EFFECTS ON TRANSCRIPTION ERRORS ON SUPERVISED LEARNING IN SPEECH RECOGNITION[1]

Ram Sundaram
Conversay
Redmond, Washington, USA
rsundaram@conversay.com

Joseph Picone
Institute for Signal and Information Processing
Mississippi State University
picone@isip.msstate.edu

ABSTRACT

Hidden Markov model-based speech recognition systems use supervised learning to train acoustic models. On difficult tasks such as conversational speech there has been concern over the impact erroneous transcriptions have on the parameter estimation process. This work analyzes the effects of mislabeled data on recognition accuracy. Training is performed using manually corrupted transcriptions, and results are presented on three tasks: TIDigits, Alphadigits and Switchboard. For Alphadigits, with 16% of the training data mislabeled, the performance of the system degrades by 12% relative to the baseline. On Switchboard, at 16% mislabeled training data, the performance of the system degrades by 8.5% relative to the baseline. An analysis of these results revealed that the Gaussian mixture model contributes significantly to the robustness of the supervised learning training process.

1. INTRODUCTION

Conversational speech is difficult to transcribe accurately [1,2]. On such tasks there has been concern over the impact erroneous transcriptions have on the parameter estimation process, especially given the cost of generating highly accurate transcriptions. The initial Switchboard (SWB) transcriptions had a word error rate (WER) of approximately 10% [2]. When initial research systems reported WERs on the order of 50% for SWB, it was conjectured that transcription errors contributed significantly to this poor performance. A three-year project was initiated to produce transcription error rates below 1% [2]. To our surprise, after reducing the transcription error rate to 1%, overall acoustic modeling accuracy did not increase significantly [3]. Hence, in this paper, we investigate the impact of clean transcriptions on the supervised learning process.

1. This material is based upon work supported by the National Science Foundation under Grant No. IIS-0085940. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Hidden Markov model-based speech recognition systems use supervised learning to train acoustic models in automatic speech recognition [4]. Typically, these learning algorithms need a set of accurate word-level transcriptions that correspond to the input training speech data. Requiring a set of error-free transcriptions increases the time required to develop new applications and is one of the most expensive aspects of data collection [4]. Other sources of transcriptions such as closed captions are readily available but are seldom used because their transcription error rates are high. Experiments have shown that is possible to have reasonable performance using these types of erroneous transcriptions [5]. Significant work has not been done to analyze and understand the underlying reasons why traditional HMM-based training algorithms are robust to mislabeled transcriptions.

2. SUPERVISED LEARNING

The details of model training using supervised learning techniques can be found in [6]. The output probability distribution at each state of an HMM is a continuous probability distribution typically modeled as a multivariate Gaussian distribution. The model parameters for each state, namely the mean and variance, are updated after every iteration. The mean calculation is given by:

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

where $L_{jm}^{r}(t)$ is the state occupancy probability, R is the total number of observations, T is the total duration of each utterance and o_t^r is the observation vector for frame t of utterance r. The state occupancy value can also be defined as the probability of the input data belonging to the model given the current model parameters. If the input data matches the model closely, it is likely that the state occupancy value will be high, and

the data contributes more to the model reestimation process and vice-versa.

3. EXPERIMENTAL DESIGN

To analyze the effect of mislabeled transcriptions on recognition accuracy, transcription errors were introduced in a controlled manner on three databases widely used within the community: TIDigits (TID), Alphadigits (AD) and Switchboard (SWB). Details about the experimental setup can be found in [7]. The experiments were performed using a publicly available speech recognition system [3]. Linguistically plausible errors were introduced in a random way designed to emulate how transcription errors occur in practice. Several methods were investigated for error generation, but all produced comparable results. A summary of the results for a range of recognition tasks is given in Table 1.

It can be observed from Table 1 that transcription errors do not make a significant impact on any of the databases. For Alphadigits, at a 2% transcription error rate, the performance of the system was not affected. With 16% of the data mislabeled, the performance of the system degrades by 12% relative to the baseline. Even for a complex database such as SWB, the word error rate degrades only by 3.5% (absolute) at a 16% transcription error rate.

To further investigate this perceived robustness, simulated experiments were performed using two one-dimensional overlapping Gaussian distributions [7]. The parameters of one Gaussian are estimated using a mixture of data generated from the in-class Gaussian and a percentage of data generated from a second, out of class Gaussian. The data was reclassified using an optimal decision boundary based on the estimated parameters, and probability of error calculated based on the estimated decision surface. These experiments produced results similar to those shown in Table 1 – parameter estimation was surprisingly robust to mislabeled data. Details of the experimental setup and results can be found in [7].

Experiments were then conducted to analyze how acoustically similar (‘b’-‘d’) and dissimilar (‘aa’-‘s’) phones perform in the presence of transcription errors. The means and variances for these phones were obtained from actual speech data. In this case, statistics collected from the AD task were used. The results are tabulated in Table 2. It can be seen that the probability of error is high even at a 0% percent transcription error rate for acoustically similar phones. This is because the distributions for these phones have significant overlap.

Note that as the transcription error increases the probability of error does not increase. In the case of acoustically dissimilar phones, the distributions have a small overlap and the probability of error is low at a 0% percent transcription error rate. As the transcription error rate increases, the probability of error marginally increases. This is due to the fact that Gaussian distributions tend to cluster around the mean of the data. Hence, even at a 20% transcription error rate, the estimate of the original distribution is not significantly different from the estimate of the original distribution for a 0% transcription error rate. In both the cases we see that corrupting the model does not increase the probability of error significantly. This behavior matches what was observed during recognition experiments on the three corpora previously described.

Database	Acoustic Models	Trans. Word Error Rate (TWER)		
		0%	2%	16%
TID	1 mixt.	3.8	4.0	5.1
	16 mixt.	0.8	1.0	2.3
AD	1 mixt.	31.9	32.3	36.2
	16 mixt.	10.8	10.8	12.1
SWB	12 mixt.	41.1	41.8	44.6

Table 1. A summary of results for systems trained with varying amounts of transcription errors (substitution errors). Each cell contains the corresponding WER for a recognition system trained at the given TWER.

4. ANALYSIS

To analyze the robustness of the training process in the presence of transcription errors, a subset of the Alphadigits database was used. This subset consisted of 4,884 utterances chosen at random. From this set, 100 utterances that had the word ‘o’ were chosen. In these 100 utterances, the word ‘o’ was replaced with the word ‘i’. Utterances with correct transcriptions were added back so that the subset now had 4,984 utterances and a transcription error rate of 7.8%. The motivation for substituting the word ‘o’ with the word ‘i’ is that both words have one phone, ‘ow’ and ‘ay’ respectively, in their pronunciations. Hence a substitution at the word level is equivalent to a substitution at the phone level.

The analysis was performed for all stages in the training process: monophone training, context-dependent

Data Error Rate (%)	Probability of Error	
	‘b’-‘d’ pair	‘aa’-‘s’ pair
0	44.1	6.84
4	44.1	6.89
8	44.1	7.25
16	44.1	7.70
20	44.1	7.87

Table 2. Probability of error for acoustically similar and dissimilar phones. Note that the probability of error does not increase significantly in either case.

training and mixture training. The acoustic models are standard three-state HMMs with self-loops and transitions to the next state. In each state, Gaussian mixtures were used to model the underlying distribution.

The results shown in Table 1 indicate that transcription errors do not degrade the performance of the recognition system significantly. Hence, the hypothesis is that the state occupancy values for the frames with erroneous data are very low and do not contribute to the model reestimation process. To verify this hypothesis, the state occupancy for the center state of the phone 'ay' was observed for the incorrect utterances (the utterances in which the word 'o' was replaced with the word 'i'). Similarly, the state occupancy for the center state of the phone 'ow' was also observed for the correct utterances (the 100 correct utterances that were added later to the list). The state occupancies were analyzed for all iterations of flat start and monophone training. Also, the state occupancy values were normalized by the number of frames for which their values were greater than zero. The normalized state occupancy values for the center state of the model 'ay' and 'ow' corresponding to the incorrect and correct utterances is shown for all stages of flat start and monophone training in Table 3.

It can be seen that the state occupancy values for the correct center state (corresponding to the model 'ow') are significantly higher than that of the incorrect center state (corresponding to the model 'ay'). Also, it was observed that the number of frames for which the state occupancies were greater than zero is significantly more for the correct state than for the incorrect state. In the utterances with transcription errors, the erroneous data typically gets mapped to the silence model. This shields the center state of the 'ay' model from the erroneous data. The incorrect data that occurs when 'ay' is substituted for 'ow' is rejected during the training process due to its low state occupancy value. Hence, the model learns very little from the incorrect data.

To verify how little the erroneous data contributes to the reestimation of the model (e.g., 'ay'), the state occupancy of the center state of the model 'ay' was analyzed from 275 correct utterances. The state occupancy for the center state of 'ay' in these 275 correct utterances was observed to be 0.53 after normalization while the state occupancy of 'ay' from the incorrect utterance is 0.148. This shows that the incorrect data does not contribute to the overall reestimation process significantly since its weights are low.

Similar experiments were performed during context-dependent training. The cross-word model 'sil-ay+ey', which had a transcription error rate of 16%, was chosen for analysis. Four out of 25 occurrences of this model were due to transcription errors. As in monophone training, one would expect the state occupancy values to be low for incorrect transcriptions and hence does not contribute significantly to the reestimation process. But, in the case of context-dependent training, each context-dependent model gets a smaller amount of training data compared to the monophone models. Hence, the percentage of incorrect data the model sees is likely to increase.

It is possible that the incorrect data contributes more to the reestimation process and the models can become corrupted. In Table 4 we see that the state occupancy values for the correct portion of the data is significantly more than the incorrect portion. But due to a relatively high transcription error rate (16% in the case of the model 'sil-ay+ey'), the state occupancy values increase after every iteration for the state in the incorrect transcription. However, this is insufficient to corrupt the model reestimation process.

During state tying, transcription errors for each model can change depending on the way the actual data was shared. After state tying is performed, the center state of 'sil-ay+ey' is shared with other models. This increases the number of instances of correct data for this model from 25 to 190 while the number of incorrect instances increases from 4 to 10. The transcription error rate for the model

	Average State Occupancy			
	Correct Transcription		Incorrect Transcription	
Iteration	Before	After	Before	After
1	0.5223	0.5829	0.0794	0.1490
2	0.5808	0.5807	0.0871	0.0851
3	0.5827	0.5913	0.1201	0.0873
4	0.5772	0.5915	0.1461	0.0873

Table 4. Average state occupancy values for the model 'sil-ay+ey' during context-dependent training before and after state tying. In both cases, the average state occupancy value for the model in the correct transcriptions is significantly more than those in the incorrect transcriptions.

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

Table 3. Average state occupancy values for the center state in the model 'ow' in the correct transcriptions and the model 'ay' in the incorrect transcriptions during monophone training. The state occupancy values are higher for the correct transcription. This difference widen after each iteration.

'sil-ay+ey' was reduced to 0.05%. Training was continued and the state occupancies were observed for the center state occurring in the correct and incorrect transcriptions.

The results are tabulated for the model 'sil-ay+ey' in Table 4. As before, the state occupancy value reduces after each iteration for the center state of the model 'sil-ay+ey' in the incorrect transcriptions. It can also be seen that the state occupancy value for the state occurring in the correct transcriptions increases after each iteration. This is because the transcription error reduces after state tying and the model is now exposed to more clean data than it was before state tying. Hence, the model effectively rejects the incorrect data.

The effect of transcription errors was also analyzed for mixture training using the same models as in context-dependent training. The results are tabulated in Table 5. It was observed that as the number of mixtures per state increases the state occupancy values decrease for the states updated by incorrect transcriptions while the values increase for states updated for correct transcriptions. As the number of mixtures is increased, the model tries to capture all the modalities in the correct portion of the data since correct data is present in much larger quantities. More details about experiments performed for analysis can be found in [7].

5. CONCLUSIONS

This paper has explored the robustness of supervised training algorithms to mislabeled data in speech recognition. The effects of different types of transcription errors were analyzed on three different databases: TIDigits, Alphadigits and Switchboard. HMM-based systems using Gaussian mixture models were shown to be robust to transcription errors. Individual transcription errors were shown to make an insignificant impact on the bias in the paramter estimates, and the estimation algorithms were shown to converge provided there were sufficient number of correct samples. Even high transcription error rates tend to be reduced in significance because these errors are split across a large number of parameters. Gaussian mixtures need a large amount of incorrect data to get corrupted. The process of iteratively training the models also adds more robustness to the acoustic models.

	Average State Occupancy	
Mixture	Correct Transcription	Incorrect Transcription
1	0.5372	0.1488
2	0.5384	0.1404
4	0.5644	0.1282

Table 5. Average state occupancy values for the model 'sil-ay+ey' after each stage of mixture training

This paper has shown that highly accurate transcriptions are not essential for training an acoustic model. It is possible to closely match the best performance by using other sources of transcriptions such as closed captions, provided there is ample data to overcome the deficiencies of the transcriptions. It would be interesting to quantify how much of these other sources of data are required to match a clean set of transcriptions in terms of system performance. For example, the system could be 90% accurate using 10 hours of clean training data on a database of interest. It is possible that this performance can be matched by using a significantly larger amount of noisy data. Quantifying the exact amount of noisy training data needed to match the performance of clean training data can be an interesting research area to explore in the future

6. REFERENCES

1. N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker and J. Picone, "Resegmentation of Switchboard," *Proceedings of the International Conference on Spoken Language Processing*, pp. 1543-1546, Sydney, Australia, November 1998.

2. J. Hamaker, N. Deshmukh, A. Ganapathiraju, J. Picone, "Resegmentation and Transcription of the SWITCHBOARD Corpus," *Speech Transcription Workshop*, Linthicum Heights, Maryland, USA, September 1998.

3. 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.

4. R.K. Moore, "A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners," *European Conference on Speech Communication and Technology*, pp. 2581-2584, Geneva, Switzerland, September 2003.

5. 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.

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

7. R. Sundaram, Effects of Transcription Errors on Supervised Learning in Speech Recognition, MS Thesis, Mississippi State University, May, 2003.