})}$$

$P(\text{sounds})$ is usually not computed ($\arg\max$)
In general

$$P(\text{sounds}) = \sum_{\text{words}} P(\text{sounds}|\text{words})P(\text{words})$$

Filler models

$$P(\text{sounds}) = \sum_{\text{words}} P(\text{sounds}|\text{words})P(\text{words})$$

- General purpose recogniser
- should be able to "fill the holes" of the target recogniser
- often loop of phones
- any word sequence is allowed (including out of vocabulary)
- can be done word by word (segmentation from target recogniser)

Word Spotting

- do not recognise all the words
- only small number of keywords
- can build models of “antiwords”

Transformation Models

Use subword units in the confidence. If a word has N phones:

$$CS(\text{word}) = \sum_{i=1}^{N} f_i(CS(\text{phone}_i))$$

where

$$f_i(x) = a_i x + b_i$$

and can be optimised on the training data

Combination Models

use combination of several features:

- word stability when changing language model parameters
- average number of active hypothesis at word end
- acoustic score per frame within words normalised to active senones
- ...

A linear classifier works well