Resource Optimized Speech Recognition using Kullback-Leibler Divergence based HMM

Ramya Rasipuram

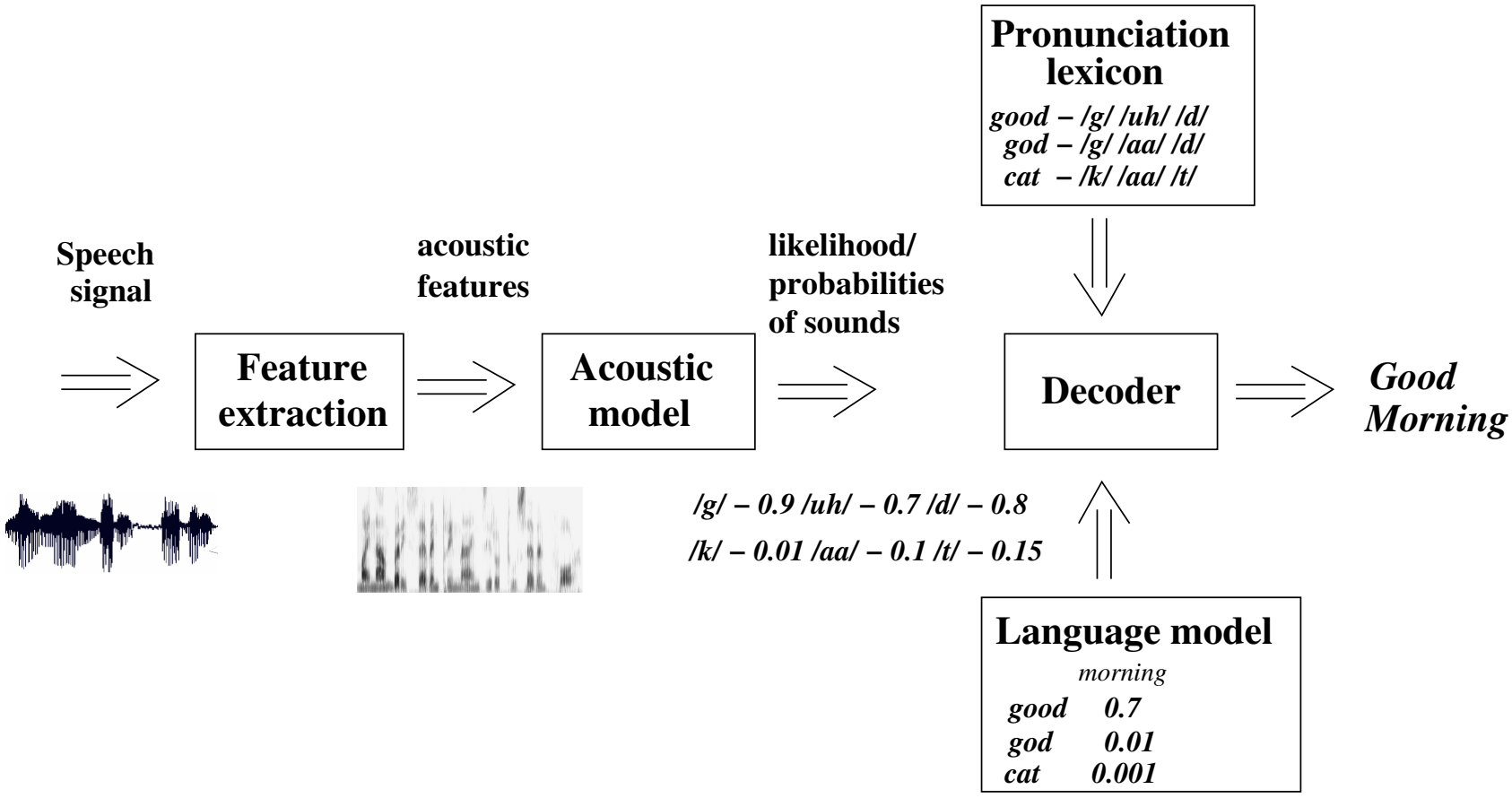
David Imseng, Marzieh Razavi, Mathew Magimai Doss, Herve Bourlard

idiap RESEARCH INSTITUTE

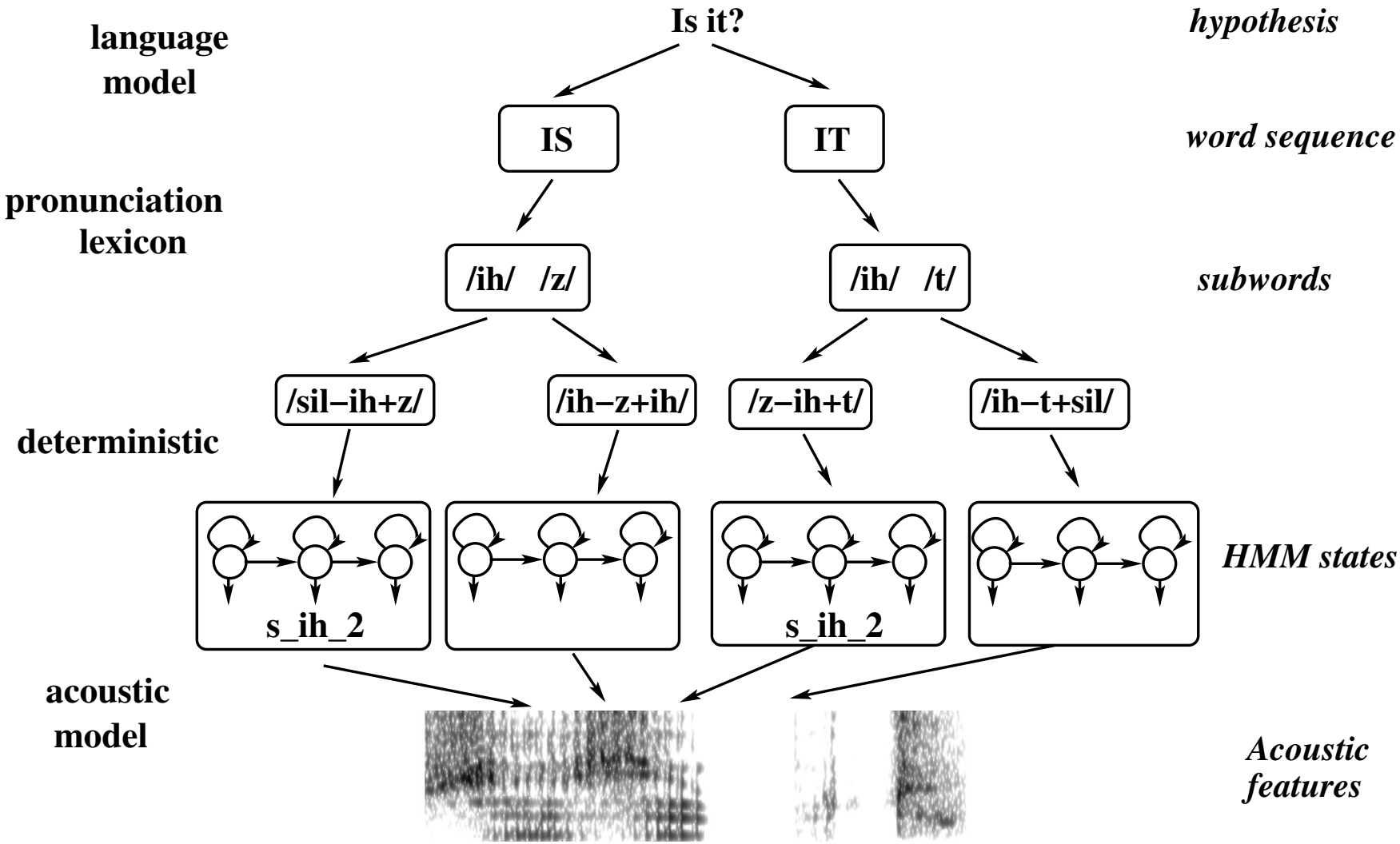
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

24 October 2014

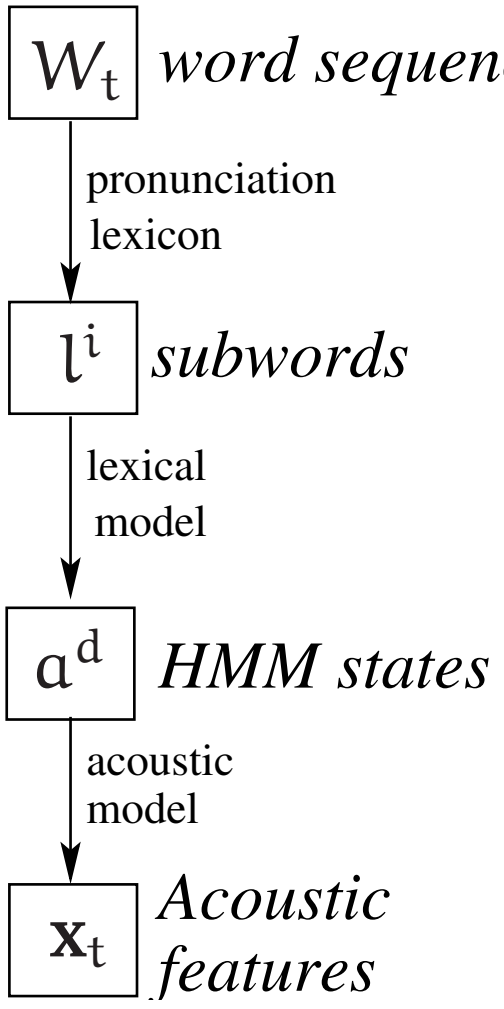
Automatic Speech Recognition (ASR)



Hidden Markov Models (HMMs) for ASR



Standard HMM-based ASR



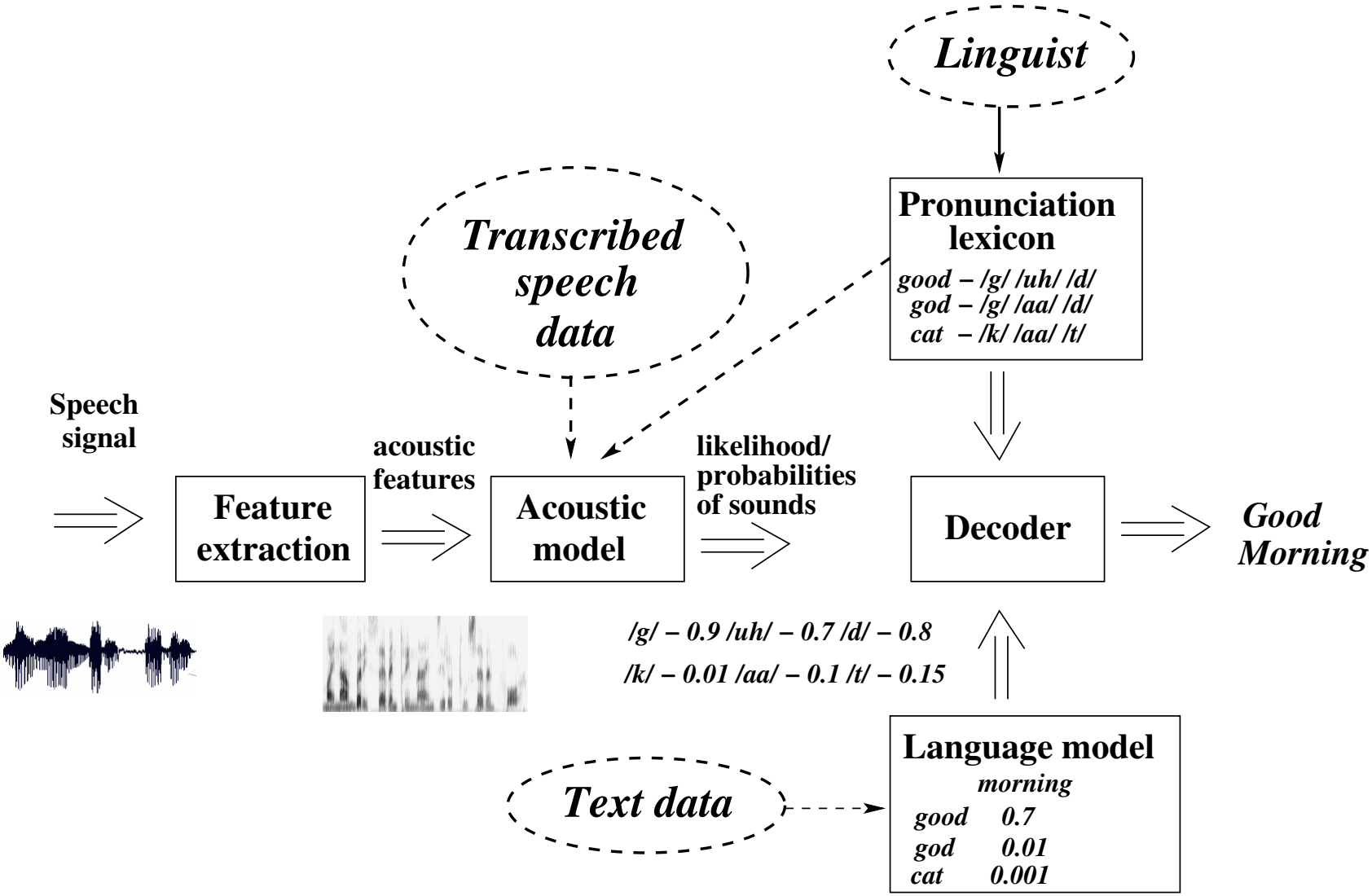
Acoustic Model:

1. GMM $\rightarrow$ HMM/GMM
2. ANN $\rightarrow$ Hybrid HMM/ANN

Lexical Model:

1. Deterministic $\rightarrow$ decision trees

Resources for ASR



ASR for Under-Resourced Languages

- Limited or no transcribed speech
- Linguistic expertise may not be available
- Limited or no text resources

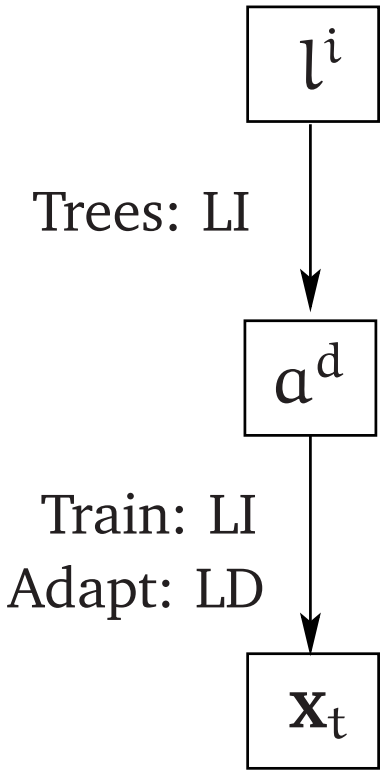
Limited Transcribed Speech Data

- Borrow resources
- Sounds of languages or phonemes can be shared across languages
- *Pronunciation lexicon is important*

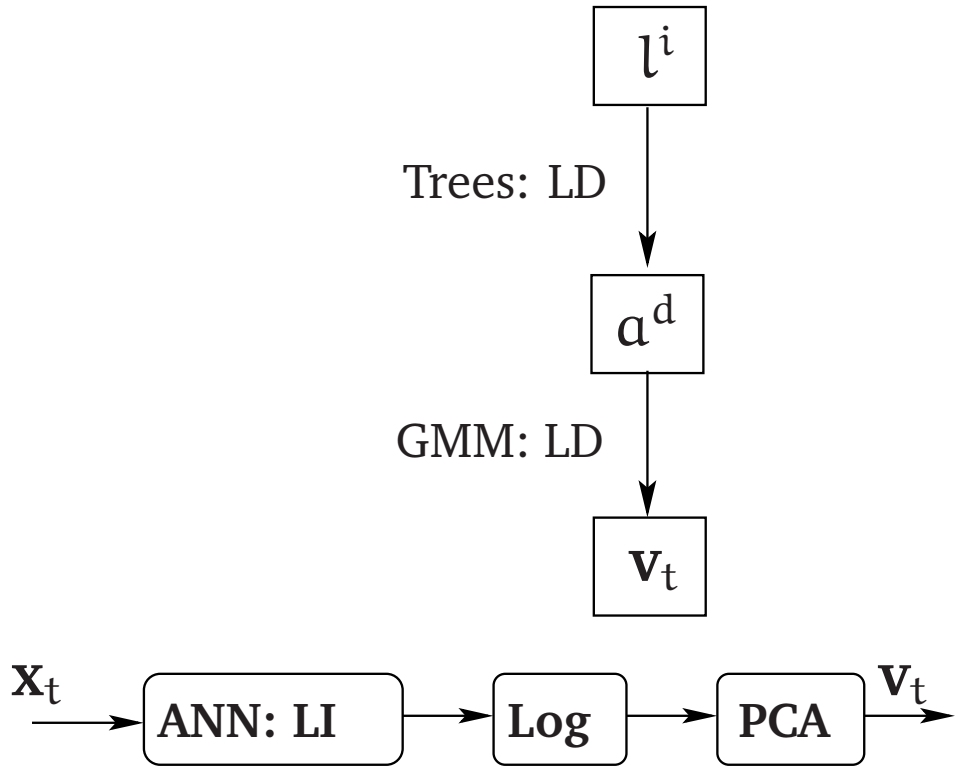
	word	**pronun**	**word**	**pronun**
English	a	eI	b	b i:
Italian		a		b @
German		a, a:		b e:
French		A, a, E		b &
Greek	α	a	β	v

Conventional Approaches

MAP adaptation



Tandem

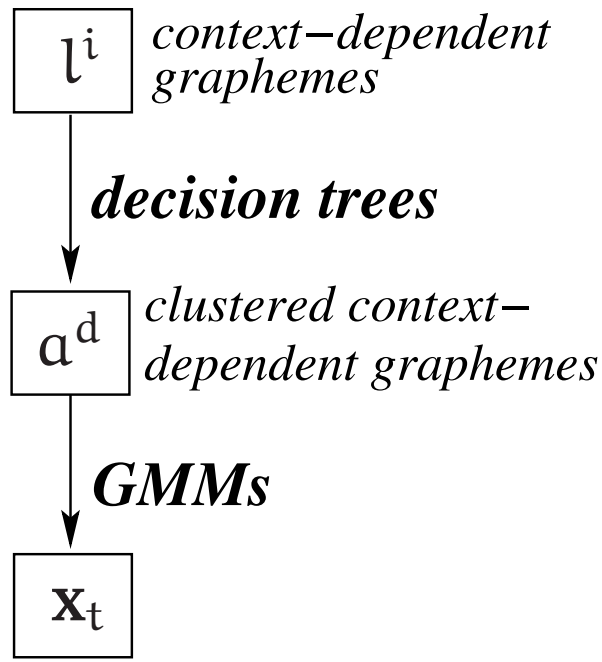


LI: language-independent data from resource-rich language(s)
LD: language-dependent data from under-resourced language

No Pronunciation Lexicon

1. Pay a linguist ← *expensive, time consuming*
2. Graphemes as subword units ← *easy, not optimal*

Word	Phone	Grapheme
Read	**r eh d** **r iy d**	**R E A D**

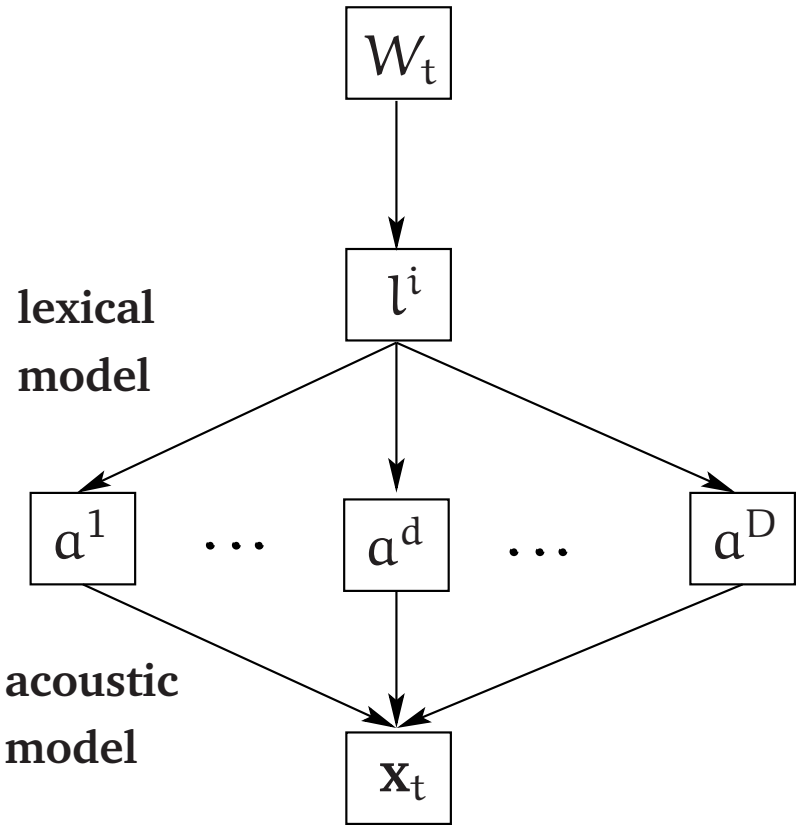


Limited Transcribed Speech and No Pronunciations

- Multilingual graphemes?
- Worse than monolingual grapheme-based ASR

Language	**word**	**pronun**	**word**	**pronun**
English	a	a	b	b
Spanish		a		b
Italian		a		b
German		a		b
French		a		
Greek	α	?	β	?

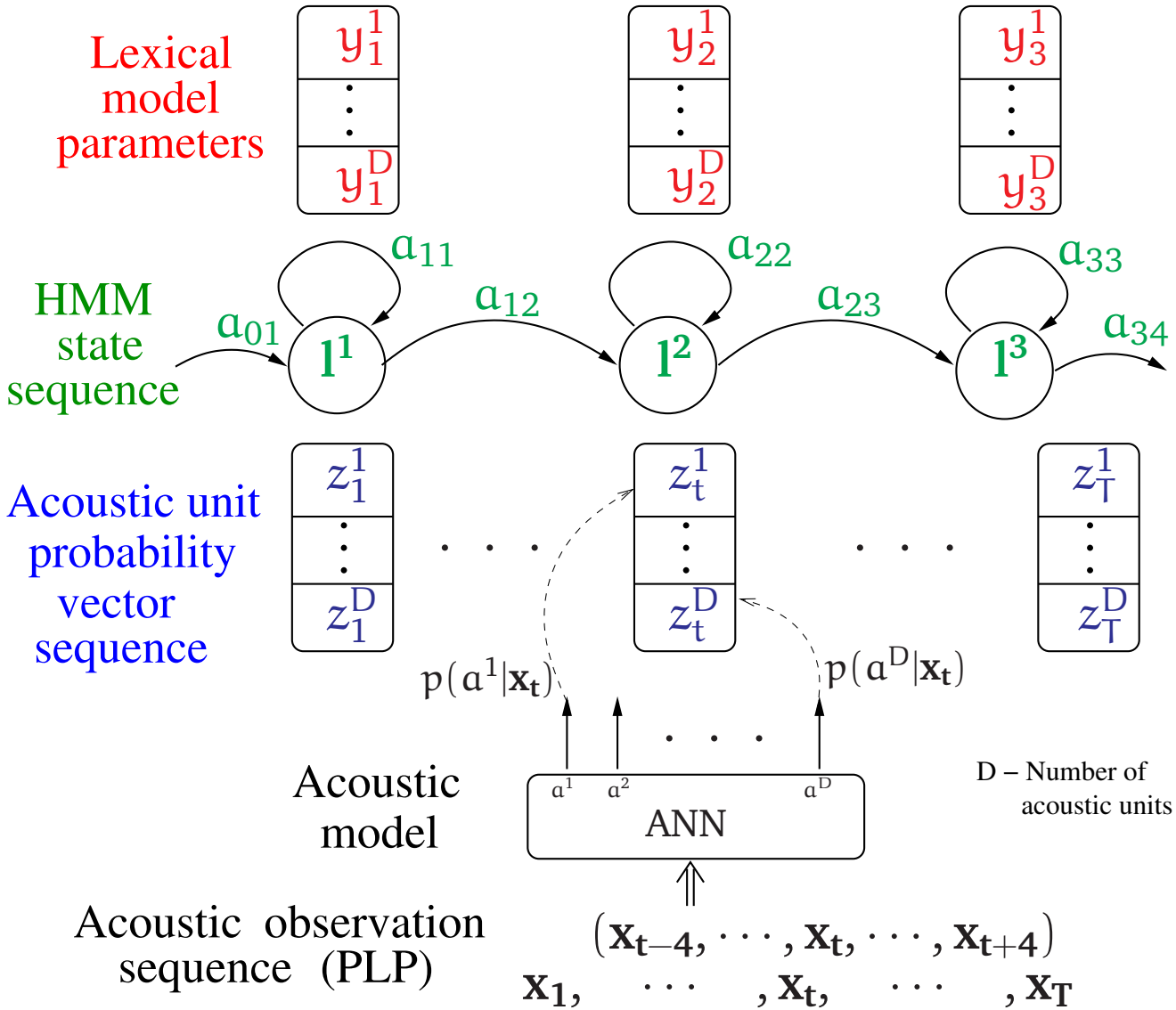
Probabilistic Lexical Modeling



- $0 < P(a^d|l^i) < 1$, $\sum_{d=1}^{D} P(a^d|l^i) = 1$
- Lexical model: $\theta_l = \{\mathbf{y}_i\}_{i=1}^{I}$ $\mathbf{y}_i = [y_i^1, \ldots, y_i^D]^T$, $y_i^d = P(a^d|l^i)$
- θ_l estimated by training a HMM
- Kullback-Leibler divergence based HMM (KL-HMM)[1]

[1] Aradilla G., "Acoustic Models for Posterior Features in Speech Recognition", EPFL PhD Thesis, 2008

KL-HMM System



KL-HMM

- Features: posterior probability estimates of acoustic units

$$\mathbf{z}_t = [z_t^1, \ldots, z_t^d, \ldots, z_t^D]^T, \quad z_t^d = p(a^d|\mathbf{x}_t)$$

- State distribution: categorical distribution

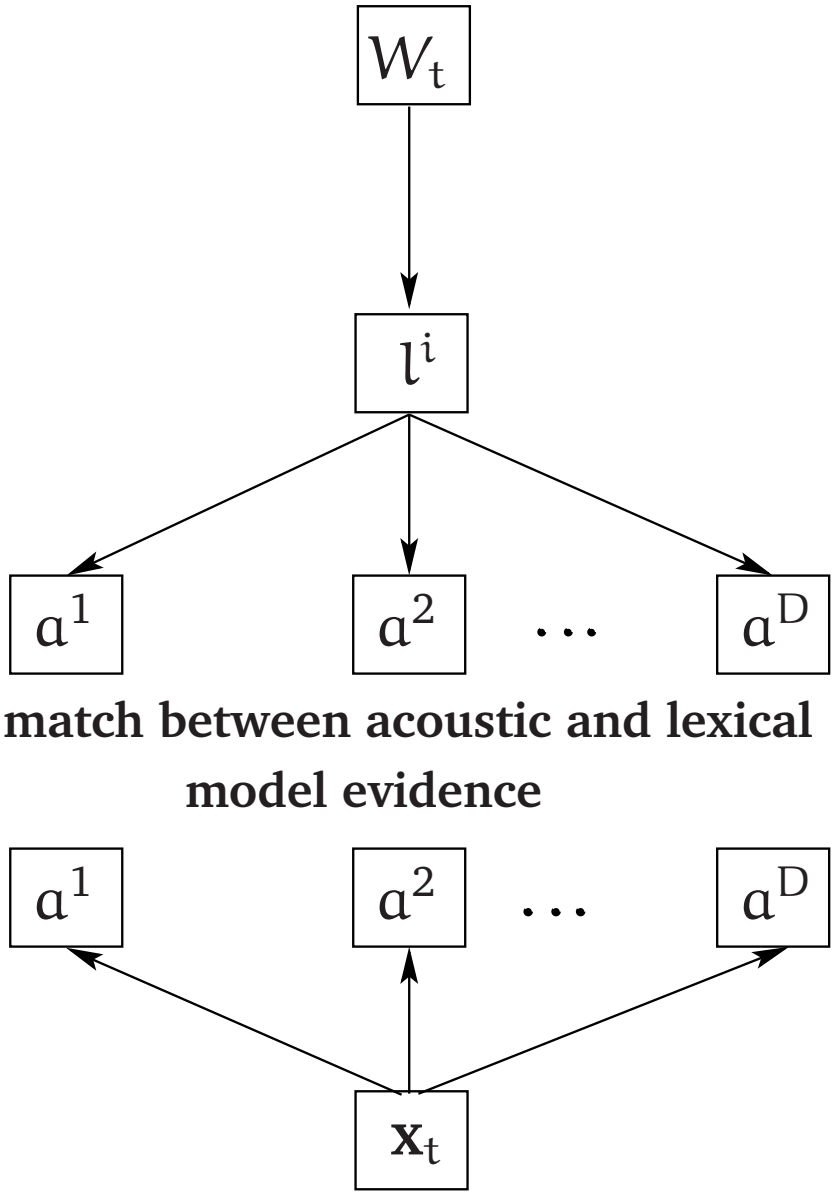
$$\mathbf{y}_i = [y_i^1, \ldots, y_i^d, \ldots, y_i^D]^T, \quad y_i^d = P(a^d|l^i)$$

- Local score: Kullback-Leibler (KL) divergence

$$S(\mathbf{z}_t, \mathbf{y}_i) = \sum_{d=1}^{D} z_t^d \log\left(\frac{z_t^d}{y_i^d}\right)$$

- Parameter estimation: Viterbi Expectation Maximization algorithm
  - cost function based on KL-divergence

Decoding



language model

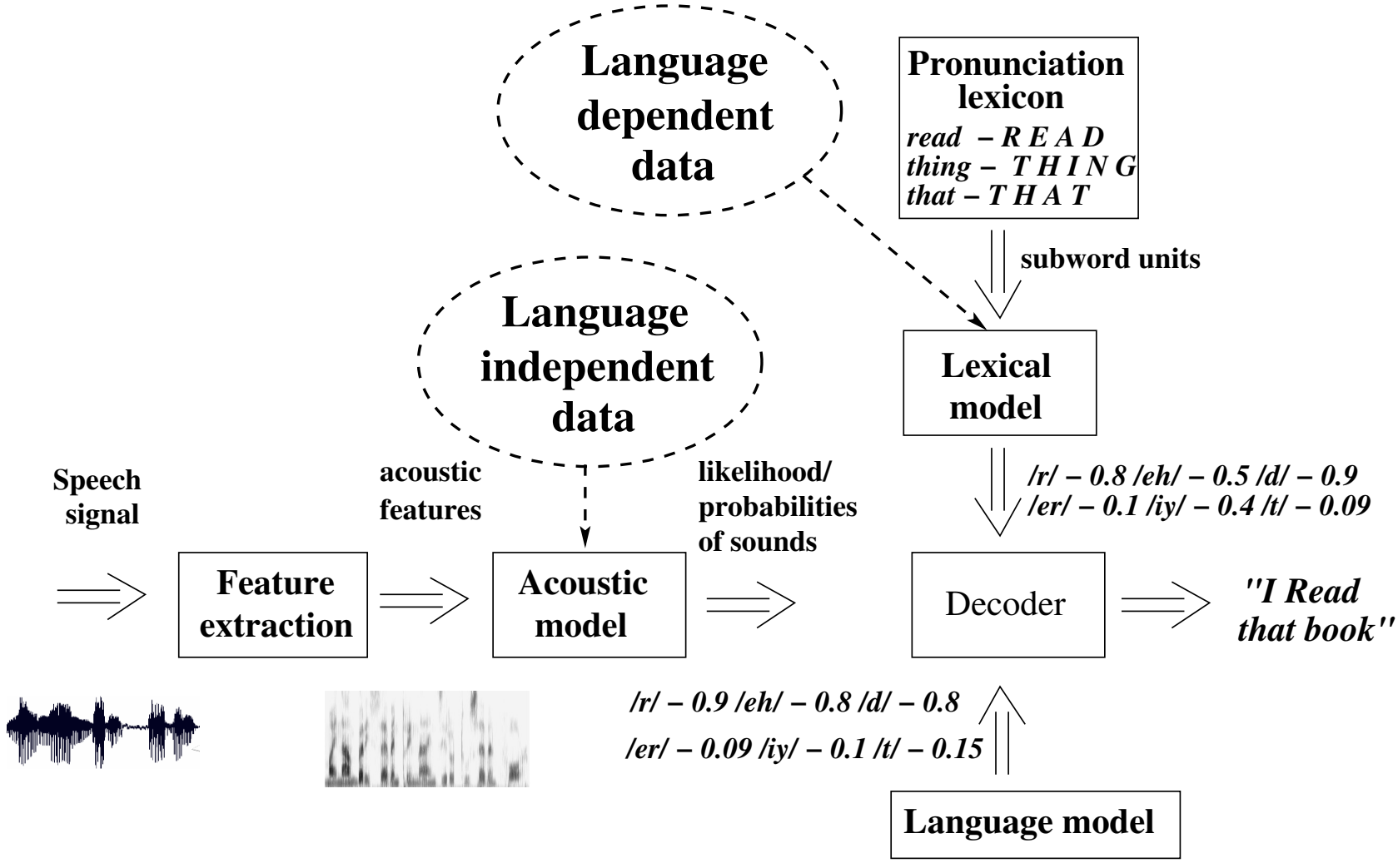
pronunciation lexicon

lexical model

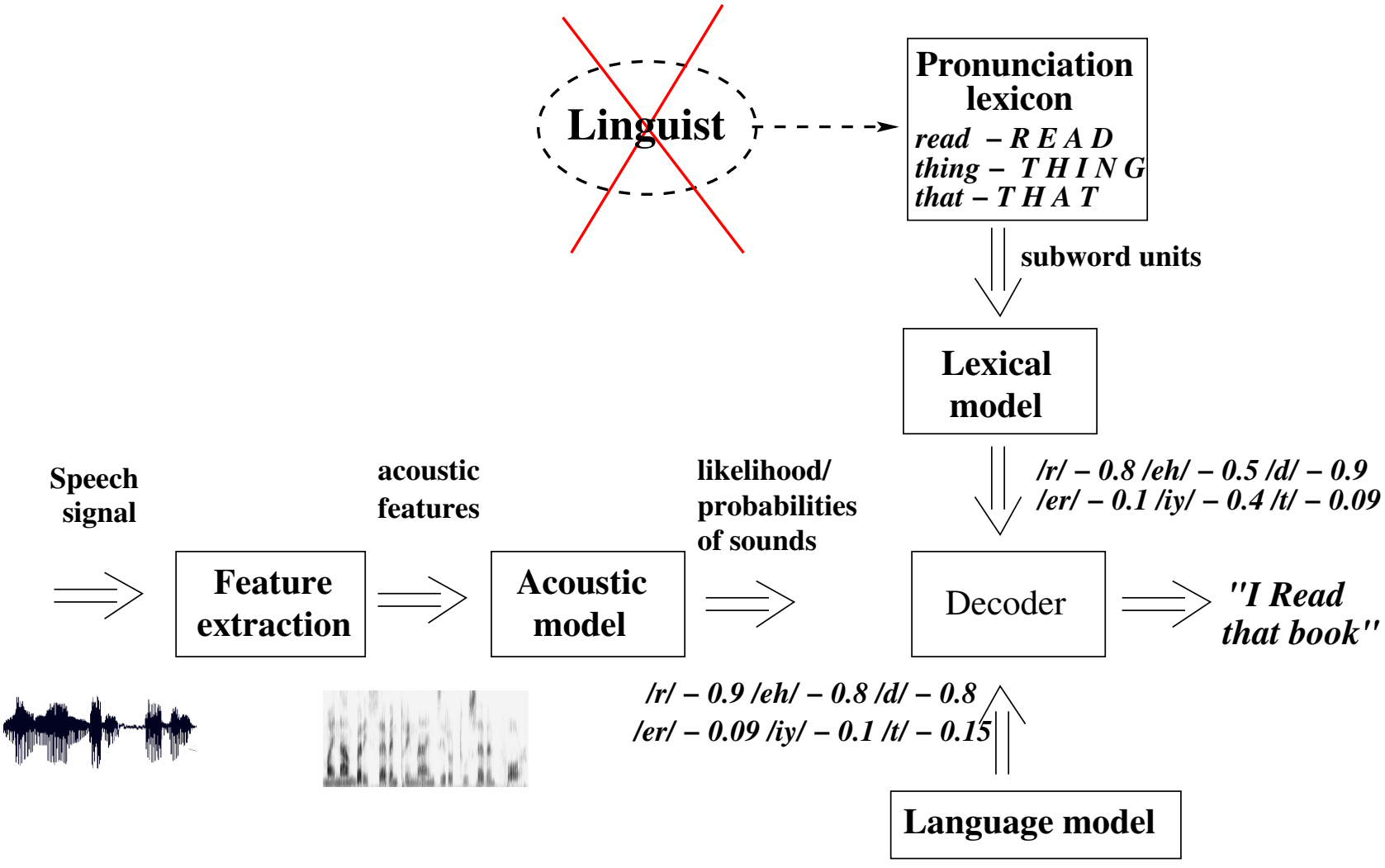
KL-divergence

acoustic model

Advantage 1: Resource Optimization



Advantage 2: Grapheme Subword Units

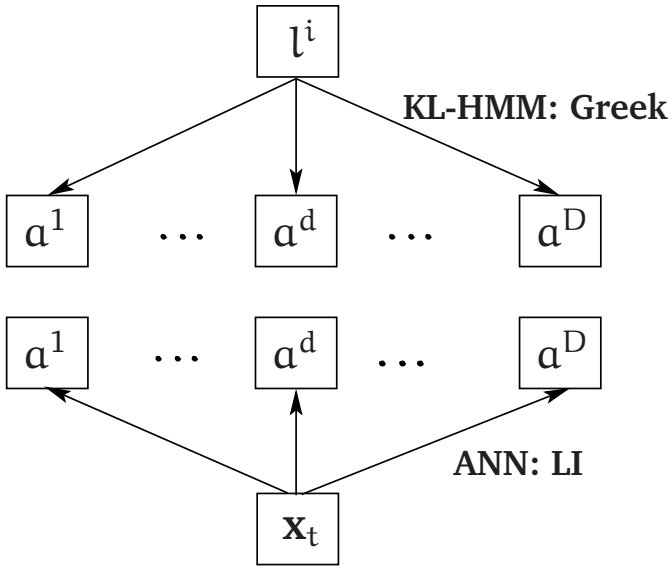


Task

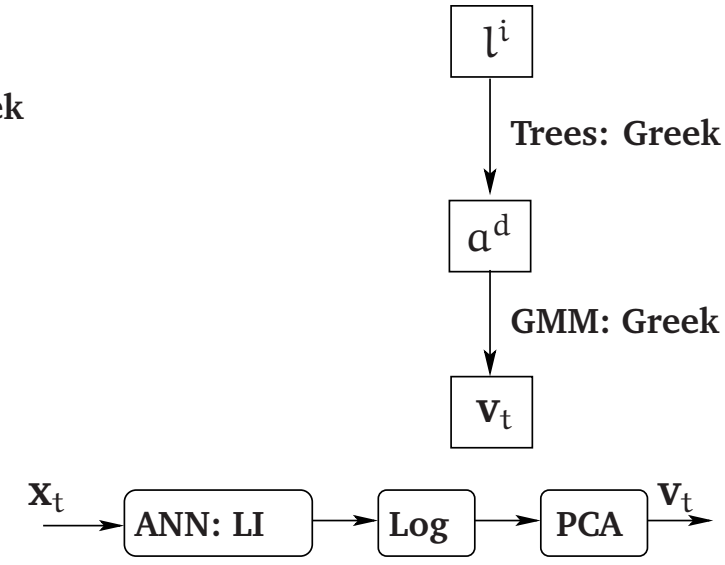
- Build speech recognition system for Greek but with
  - Limited transcribed speech data
  - No pronunciation lexicon
- Borrow resources from French, German, Italian, Spanish and English $\leftarrow$ language independent (LI) data

Systems

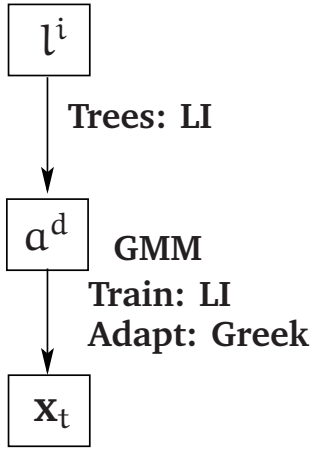
KL–HMM



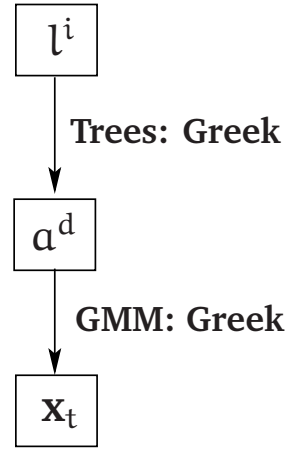
Tandem



MAP adaptation

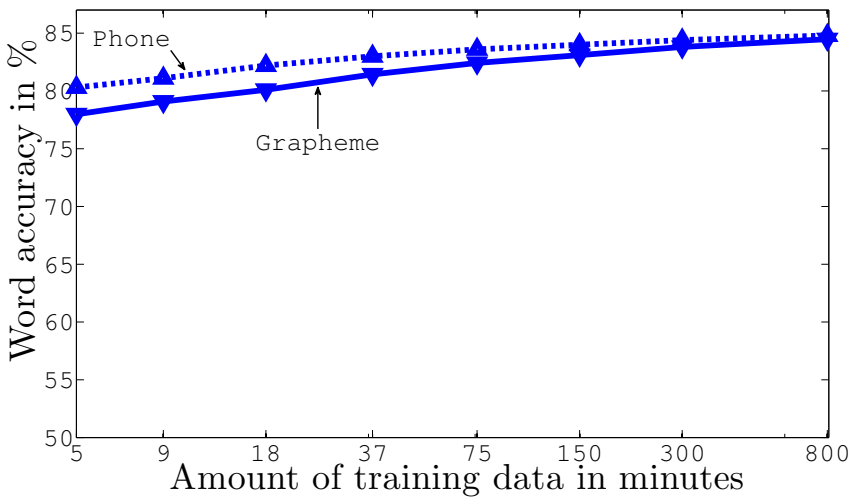


HMM/GMM

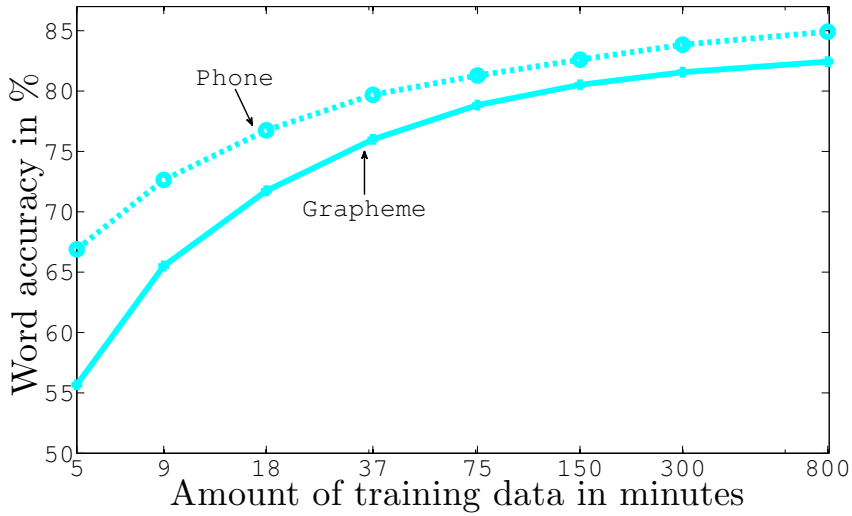


Results

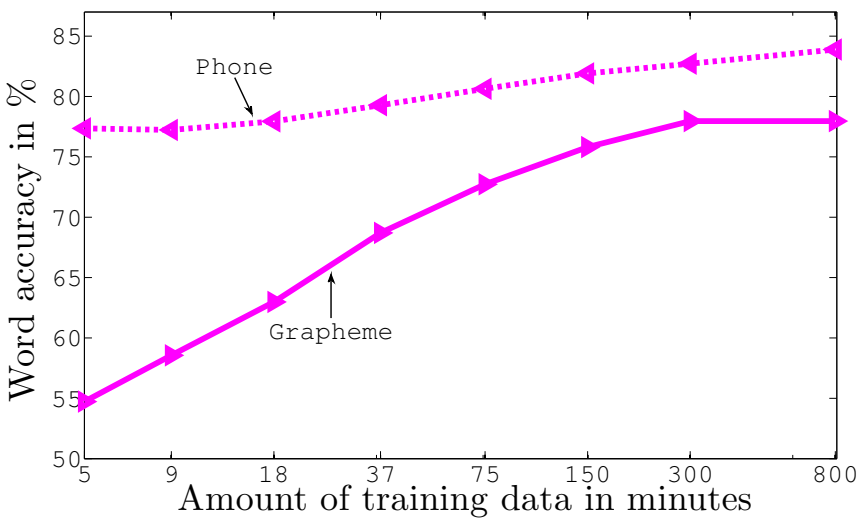
KL-HMM



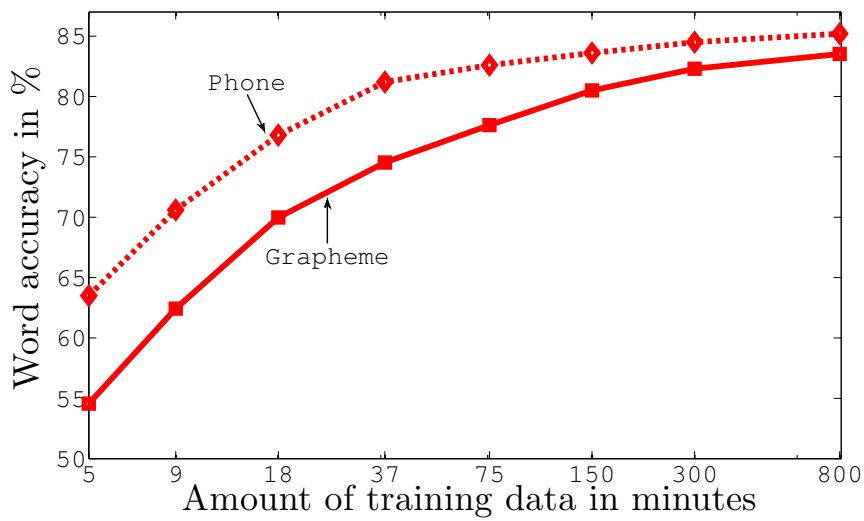
Tandem



MAP adaptation



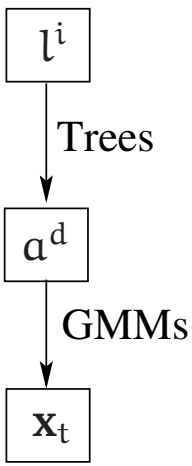
HMM/GMM



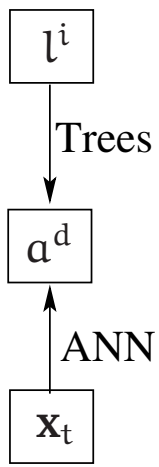
Advantage 3: Pronunciation Variability Modeling

- Train data: Native speech
- Test data: Native and non-native speech

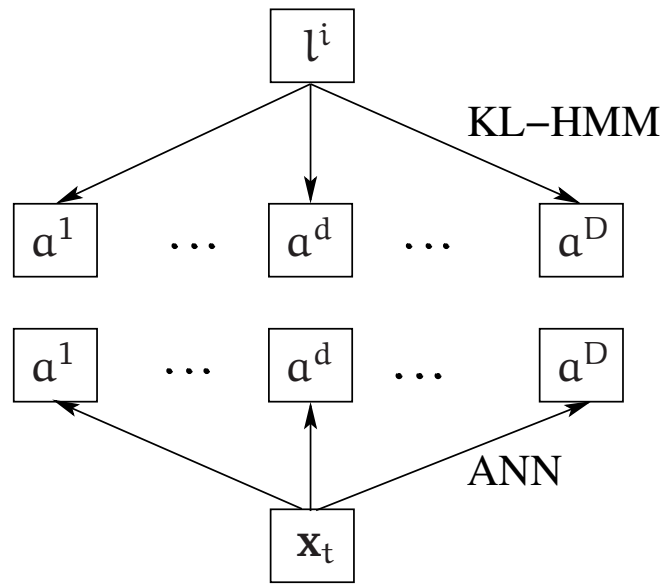
HMM/GMM



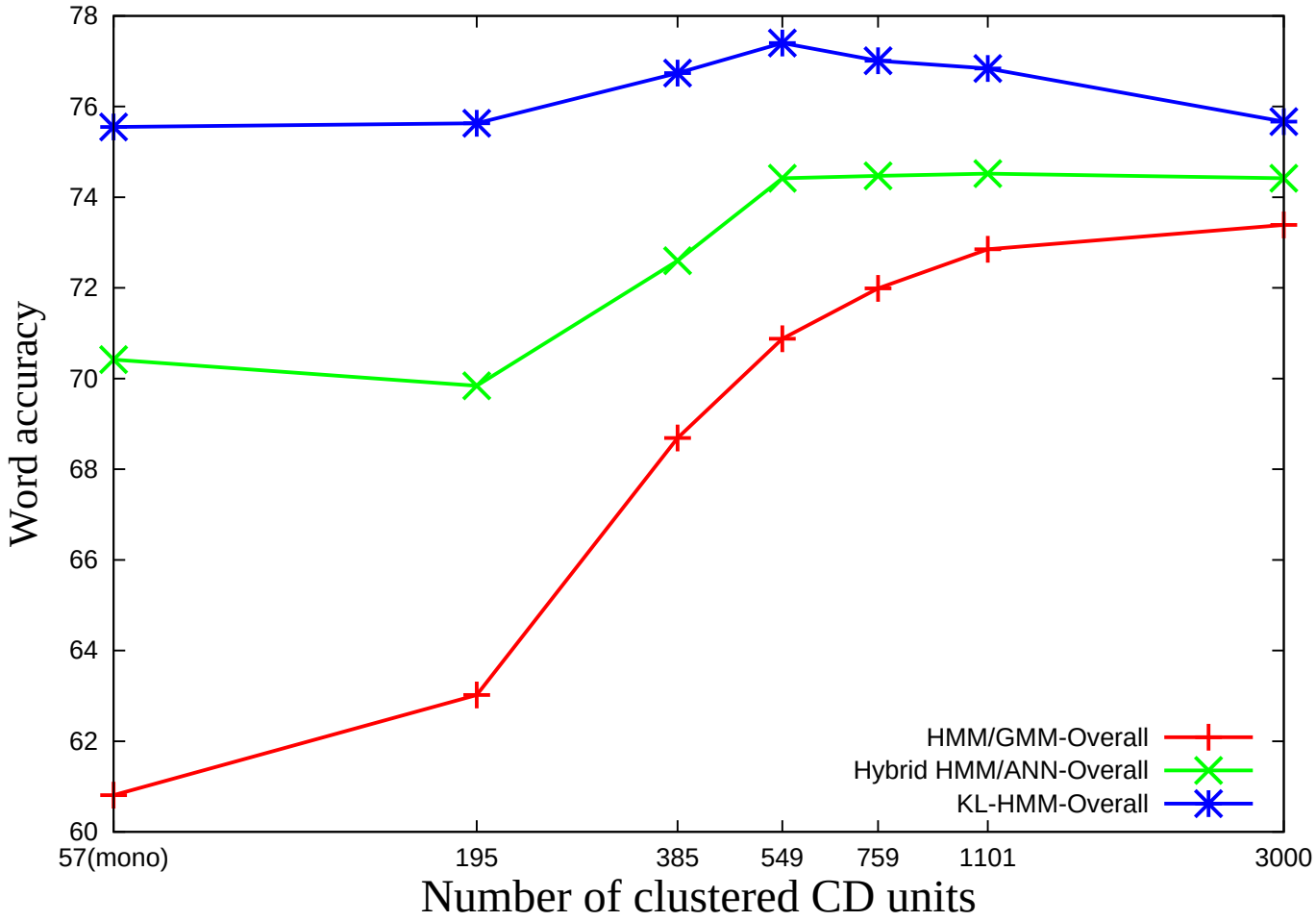
Hybrid HMM/ANN



KL-HMM



Results


78
76
74
72
70
68
66
64
62
60
Word accuracy
57(mono)
195
385
549
759
1101
3000
Number of clustered CD units
HMM/GMM-Overall
Hybrid HMM/ANN-Overall
KL-HMM-Overall

Conclusions

KL-HMM approach for speech recognition:

1. Efficient resource sharing
2. Suitable for both grapheme and phone based pronunciation lexicon
3. Suitable when task is challenged by both transcribed speech and pronunciation resource constraints
4. Performs better or comparable in well-resourced conditions

Thank you for your attention

Questions?