Speech Recognition

Lecture 10: Pronunciation Models.

Mehryar Mohri
Courant Institute of Mathematical Sciences
mohri@cims.nyu.edu

Speech Recognition Components

- Acoustic and pronunciation model:

$$\Pr(o \mid w) = \sum_{d,c,p} \Pr(o \mid d) \Pr(d \mid c) \Pr(c \mid p) \Pr(p \mid w).$$

acoustic model

- $\Pr(o \mid d)$: observation seq. $\leftarrow$ distribution seq.
- $\Pr(d \mid c)$: distribution seq. $\leftarrow$ CD phone seq.
- $\Pr(c \mid p)$: CD phone seq. $\leftarrow$ phoneme seq.
- $\Pr(p \mid w)$: phoneme seq. $\leftarrow$ word seq.

- Language model: $\Pr(w)$, distribution over word seq.

Terminology

- Phonemes: abstract units representing sounds in words or word sequences, e.g., /aa/, or /t/.
- Phones: acoustic realizations of phonemes, e.g., [t].
- Allophones: distinct realizations of the same phoneme, typically due to specific dialect, phonemic context, or speaking rate.
  - Example: [dx] and [t] can be realizations of /t/ in American English as in [s aa dx el] or [s aa t el].

Pronunciation Problems

- Problems: different sources of variability.
  - phonetic context: context-dependent models.
  - speaker variation (gender, dialect, accent, individual, emotion): speaker adaptation with about 10 s.
  - task or genre variation: domain adaptation.
  - effects of environment (source, channel, Lombard effect): cepstral mean normalization, microphone arrays.

Context-Dependent Phones

(Lee, 1990; Young et al., 1994)

- Idea:
  - phoneme pronunciation depends on environment (allophones, co-articulation).
  - model phone in context $\rightarrow$ better accuracy.
- Context-dependent rules:
  - Context-dependent units: $ae/b\_\_\_d \rightarrow ae_{b,d}$.
  - Allophonic rules: $t/V'\_\_\_V \rightarrow dx$.
  - Complex contexts: regular expressions.

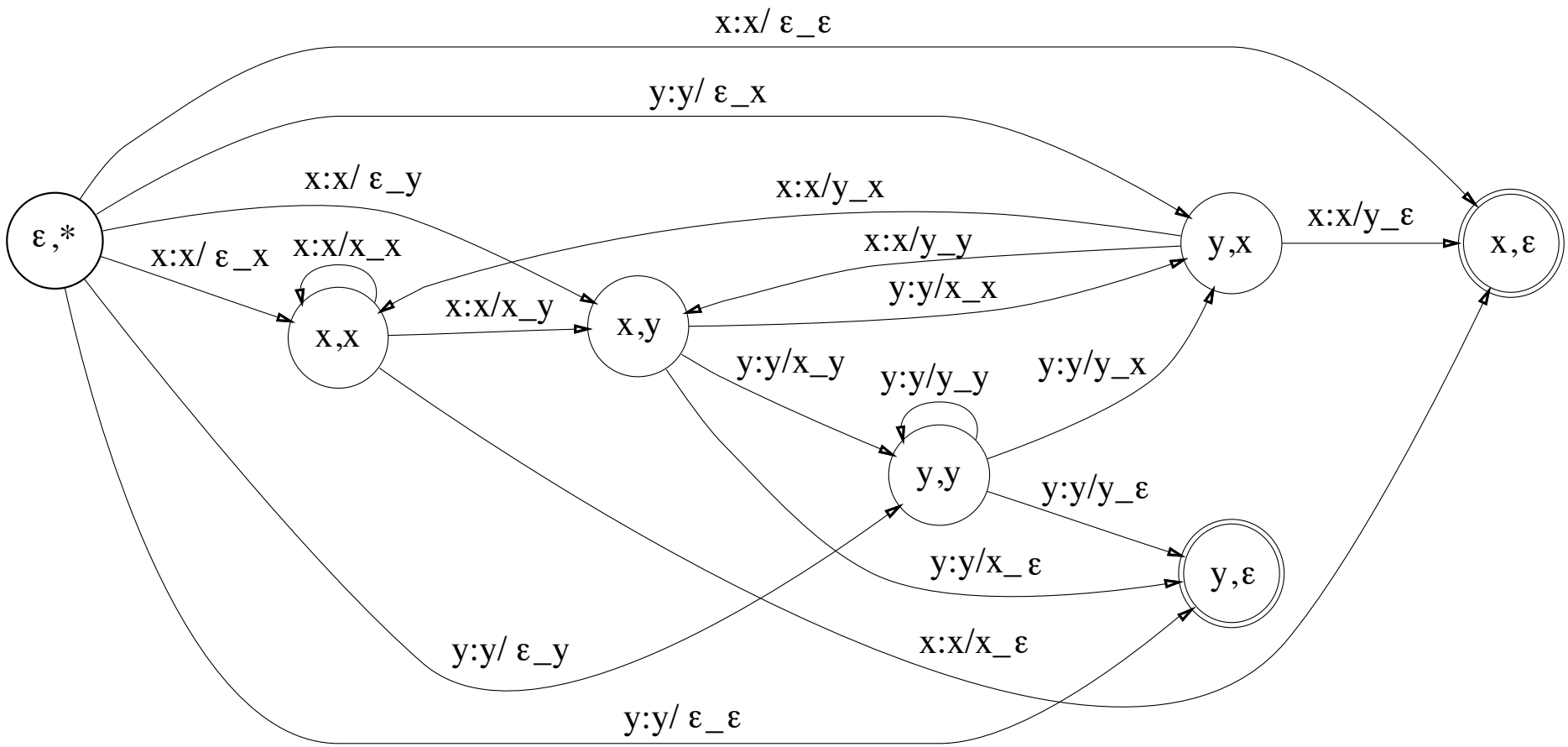
CD Phones - Speech Recognition

- Triphones: simplest and most widely used model.
  - context: window of length three.
  - example: *cat*, $pause\ _{pause}k_{ae}\ _{k}ae_{t}\ _{ae}t_{pause}\ pause$.
  - cross-word triphones: context spanning word boundaries, important for accurate modeling.
  - older systems: only word-internal triphones.
- Extensions: quinphones (window of size five), gender-dependent models.

CD Model Representation

(MM, Pereira, Riley, 2007)

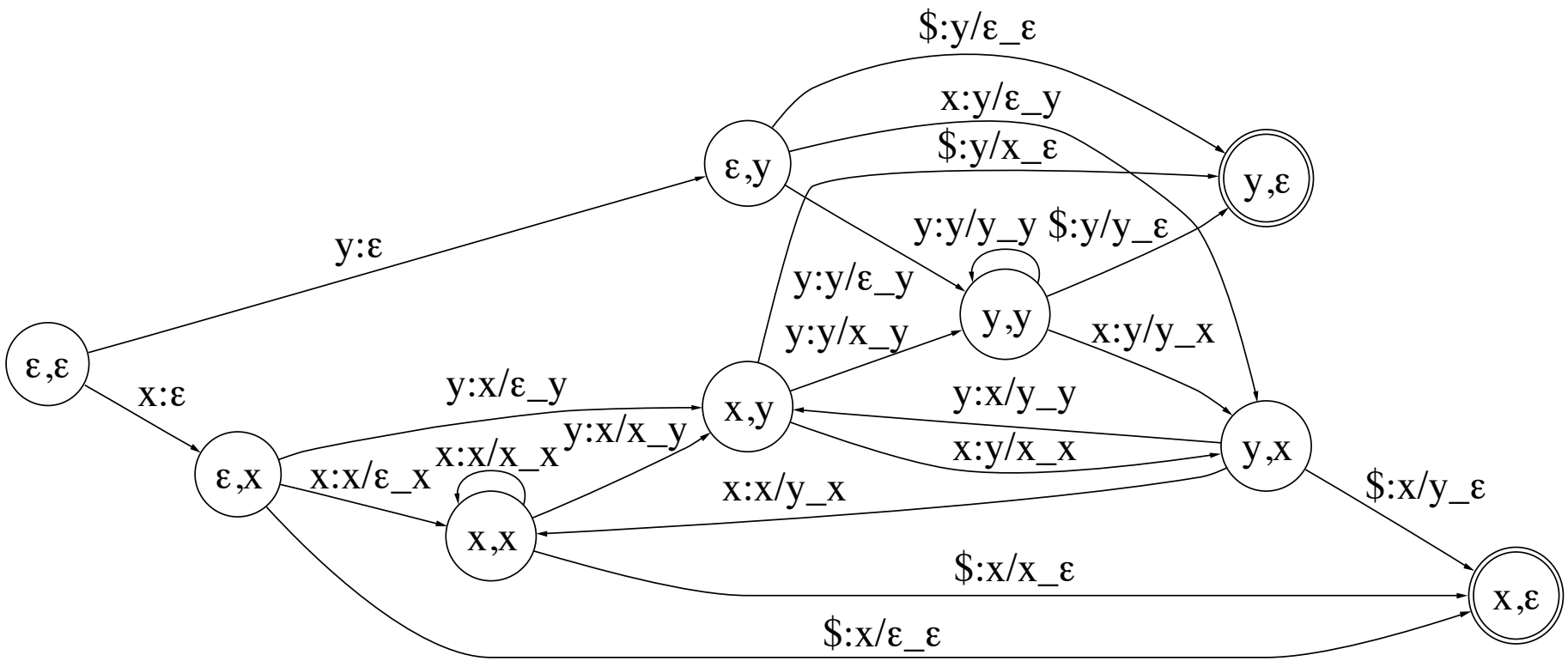
- Non-deterministic transducer representation



CD Model Representation

(MM, Pereira, Riley, 2007)

- Deterministic transducer representation



Modeling Problems

- Parameters: very large numbers for VLVR.
  - Number of phones: about 50 .
  - Number of CD phones: possibly $50^3 = 125{,}000$, but not all of them occur (phonotactic constraints). In practice, about $60{,}000$.
  - Number of HMM parameters: with 16 mixture components, $60000 \times 3 \times (39 \times 16 \times 2 + 16) \approx 228\text{M}.$
- Data sparsity: some triphones, particularly cross-word triphones, do not appear in sample.

Solutions

- Backing-off: use simpler models with shorter contexts when triphone not available.
- Interpolation: with simpler models such as monophone or biphone models.
- Parameter reduction: cluster parameters with similar characteristics ('parameter tying').
  - clustering HMM states.
  - better estimates for each state distribution.

  $\longrightarrow$ decision trees.

Clustering Method

- Initially, group together all triphones for the same phoneme.
- Split group according to decision tree questions based on left or right phonetic context.
- All triphones (HMM states) at the same leaf are clustered (tied).
- Advantage: even unseen triphones are assigned to a cluster and thus a model.
- Questions: which DT questions? Which criterion?

Questions

- Simple discrete pre-defined binary questions. Examples:
  - is the phoneme to the left an /l/?
  - is the phoneme to the right a nasal?
  - is the previous phoneme an unvoiced stop?

Phonetics

- **Vowels:**

Short vowels:	**pit**	**pet**	**pat**	**putt**	**pot**	**put**	**another**	
	ɪ	e	æ	ʌ	ɒ	ʊ	ə	ə
Long vowels:	**bean**	**barn**	**born**	**boon**	**burn**			
	iː	ɑː	ɔː	uː	ɜː			
Diphthongs:	**bay**	**buy**	**boy**	**no**	**now**	**peer**	**pair**	**poor**
	eɪ	ai	ɔɪ	əʊ	aʊ	ɪə	eə	ʊə

- **Consonants:**

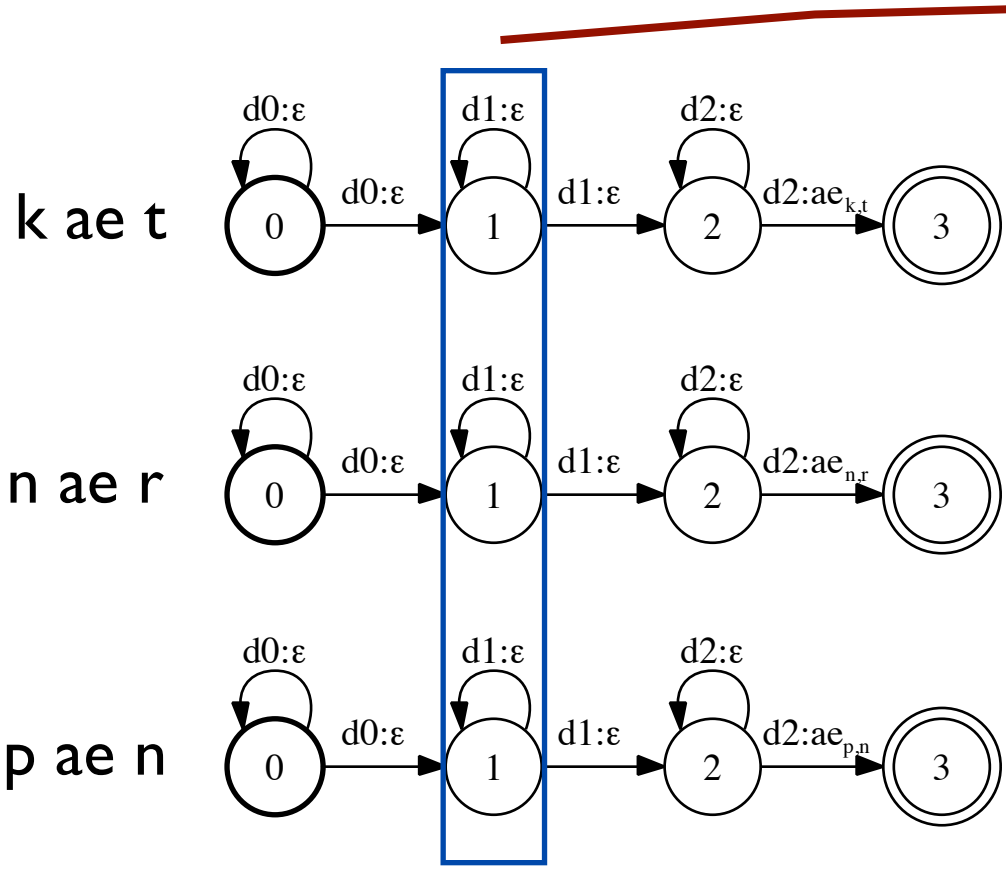
plosives/stops:	p	b	t	d	k	g	affricates:	ʧ	ʤ	
	pin	**b**in	**t**in	**d**in	**k**in	**g**um		**ch**ain	**J**ane	
fricatives:	f	v	θ	ð	s	z	ʃ	ʒ	h	
	fine	vine	**th**ink	**th**is	seal	zeal	**sh**eep	mea**s**ure	**h**ow	
nasals:	m	n	ŋ							
	mum	**n**un	su**ng**							
approximants/liquids:	l	r	w	j						
	light	**r**ight	**w**et	**y**et						

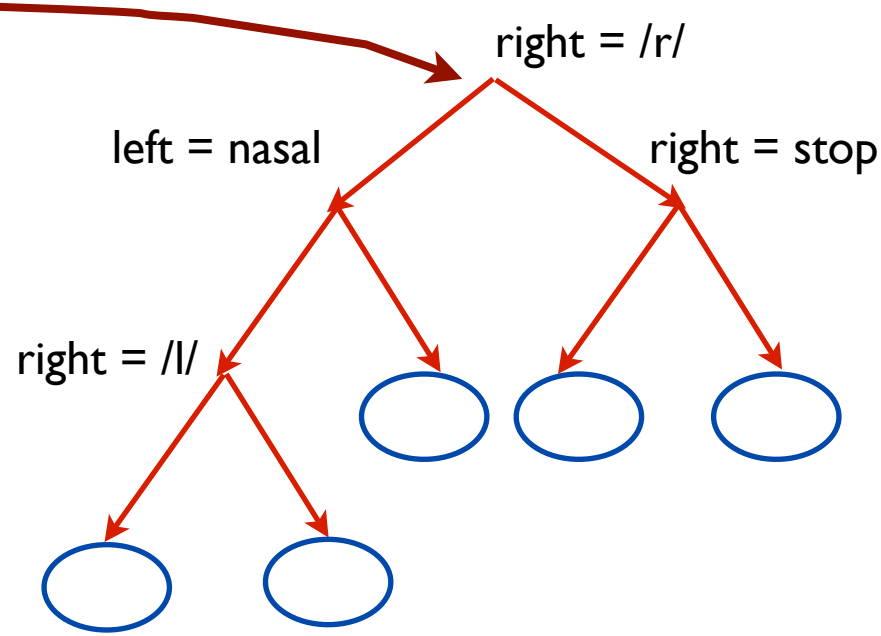
Sound Features

- Example: voiced sound (vocal cords vibrate), nasals (e.g., /m/, /n/), approximants (e.g., /l/, /r/, /w/, /j/), vowels.

Voiceless consonant (surd)	Voiced equivalent
[p] (***p**in*)	[b] (***b**in*)
[t] (***t**en*)	[d] (***d**en*)
[k] (***c**on*)	[g] (***g**one*)
[tʃ] (***ch**in*)	[dʒ] (***g**in*)
[f] (***f**an*)	[v](***v**an*)
[θ] (*(**th**in, **th**igh)*	[ð] *(**th**en, **th**y)*
[s] (***s**ip*)	[z] (***z**ip*)
[ʃ] *(pres**s**ure)*	[ʒ] *(plea**s**ure)*

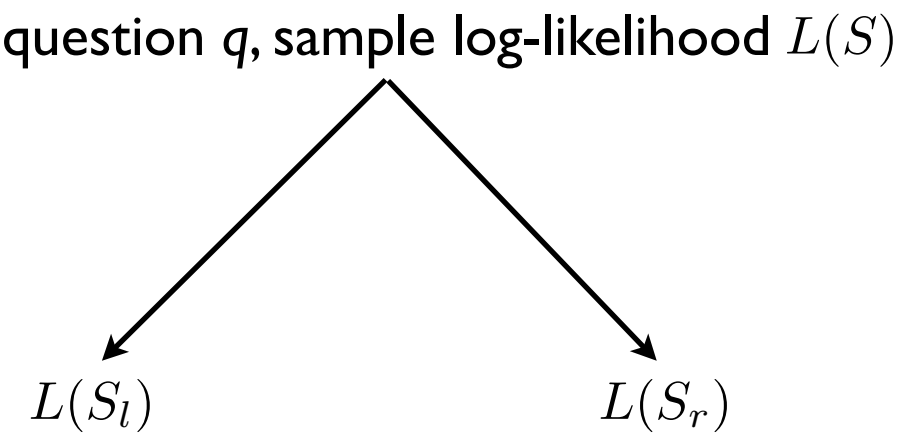
Clustering CD Phones





Criterion

- Criterion: best question is one that maximizes sample likelihood after splitting.
- ML evaluation: requires single Gaussians with diagonal covariance matrices trained on sample.



Log-likelihood difference: $\Delta L(q) = \big[L(S_l) + L(S_r)\big] - L(S).$

Best question: $q^* = \underset{q}{\operatorname{argmax}}\, L(S_l) + L(S_r).$

Log-Likelihood

- Sample $S = (x_1, \ldots, x_m) \in (\mathbb{R}^N)^m$.
- Diagonal covariance Gaussian:

$$\Pr[x] = \frac{1}{\prod_{k=1}^{N} (2\pi\sigma_k^2)^{1/2}} \prod_{k=1}^{N} \exp\Big(-\frac{1}{2}\frac{(x_k - \mu_k)^2}{\sigma_k^2}\Big).$$

- Log-likelihood for diagonal covariance Gaussian:

$$\begin{aligned} L(S) &= -\frac{1}{2}\sum_{i=1}^{m}\Big[\sum_{k=1}^{N}\log(2\pi\sigma_k^2) + \sum_{k=1}^{N}\frac{(x_{ik} - \mu_k)^2}{\sigma_k^2}\Big] \\ &= -\frac{1}{2}\Big[m\sum_{k=1}^{N}\log(2\pi\sigma_k^2) + m\sum_{k=1}^{N}\frac{\sigma_k^2}{\sigma_k^2}\Big] \\ &= -\frac{1}{2}\Big[mN(1 + \log(2\pi)) + m\sum_{k=1}^{N}\log(\sigma_k^2)\Big]. \end{aligned}$$

Decision Tree Split

- Log-likelihood difference:

$$L(S_l) + L(S_r) = -\frac{1}{2} mN(1 + \log(2\pi)) - \frac{1}{2}\left[m_l \sum_{k=1}^{N} \log(\sigma_{lk}^2) + m_r \sum_{k=1}^{N} \log(\sigma_{rk}^2)\right].$$

- Best question:

$$q^* = \operatorname*{argmin}_q \left[m_l \sum_{k=1}^{N} \log(\sigma_{lk}^2) + m_r \sum_{k=1}^{N} \log(\sigma_{rk}^2)\right],$$

with $$\sigma_{lk}^2 = \frac{1}{m_l} \sum_{x \in S_l} x_k^2 - \frac{1}{m_l^2} (\sum_{x \in S_l} x_k)^2$$

$$\sigma_{rk}^2 = \frac{1}{m_r} \sum_{x \in S_r} x_k^2 - \frac{1}{m_r^2} (\sum_{x \in S_r} x_k)^2.$$

Stopping Criteria

- Grow-then-prune strategy with cross-validation using held-out data set.
- Heuristics in VLVR:
  - question does not yield significant increase in log-likelihood.
  - insufficient data for questions.
  - computational limitations.

Full Training Process

- Train CI phone HMMs with single Gaussians and diagonal covariance.
- Create triphone HMMs by replicating CI phone models and reestimate parameters.
- Apply decision tree clustering to the set of triphones representing the same phoneme.
- Create Gaussian mixture model using mixture splitting technique for each cluster.

References

- Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. *Classifications and Regression Trees*. Chapman & Hall, 1984.
- Chen, F. Identification of contextual factors for pronunciation networks. In Proceedings of ICASSP (1990), S14.9.
- Luc Devroye, Laszlo Gyorfi, Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.
- Ladefoged, P., A Course in Phonetics., New York: Harcourt, Brace, and Jovanovich, 1982. Automatic Generation of Lexicons 17.
- Kai-Fu Lee. Context-Dependent Phonetic Hidden Markov Models for Continuous Speech Recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 38(4): 599-609, 1990.
- Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. Speech Recognition with Weighted Finite-State Transducers. In Larry Rabiner and Fred Juang, editors, Handbook on Speech Processing and Speech Communication, Part E: Speech recognition. volume to appear. Springer-Verlag, Heidelberg, Germany, 2007.

References

- Quinlan, J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.
- Quinlan, J. R. Induction of Decision Trees, in *Machine Learning, Volume 1*, pages 81-106, 1986.
- Randolph, M. "A data-driven method for discover and predicting allophonic variation," Proc. ICASSP `90, S14.10, 1990.
- Michael Riley and Andrej Ljolje. Lexical access with a statistically-derived phonetic network. In Proceedings of the European Conference on Speech Communication and Technology, pages 585-588, 1991.
- M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, and D. Bell, "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," Proc. ICASSP '89, pp. 699--702, Glasgow, Scotland, May, 1989.
- Steve Young, J. Odell, and Phil Woodland. Tree-Based State-Tying for High Accuracy Acoustic Modelling. In *Proceedings of ARPA Human Language Technology Workshop*, Morgan Kaufmann, San Francisco, 1994.