Speaker-Independent Phone Recognition Using Hidden Markov Models

Author: KAI-FU LEE and HSIAO-WUEN HON

Presenter: Xiaoqing Zhao
Media Informatics
RWTH Aahcen University

February 25, 2015

Outline

1. Introduction
2. The TIMIT Database
3. Speech Processing
4. Context-Independent HMM Training
5. Context-Dependent HMM Training
6. HMM Recognition
7. Experiment Results
8. Conclusion

Introduction

Hidden Markov models(HMM) are successfully applied to various constrained tasks, such as speaker-dependent recognition of isolated words, continuous speech, and phones. So we intend to apply HMM's to speaker-independent phone recognition in continuous speech.

With context-independent phone models, we attained a recognition rate of 64.07%, and with right-context-dependent phone models, the recognition rate improved to 73.80%. The results are also compared well with other approaches to speaker-independent phone recognition.

The TIMIT Database

630 speakers, 10 sentences said each.

- 2 "sa" sentences, which are the same across all speakers.
- 5 "sx" sentences, which were read from a list of 450 phonetically balanced sentences selected by MIT.
- 3 "si" sentences, which were randomly selected by TI.

We actually used 2830 sentences by 357 speakers for training data, and 160 sentences by 20 speakers for test data. After folded some phones, 39 effective phones left.

The TIMIT Database

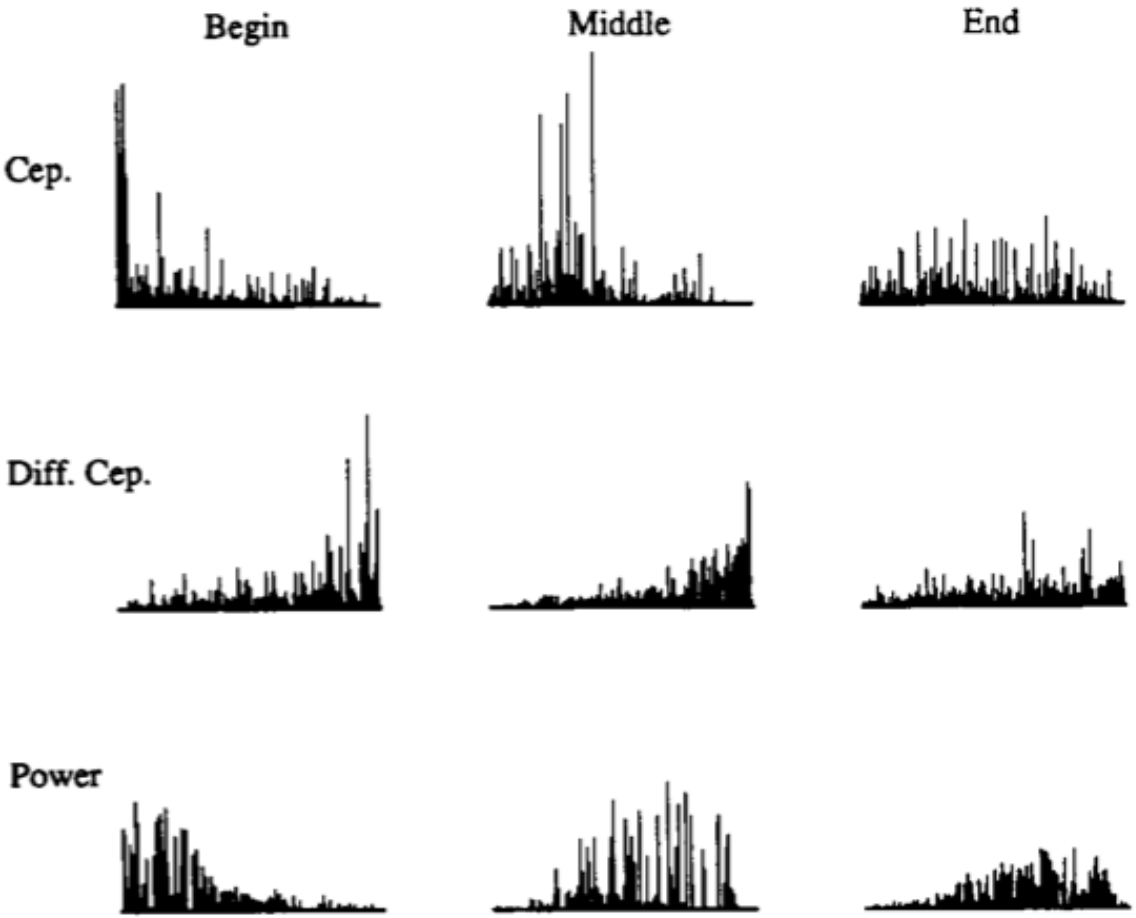
LIST OF THE PHONES USED IN OUR PHONE RECOGNITION TASK

Phone	Example	Folded	Phone	Example	Folded
iy	*beat*		en	*button*	
ih	*bit*		ng	*sing*	eng
eh	*bet*		ch	*church*	
ae	*bat*		jh	*judge*	
ix	*roses*		dh	*they*	
ax	*the*		b	*bob*	
ah	*butt*		d	*dad*	
uw	*boot*	ux	dx	(butter)	
uh	*book*		g	*gag*	
ao	*about*		p	*pop*	
aa	*cot*		t	*tot*	
ey	*bait*		k	*kick*	
ay	*bite*		z	*zoo*	
oy	*boy*		zh	*measure*	
aw	*bough*		v	*very*	
ow	*boat*		f	*fief*	
l	*led*		th	*thief*	
el	*bottle*		s	*sis*	
r	*red*		sh	*shoe*	
y	*yet*		hh	*hay*	hv
w	*wet*		cl (sil)	(unvoiced closure)	pcl,tcl,kcl,qcl
er	*bird*	axr	vcl (sil)	(voiced closure)	bcl,dcl,gcl
m	*mom*	em	epi (sil)	(epinthetic closure)	
n	*non*	nx	sil	(silence)	h#,#h,pau

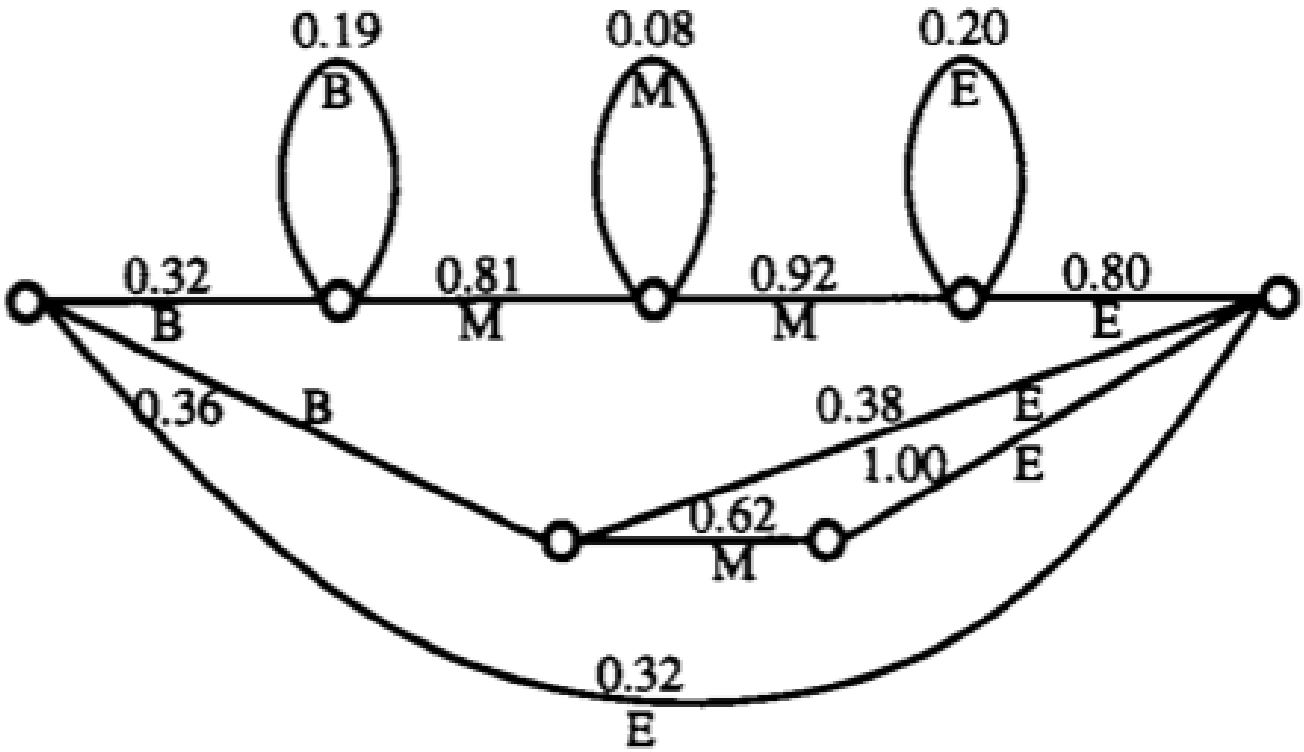
Speech Processing

- A set of 12 LPC-derived cepstral coefficients are computed from the LPC coefficients. 12 coefficients are vector quantized into a codebook of 256 prototype vectors of LPC cepstral coefficients.
- 13th codebook is vector quantized from differential coefficients.
- The last codebook is vector quantized from energy and differential energy values.

Context-Independent HMM Training



Context-Independent HMM Training



Co-occurrence Probability

The output parameters are smoothed using a novel smoothing technique: co-occurrence smoothing.

$$CP(i|j) = \frac{\sum_{p=1}^{NP} \sum_{d=1}^{ND(p)} P(i|p,d) \cdot P(j|p,d) \cdot P(p) \cdot P(d)}{\sum_{k=1}^{NC} \sum_{p=1}^{NP} \sum_{d=1}^{ND(p)} P(k|p,d) \cdot P(j|p,d) \cdot P(p) \cdot P(d)}$$

When codeword j is observed, how often is codeword i observed in similar contexts.

Co-occurrence Probability

Use the co-occurrence probability(CP) to smooth the output pdf's(P) into a smoothed pdf(SP):

$$SP(k|p,d) = \sum_{i=1}^{NC} CP(k|i) \cdot P(i|p,d)$$

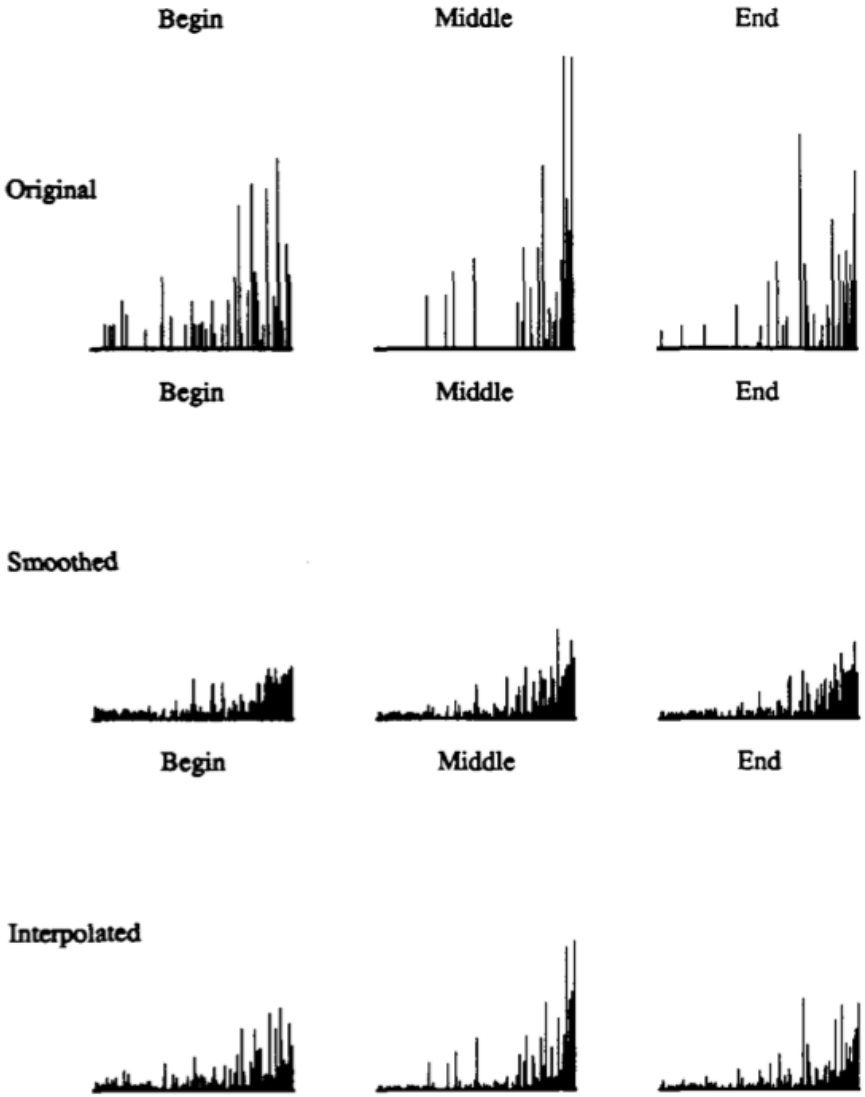
Co-occurrence Probability

In case of the result is over-smoothed, a compromise can be reached by combining the two pdf's:

$$MP(k|p,d) = \lambda_c \cdot P(k|p,d) + (1-\lambda_c) \cdot SP(k|p,d).$$

Co-occurrence Probability

Poorly trained pdf after smoothed by co-occurrence:



Co-occurrence Probability

Advantages:

- Working on continuous speech and does not require segmentation.
- Text-independent.
- Operating directly on the output pdf's.
- Less training data needed.
- Combining the two matrices gives superior results than using either one alone.

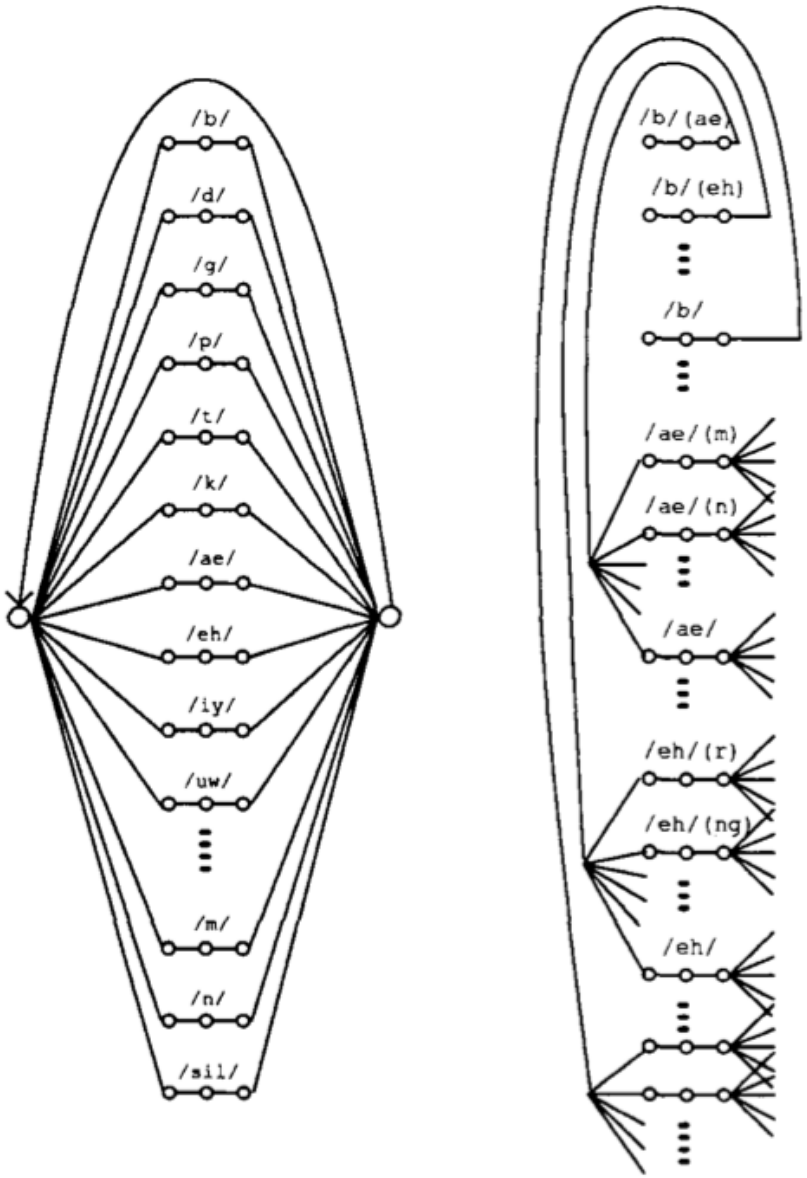
Context-Dependent HMM Training

Right-context dependent phone modeling:

before: /sil hh ix dx en sil m ...../

after: /sil(hh) hh(ix) ix(dx) dx(en) en(sil) sil(m) ...../

HMM Recognition



Phone Recognition

	Context-Indep.	Context-Dep.
Correct	64.07% (3883)	73.80% (4473)
Substitutions	26.22% (1589)	19.62% (1189)
Deletions	9.72% (589)	6.58% (399)
Insertions	10.79% (654)	7.72% (468)

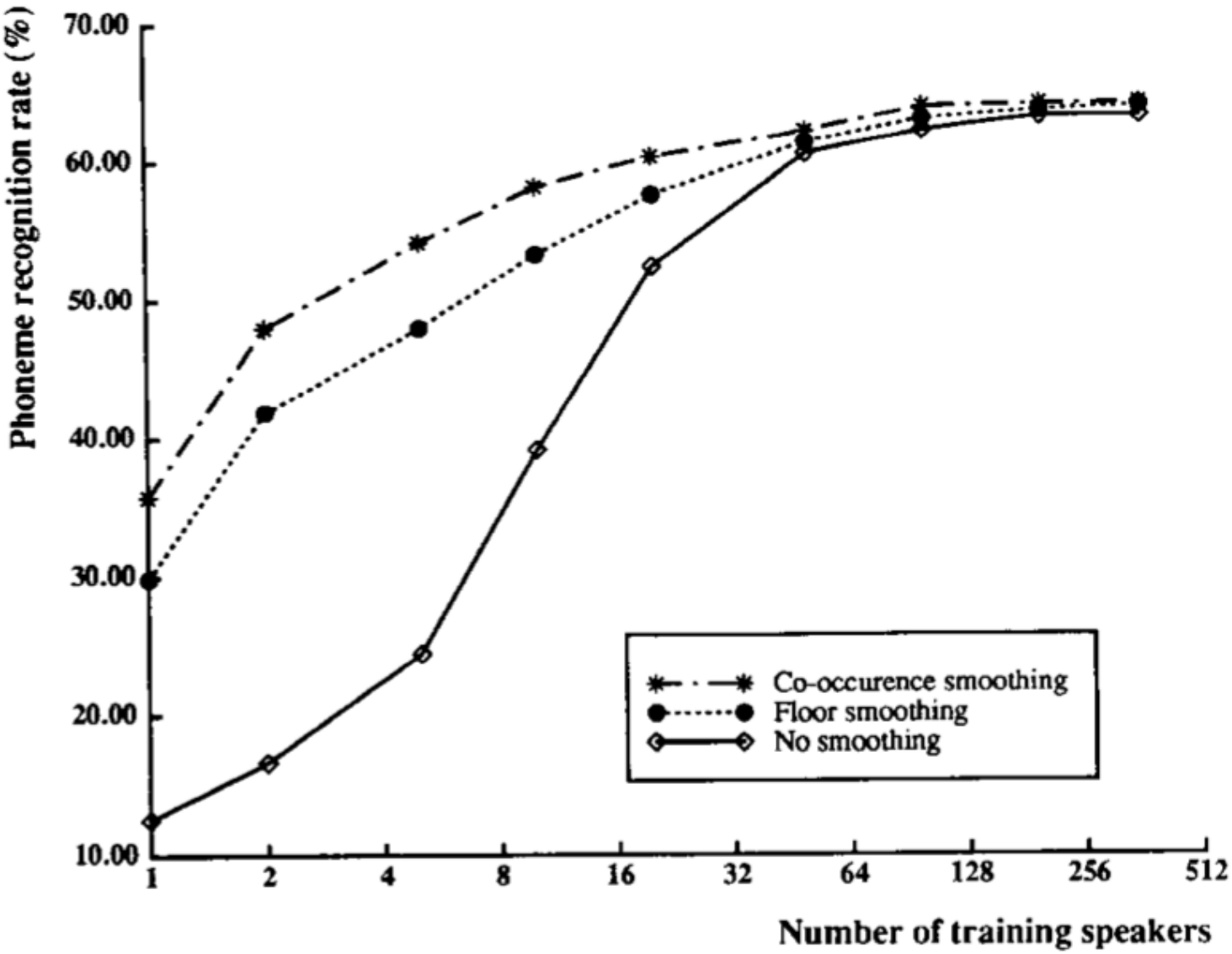
Phone Recognition

Class	Occurrences	Context-Indep.	Context-Dep.
Sonorant	3027	53.68% (1625)	65.71% (1989)
Stop	1014	58.09% (589)	69.92% (709)
Fricative	736	66.03% (486)	78.40% (577)
Closure	1284	92.13% (1183)	93.30% (1198)

Phone Language Model

Language Model	Context-Independent Recognition Rate	Context-Dependent Recognition Rate
Bigram	64.07%	73.80%
Unigram	60.91%	70.38%
None	58.77%	69.51%

Co-occurrence Smoothing



Conclusion

- Speaker-independent phone recognition.
- 73.80% speaker-independent phone recognition accuracy in continuous speech.
- Using TIMIT database fro evaluating.