Effects of Transcription Errors on Supervised Learning in Speech Recognition

May 23rd, 2003

Ram Sundaram

Candidate for Master of Science in Electrical Engineering
Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University

Organization of Presentation

- * Motivation for transcription error analysis
- * Acoustic model training
- * Databases and types of transcription errors
- * Experimental setup
- * Results on corrupted databases
- * Simulated experiments
- * Experimental analysis
- * Conclusions and future work

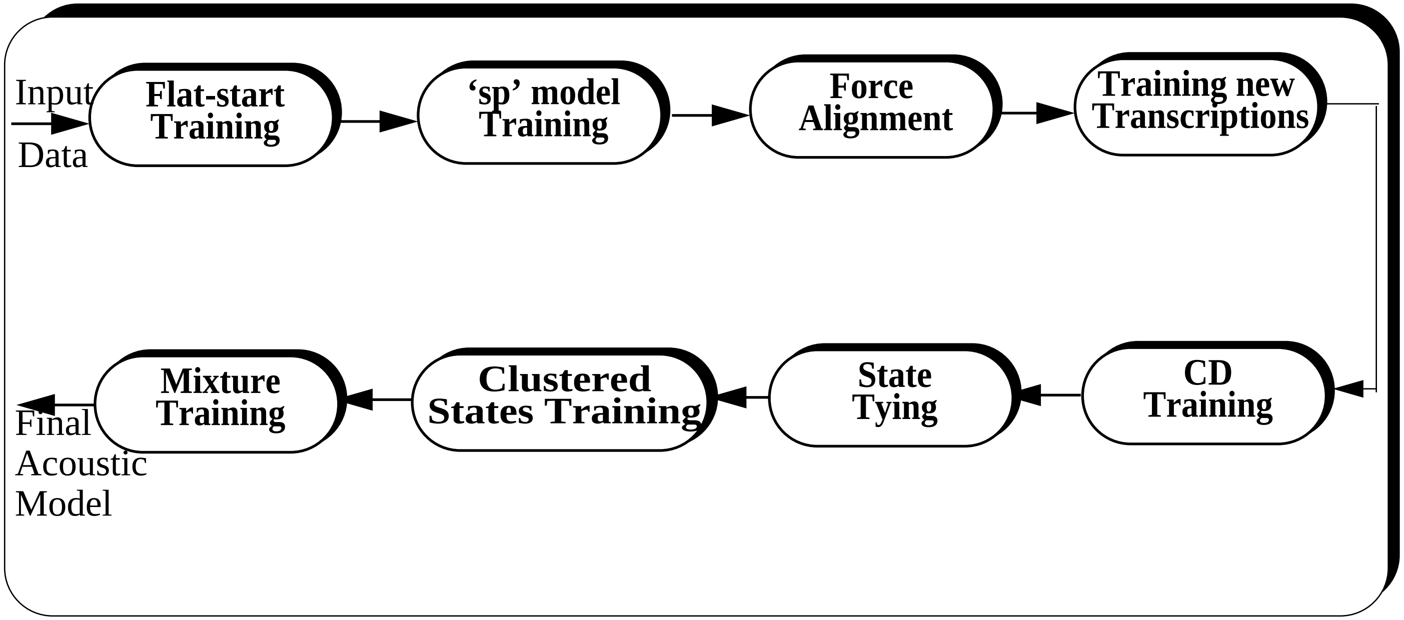
Motivation

- Cleaner training data yielded no measurable performance improvement on conversational speech tasks

- Readily available transcriptions do not degrade performance significantly

- Need to understand the robustness of the training process

Acoustic Model Training



* Define a phone set and create a pronunciation lexicon

* Define the HMM topology (typically 3 state HMMs)

* Gaussian distributions are used as the underlying distribution

* Multiple iterations for each stage in the training process

Updating Parameters

- * Mean update

$$\hat{\mu}_{jm} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} L_{jm}^{r}(t) o_t^r}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} L_{jm}^{r}(t)}$$

- * $L_{jm}^{r}(t)$ is the state occupancy probability for the mth mixture in the jth state in the rth utterance at time t

- * The state occupancy value can also be defined as the probability of the input data belonging to the model given the current model parameters

Types of Transcriptions

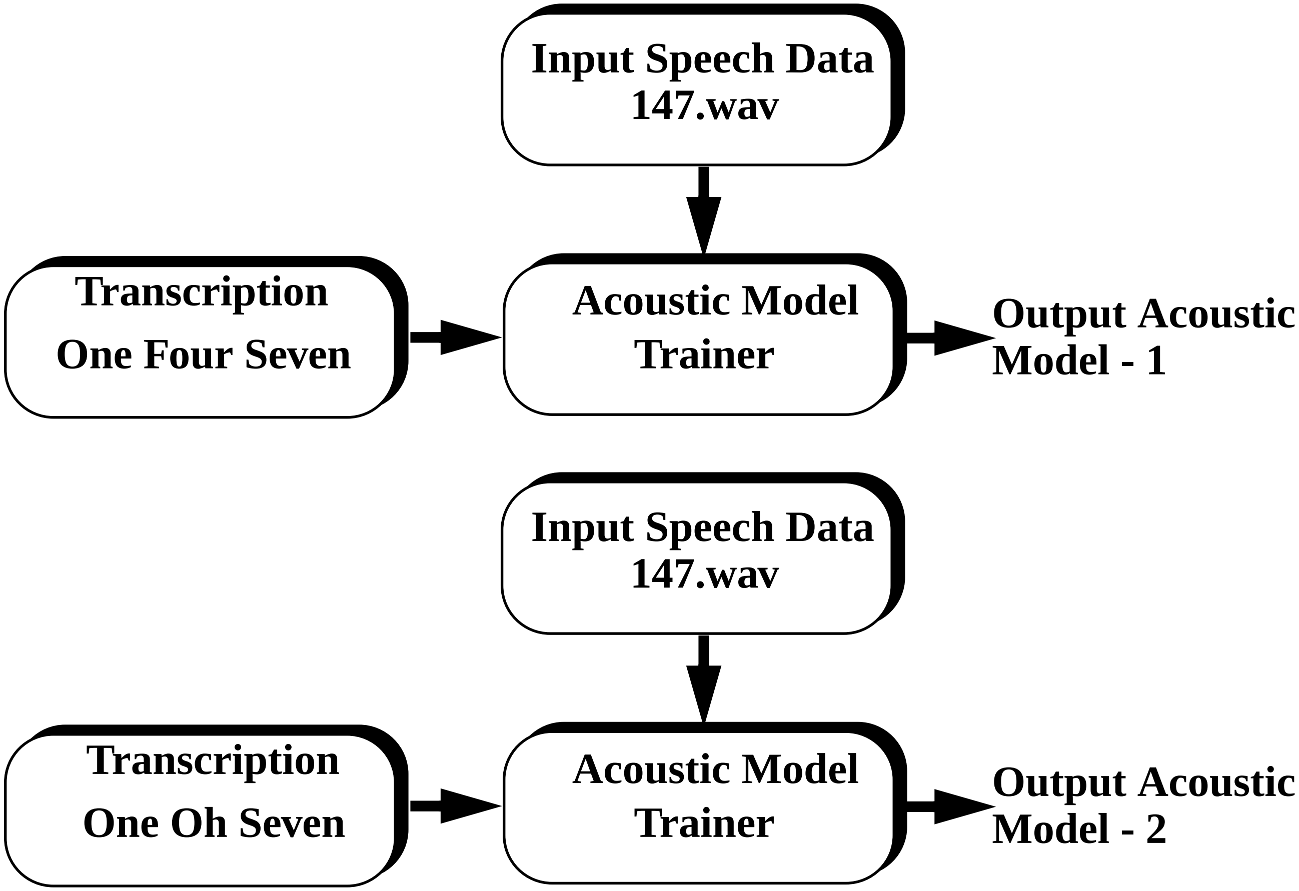
words	**how did you**
monophone	**hh aw d ih d y uw**
word-internal	**hh+aw hh-aw d+ih d-ih+d ih-d y+uw y-uw**
cross-word	**hh+aw hh-aw+d aw-d+ih d-ih+d ih-d+y d-y+uw y-uw**

* 35 to 45 phones used in US English phone set

* State clustering is performed to avoid sparse data problems

* Types of models dependent on the application

Thesis Focus



* How does the transcription error affect acoustic model generation?
* Does it have a significant effect on recognition accuracy?

Experimental Setup - (I)

There’s	not a whole lot of	**fabric**	variety		there
He	not a whole lot of		variety	**in**	there
Subs		*Dels*		*Ins*	

* Ways to introduce errors:

— Introducing errors from a validator's point of view

— Random introduction of errors

* Errors randomly distributed across the database

* Corrupt the transcriptions for existing databases namely TIDigits, Alphadigits and Switchboard

Experimental Setup - (II)

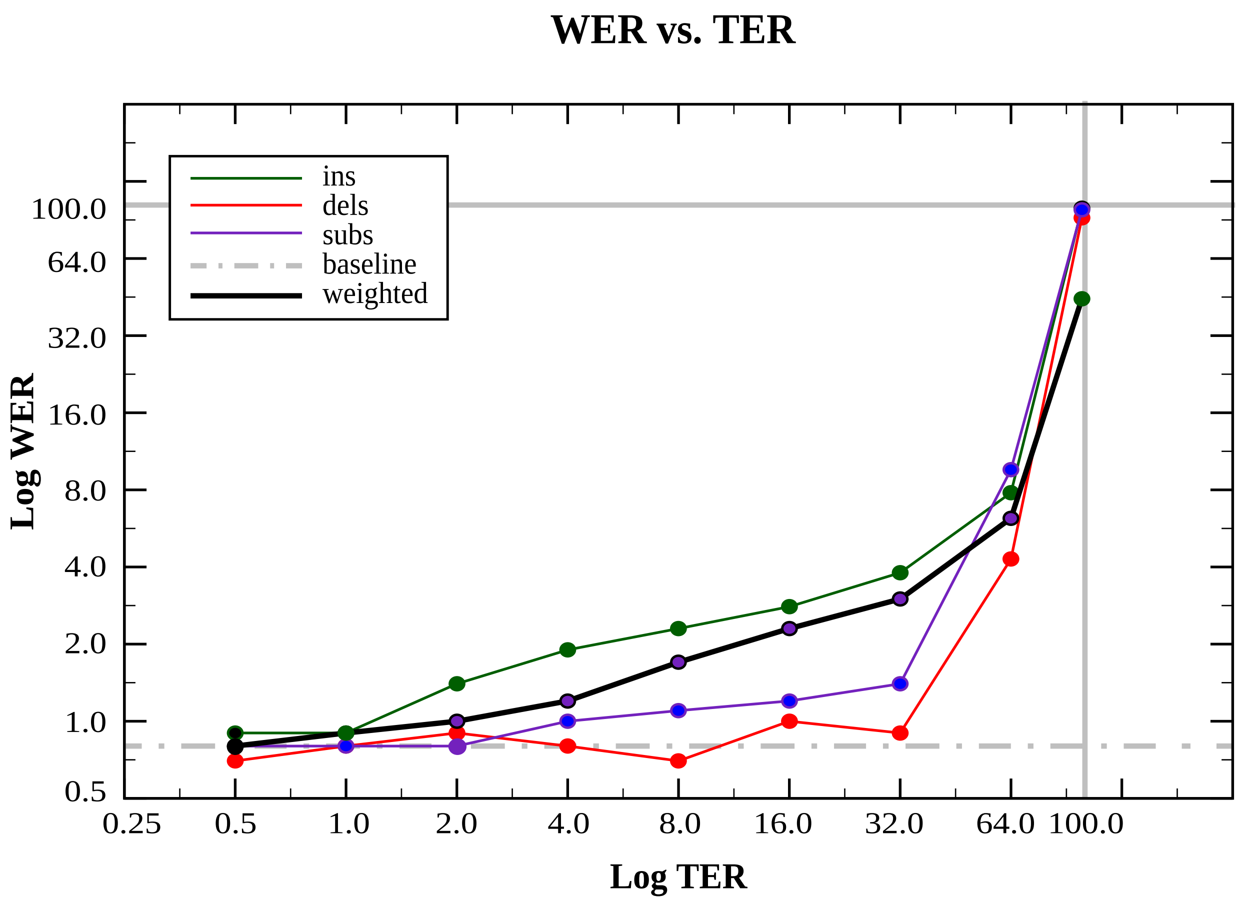
- * TIDigits: small digits-only vocabulary with over 300 speakers and over 12000 training and test utterances

- * Alphadigits: vocabulary includes alphabets and digits with several male and female speakers and over 50 hours of training and 3 hours of testing data

- * Switchboard: a large vocabulary (over 100,000 words) task involving telephone recordings of conversations involving several speakers, 2438 conversations used for training and 30 minute test data

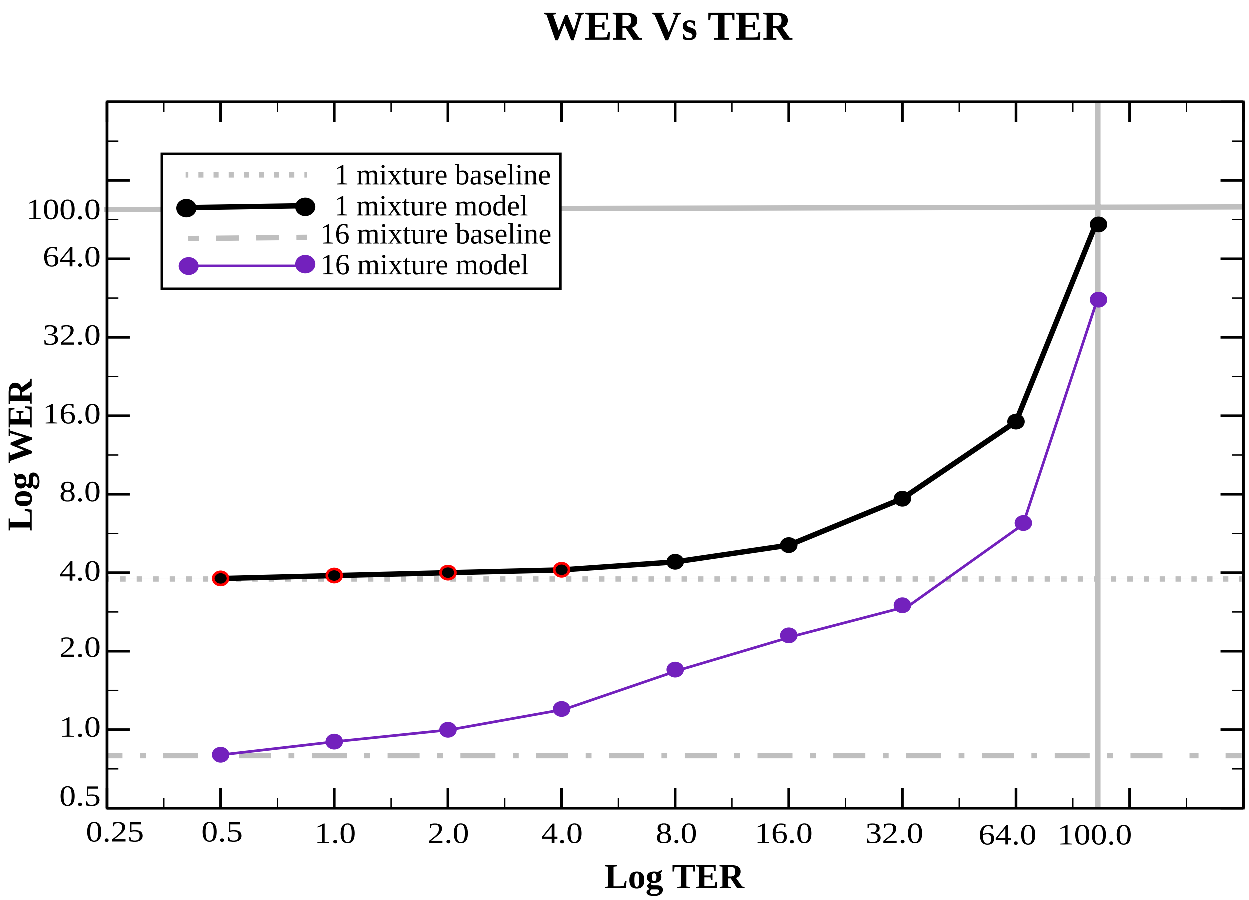
- * Substitutions for SWB randomly chosen; Substitutions for TIDIGITS/Alphadigits uniformly chosen across all words

TIDigits Results (I)



- * TIDigits training performed using word models
- * No significant degradation in word error rate (WER) until 16% transcription error rate (TER) for any type of transcription error

TIDigits Results (II)



- * Monophone and mixture models show the same trend
- * Small, gradual degradation in WER with increase in TER

Comparative Results

Corpora	Acoustic Models	Transcription Error Rate WER		
		0%	2%	16%
TIDIGITS	1 mixture word	3.8	4.0 (+5.2)	5.1 (+34.2)
	16 mixture word	0.8	1.0 (+25.3)	2.3 (+187.5)
Alphadigits	1 mixture xwrd	31.9	32.3 (+1.2)	36.2 (+13.4)
	16 mixture xwrd	10.8	10.8 (+0.0)	12.1 (+12.0)
SWB	12 mixture xwrd	41.1	41.8 (+1.7)	44.6 (+8.5)

* Cross-word models used in training Alphadigits and Switchboard

* No significant change in WER for low TER

* Alphadigits: Phonetic mixture models are more robust to transcription errors by 11% relative

Simulated Experiments

* Need for simulated experiments

— Robustness to transcription errors cannot be attributed to a single phenomenon

— High dimensionality makes the computations intractable

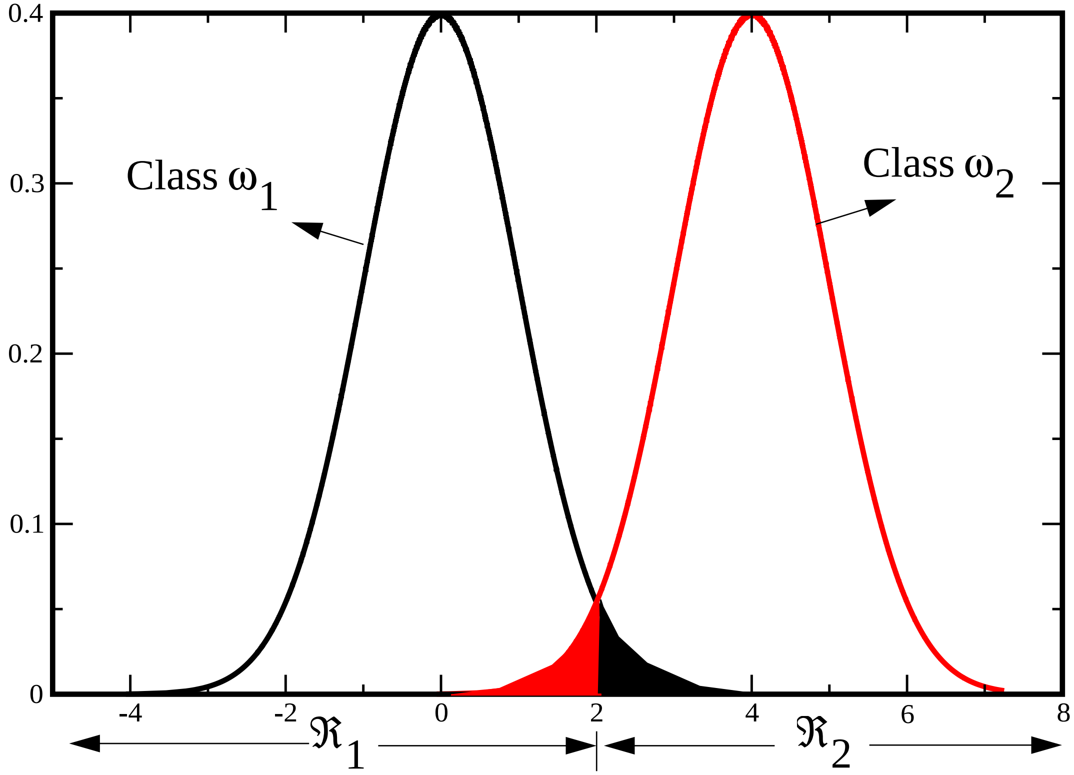
— A simpler setup using a two-model scenario

* Quantify the effect of erroneous data on Gaussian distributions

— Kullbeck-Leibler distance

— Probability of Error

Experimental Design



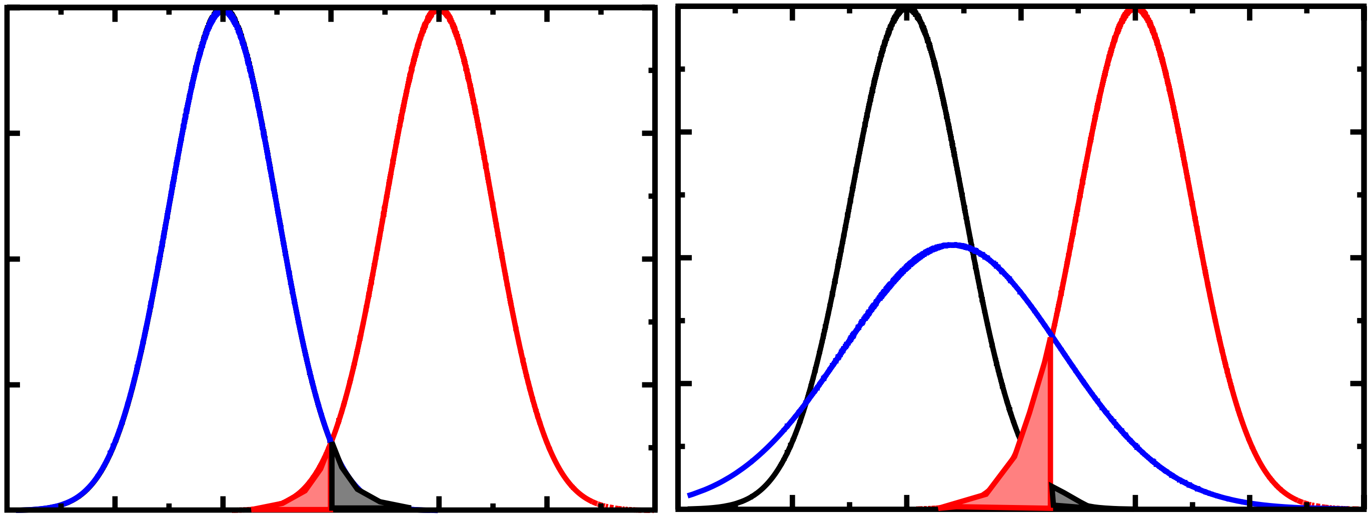
- Probability of error given by:

$$P(e) = P(x \in \Re_2, \omega_1) + P(x \in \Re_1, \omega_2)$$

- Corrupt one distribution in a controlled manner

- Estimate the parameters of the distribution

Simulated Experiments - Results (I)



- Original distribution (black), corrupting distribution (red), new estimated distribution (blue)
- Probability of error calculated for zero and twenty percent corrupted data
- Probability of error increases but not significantly

Simulated Experiments - Results (II)

Data Error Rate	Probability of Error	
	‘b’ - ‘d’ (acoustically similar pair)	‘aa’ - ‘s’ (acoustically dissimilar pair)
0	44.1	6.84
2	44.1	6.89
4	44.1	7.01
6	44.1	7.12
8	44.1	7.25
10	44.1	7.37
12	44.1	7.49
14	44.1	7.60
16	44.1	7.70
18	44.1	7.79
20	44.1	7.87

* first feature of the phones were chosen from Alphadigits

* 'aa'-'s' pair:

  Mean: [4.038, -5.717]
  Variance: [9.381,12.259]

* 'b' - 'd' pair:

  Mean: [0.704,-0.461]
  Variance:[21.119,16.406]

Analysis — Setup

* Need to understand the robustness of the training process at a fundamental level

* Experimental Setup

— 4884 utterances from Alphadigits were used

— 100 utterances with the word “o” were chosen

— “o” was replaced with “i” in these 100 utterances

— The 100 utterances without transcription errors were added

— The subset now has 4984 utterances with 7.8% transcription error rate

Analysis - Hypotheses

- How much does an incorrect model learn from the erroneous data?
  - Analyzed by observing the state occupancy values of the incorrect model (model ‘ay’ that occurs in the utterance with transcription error) and comparing it with the state occupancy values of the correct model (equivalent model ‘ow’ that occurs in the same utterance but with no transcription errors).
- How much does the erroneous portion of the data contribute to the model reestimation process?
  - Analyzed by observing the state occupancies of the incorrect model (model ‘ay’) in the utterances with transcription errors and comparing it with the state occupancies of the same model in other utterances without transcription errors.

Monophone Training

Iteration	Center State of 'ow'	Center State of 'ay'
1	0.037	0.037
2	0.122	0.057
3	0.355	0.078
4	0.590	0.150
5	0.633	0.150
6	0.634	0.173
7	0.641	0.159
8	0.639	0.153
9	0.660	0.143
10	0.655	0.153
11	0.659	0.155
12	0.660	0.151

- State occupancy values expected to be low based on previous results on databases
- State occupancy values were observed for the incorrect model 'ay' and correct model 'ow'
- Incorrect model has low state occupancy value and learns little from the erroneous data

Monophone Training (II)

* How much does the erroneous portion of the data contribute to the model reestimation process?

— State occupancy values for 275 correct utterances for the model ‘ay’ was observed to be 0.53

— State occupancy values for 100 incorrect utterances for the model ‘ay’ was observed to be 0.15

* Erroneous data does not contribute significantly to the reestimation process

* Model ‘ay’ is left largely uncorrupted

Context-Dependent Training - (I)

Iteration	Average State Occupancy for Correct Transcriptions	Average State Occupancy for Incorrect Transcriptions
1	0.5223	0.0794
2	0.5808	0.0871
3	0.5827	0.1201
4	0.5772	0.1461

- Each context-dependent model gets less data
- Likely that the erroneous data may have more impact than in monophone training
- Models sil-ay+ey and f-ay+eh were observed
- Each model had a transcription error rate of 16% and 66% respectively
- State occupancy values are low for the ‘sil-ay+ey’ model in incorrect utterances but seem to increase after each iteration for the incorrect model

Context-Dependent Training - (II)

Iterations	Average State Occupancy for Correct Transcription	Average State Occupancy for Incorrect Transcription
1	0.5829	0.1490
2	0.5807	0.0851
3	0.5913	0.0873
4	0.5915	0.0873
5	0.5910	0.0876

- State clustering is performed to share data
- Percentage of transcription error likely to change based on how the states are shared
- TER for sil-ay+ey decreases to 0.05%
- State occupancy value decreases for the model in incorrect transcriptions

Context-dependent Training - (III)

Iterations	Average State Occupancy in Incorrect Transcriptions
1	0.3246
2	0.2020
3	0.2059
4	0.1726
5	0.1621

- CD model ‘f-ay+eh’ had 66% TER prior to state-tying
- Average state occupancy value is 0.56
- After state-tying, the TER decreases significantly

- State occupancy drops from 0.56 to 0.16 after state-tying
- State-tying helps in decreasing the TER and increasing the state occupancy values for the models in correct transcriptions

Mixture Training

Training Stage	State occupancy in correct transcriptions	State Occupancy in incorrect transcriptions
After 1mixture	**0.5372**	**0.1488**
After 2mixture	**0.5384**	**0.1404**
After 4 mixture	**0.5644**	**0.1282**

* Mixture training is performed to model the variations in the data

* State occupancy values can increase if erroneous data is modeled by a mixture

* State occupancy are low for the incorrect transcriptions and decreases as number of mixtures are increased

* Mixtures model other variations in the correct portion of the data and seem to ignore the erroneous data further

Conclusions

- Transcription errors do not corrupt the acoustic models significantly
- Alphadigits - at 16% TER, WER degrades only by 12%
- SWB - at 16% TER, WER degrades only by 8.5%
- Robustness to erroneous data mainly due to Gaussian distribution
- State-tying helps in decreasing the TER during the context-dependent modeling stage
- Mixture training adds more robustness by modeling other variations in the correct portion of the data

Future Work

* Best performance is obtained by using a clean set of data

— Need to analyze how much more erroneous data is required to match the performance of clean data

* Gaussian distribution adds significant robustness to the training process

— What happens if other distributions (e.g., Laplacian) are used to model the data?

Achievements

Accomplishments:

- Led ISIP's 2000 and 2001 Hub 5E evaluation efforts
- Led ISIP's 2000 SPeech In Noisy Environments (SPINE) evaluation efforts
- Developed HMM training for the ISIP prototype system
- Developed front end normalization algorithms

Publications:

- R. Sundaram and J. Picone, "The Effects of Transcription Errors," Proceedings of the Speech Transcription Workshop, Linthicum Heights, Maryland, USA, May 2001.
- R. Sundaram, J. Hamaker, and J. Picone, "TWISTER: The ISIP 2001 Conversational Speech Evaluation System," Proceedings of the Speech Transcription Workshop, Linthicum Heights, Maryland, USA, May 2001.
- B. George, B. Necioglu, J. Picone, G. Shuttic, and R. Sundaram, "The 2000 NRL Evaluation for Recognition of Speech in Noisy Environments," presented at the Speech In Noisy Environments (SPINE) Workshop, Naval Research Laboratory, Alexandria, Virginia, USA, October, 2000.
- R. Sundaram, A. Ganapathiraju, J. Hamaker and J. Picone, "ISIP 2000 Conversational Speech Evaluation System," Speech TranscriptionWorkshop, College Park, Maryland, USA, May 2000.

PROGRAM OF STUDY

Course No.	Title	Semester
EE 6773	Digital Signal Processing	Fall 1999
EE 7000	Advanced Topics in Speech Recognition (DIV)	Fall 1999
EE 8990	Fundamentals of Speech Recognition (Special Topics)	Spring 2000
EE 6713	Computer Architecture	Spring 2000
ST 8114	Statistical Methods	Spring 2000
CS 8633	Natural Language Processing	Fall 2000
ST 8253	Multivariate Methods	Fall 2000
EE 8990	Pattern Recognition	Spring 2001
ECE 8000	Research/Thesis	

Acknowledgements

I would like to thank Dr. Joe Picone for his mentoring and guidance throughout my graduate studies. I would also like to express my gratitude to Dr. Aravind Ganapathiraju and Jon Hamaker for their valuable suggestions.

References

1. L. R. Rabiner, B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, New Jersey, USA, 1993.

2. L. Lamel, J. L. Gauvain, G. Adda, "Lightly Supervised Acoustic Model Training," *Proceedings of the ISCA ITRW ASR2000*, Paris, France, September 2001.

3. G. Zavaliagkos, T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance," *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, Virginia, February 1998.

4. P. Placeway, J. Lafferty, "Cheating with Imperfect Transcripts," *Proceedings of the International Conference on Speech and Language Processing*, vol. 4, pp. 2115-2118, 1996.

5. T. Kemp, A. Waibel, "Unsupervised Training of a Speech Recognizer Recent Experiments," *Proceedings of ESCA Eurospeech'99*, pp. 2725-2728, Budapest, Hungary, September 1999.