RECENT IMPROVEMENTS IN VOICEMAIL TRANSCRIPTION

M. Padmanabhan, G. Saon, S. Basu, J. Huang, G. Zweig

IBM T. J. Watson Research Center P. O. Box 218,
Yorktown Heights, NY 10598
(mukund,gsaon,sbasu,jhuang02,gzweig)@us.ibm.com

ABSTRACT

In this paper we report recent improvements in voicemail transcription. Last year, the speaker independent and speaker adapted word error rates (WER) on the Voicemail Transcription task were reported at 41.94% and 38.18% respectively. This year, we report a relative improvement of 18% in the speaker independent performance and 11% in the speaker adapted performance over last year. This improvement is a result of some new algorithms and an increase in the amount of training data. In the following sections, we describe the contribution of several components to improving the word error rate.

1. INTRODUCTION

In this paper we report recent improvements in voicemail transcription. The voicemail transcription task was introduced last year [1] as representing a style of conversational telephone speech that is somewhat different from the Switchboard and CallHome databases. Last year, the speaker independent and speaker adapted word error rates (WER) on this task were reported at 41.94% and 38.18% respectively, in [1]. This year, we report a relative improvement of 18% in the speaker independent performance and 11% in the speaker adapted performance over last year [2]. This improvement is a result of some new algorithms and an increase in the amount of training data. In the following sections, we describe the contribution of several components to improving the word error rate.

2. ACOUSTIC MODELS

2.1. Training/Test data

The starting point for the experiments reported in this paper was the system described in [1]. This system was trained on 20 hours of voicemail data (a superset of the Voicemail Corpus 1 available through the LDC - www.ldc.upenn.edu), and had a speaker independent error rate of 41.94% and speaker adapted performance of 38.18%. These error rates were reported on a test set comprising of 43 voicemail messages; this will be used as the development test set for the purpose of reporting results on various algorithms in the following sections.

We have continued our efforts to collect voicemail training data, and have succeeded in doubling the size of the database that was used last year. The training database now comprises 40 hours of speech, (400k words of text) and the size of the vocabulary has increased from 10k to 14k words.

2.2. System Description

The speech recognition system uses a phonetic representation of the words in the vocabulary (with an alphabet of 62 phones). Each phone is modelled with a 3-state left-to-right HMM. Further, we identify the variants of each state that are acoustically dissimilar by asking questions about the phonetic context in which the state occurs. The questions are arranged heirarchically in the form of a decision tree, and its leaves correspond to the bacis acoustic units that we model. A feature vector is extracted every 10 ms, and we model the pdf of each leaf of the decision tree, with a mixture of gaussians. The features we used were smoothed estimates of the Mel cepstra described in [3].

2.3. Model Complexity Adaptation

In our system, each leaf of the decision tree is modelled by a mixture of gaussians. In an earlier paper [4], we had described how to select the number of gaussians for a leaf. The essence of the algorithm is to start with a small baseline system, S1, and evaluate the probabaility of correct classification of the leaf in the training data. If this probability is below a threshold, t, it implies that the model for the leaf does not match the data for the leaf very well; hence, the resolution of the model for the leaf is increased by using the model for

We would like to acknowledge the support of DARPA under Grant MDA972-97-C-0012 for funding part of this work.

the leaf from a larger system, S2. The corresponding adapted system is referred to as S1xS2-t. The results are tabulated in Table II, and graphed in Fig. 1, and indicate that the performance of the adapted system is always somewhere between the performance of the S1 and S2 systems, and generally provides better performance for the same number of gaussians. Hence, it appears to be an efficient way of compacting a system, rather than improving on the best performance as obtained with our standard techniques. We also compare this model compression strategy with other techniques that use classical model selection criteria [5], such as BIC in determining the optimum number of gaussians. The results show that the MCA adapted systems for a threshold of 0.55 gives similar performance to the BIC system, however, it provides greater control of the tradeoff of complexity vs performance, through control of the threshold.

Table II

Old system					
Desc	10	20	40	100	150
Size	24k	44k	75k	126k	148k
WER	41.09	39.27	38.97	37.11	37.61
MCA system					
Desc	10x100 -0.55	20x100 -0.55	40x100 -0.55	100x150 -0.55	
Size	34k	52k	80k	127k	
WER	39.07	38.02	37.92	38.12	
Desc	10x100 -0.45	20x100 -0.45	40x100 -0.45	100x150 -0.45	
Size	27k	46k	76k	126k	
WER	40.58	39.17	37.41	37.71	
BIC System					
Size	45k	-			
WER	37.92	-			

2.4. Post-processing recognizer outputs using ROVER

In order to exploit the differences in the errors made by our systems, we used NIST's ROVER (Recognizer Output Voting Error Reduction) [6] as a post-processor of the various word hypotheses scripts provided by these systems. Such voting schemes create a consensus alignment from multiple scripts, and will work well when the systems being combined make independent errors. In the following, we will briefly describe the systems which were combined:

- The first system (BL) had 127k gaussians and 2709 leaves and represented the baseline (10ms frame rate, decision trees use left context and within-word right context only to predict context dependent variation of a phone).

- The second system (HF) is the equivalent of the above system, but uses a higher frame rate of 5 ms. Further, the HMM topologies were changed to preserve the same minimum duration for all phones as for the baseline. This system is used to rescore the top 100 hypotheses produced by the first system.

- The third system (RC) uses decision trees that use both left and right context across word boundaries. This system had 3017 leaves and 125k gaussians.

- The last system (SA) is a speaker adapted system described in [7].

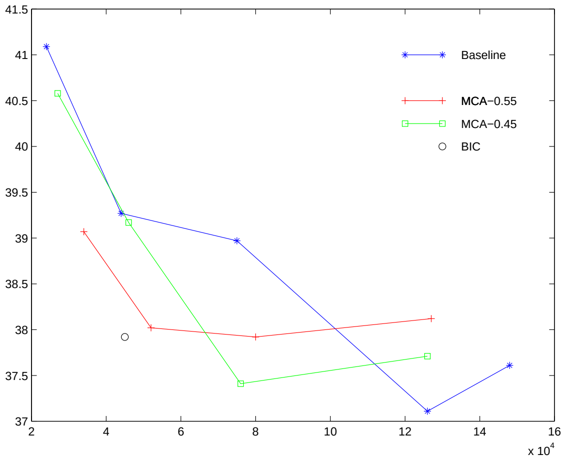

Figure 1: Word error rate vs number of gaussians

Empty scripts are used to avoid possible insertions and to transform substitutions into deletions if all the word hypotheses at a current step in the WTN are different.

Table III

Individual systems	
Baseline (BL) Right-context (RC) Higher frame rate (HF)	37.01% 38.47% 36.51%
Speaker-adapted (SA)	33.99%
Rover voting	
Rover1 = HF + BL + RC + empty SA + Rover1 + empty	35.45% 32.88%

As can be seen from Table III, ROVER reduces the speaker adapted WER by an additional 3.37% (relative).

2.5. Data driven approach to designing compound words

One way of modelling the pronuniciation variablility and co-articulation effects in spontaneous speech is to construct compound words [1, 8] and explicitly include baseforms to model the deviant pronunciations of the compound word. In [1], these were constructed by flagging actual instances of co-articulation in the training data, and in [8], by designing the words to reduce the overall perplexity. Here we experimented with alternative data driven approaches :

Denoting two successive words in the training corpus as W_t and W_{t+1}, we created compound word models for word pairs that had the largest values of the following quantity.

$$\sqrt{P(W_t = w_i/W_{t+1} = w_j)P(W_{t+1} = w_j/W_t = w_i)} \tag{1}$$

This can be seen to be just the square root of the product of the bigram probability $P(W_{t+1} = w_j/W_t = w_i)$, and a reverse bigram probability, $P(W_t = w_i/W_{t+1} = w_j)$. The rationale behind the use of this score is as follows : the words within the pair have to occur frequently together and more rarely in the pair context of other words. This requirement is necessary since one very frequent word, say a, can be part of several different frequent pairs, say $(a, b_1), \ldots, (a, b_n)$, $(b_{n+1}, a), \ldots, (b_m, a)$. If all these pairs were to be added to the vocabulary then the confusability between b_i and the pair (a, b_i) or (b_i, a) would be increased especially if word a has a short phone sequence.

The results with the use of this strategy to select compound words are tabulated in Table IV. The measure was applied iteratively. After one iteration, the pairs that score more than a threshold were transformed into compound words and all instances of the pairs in the training data were replaced by these new words.

Table IV

Iteration	# cpd words	Perplexity	WER
0	0	78	39.42%
1	42	103	37.45%
2	19	114	36.79%
3	9	117	36.64%

2.6. Modelling pdf's with non-gaussian models

Purely gaussian densities have been know to be inadequate for the purpose of modelling pdf's in speech recognition systems due to the heavy tailed distributions observed by speech feature vectors. In most of the speech recognition literature, pdf's are modelled as mixtures of gaussian densities. The only attempt to model the phonetic units in speech with nongaussian mixture densities is [10], where Laplacian densities were used with a heuristic estimation algorithm.

In [9] we attempted to address this problem by considering mixture models with the components defined as

$$p(x/\mu\Sigma) = \rho_d \frac{1}{\sqrt{det\Sigma}} exp\left[-\left((x-\mu)^T\Sigma^{-1}(x-\mu)\right)^{\alpha}\right] \tag{2}$$

The case $\alpha = 2$ corresponds to the gaussian density, whereas the laplacian case considered in [10] corresponds to $\alpha = 1$. Smaller values of α correspond to more peaked distributions ($\alpha \to 0$ yields the δ-function), whereas larger values of α correspond to distributions with flat-tops ($\alpha \to \infty$ yields the uniform distribution over elliptical regions). For more details about these issues see [9]. This particular choice of family of densities has been studied in the literature and referred to in various ways e.g., α-stable densities as well as power exponential distributions, cf. [11]. More recently, we have also become interested in automatically finding the 'best' value of α directly from the data.

Recognition experiments were carried out on the voicemail as well as the broadcast transcription task HUB4'98 by allowing different mixture components to have different values of the parameter α as compared with the fixed values $\alpha = 1$ and $\alpha = 2$. The preferred values of α tends to be less that 1.0, both for the voicemail and for the HUB4 task confirming on a systematic basis that nongaussian mixture components are preferred. An additional interesting point was that the distribution of the α values was much wider for the voicemail task than the HUB4 task. The reason for this could be the highly variable nature of the voicemail data.

Table I

Performance of α densities

Baseline (BL)	39.7%
$\alpha = 1$ (20 iterations)	38.5%
Prototype dependent α	38.8%

2.7. Other experiments

In addition to the experiments described above, we also experimented with a number of other techniques including (i) using Bayesian networks to incorporate dependencies on hidden variables that are not related to the normal linguistic quantities (ii) increasing the conditioning of the probability computation for a feature vector not just on the leaf at the current time, but also leaves at adjacent times, and (iii) adaptation techniques where we attempt to obtain better performance by starting from models that are better matched to the test speaker than a speaker-independent model, by clustering the training speakers into homogenous clusters. Details of these experiments are given elsewhere [12, 2, 7] and will not be repeated here.

3. CONCLUSION

In this paper we report recent improvements in voicemail transcription compared to last year. The overall performance (word error rate) on this task improved by 18% (relative) for our speaker independent system, and 11% for our speaker adapted system, respectively, and we describe the various components that brought about this improvement. Specifically, we experimented with

- model selection schemes for selecting the number of gaussians used to model the pdf - both the discriminant measure based scheme and BIC allow the model size to be compressed without loss in accuracy
- the use of voting schemes such as ROVER to create a consensus alignment from multiple scripts. This gave us a 3.4% relative improvement in performance
- data driven approach to designing compound words. This gave us a 7% relative improvement in performance
- use of non-gaussian parametric models to model the pdf
- use of Bayesian networks to introduce dependencies not related to linguistic quantities
- speaker adaptation techniques based on speaker clustering that provide a 10.5% relative improvement over the speaker independent performance

4. REFERENCES

[1] M. Padmanabhan et al., "Transcription of new speaking styles - Voicemail", Proceedings ARPA Hub4 Workshop, Lansdowne VA, Feb 1998. Also available at http://www.nist.gov/speech.

[2] G. Zweig et al., "Recent Improvements in Voicemail Transcription", Proceedings ARPA Hub4 Workshop, Herndon, VA, Feb 1999. Also available at http://www.nist.gov/speech.

[3] S. Dharanipragada et al., "Techniques for capturing temporal variations in speech signals with fixed-rate processing", Proceedings of the ICSLP 1998.

[4] L. R. Bahl and M. Padmanabhan, A discriminat measure for model complexity adaptation", Proceedings of the ICASSP, 1998.

[5] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian Information Criterion with Applications in Speech Recognition", Proceedings of the ICASSP, pp 645-648, 1998.

[6] J. G. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)", Proceedings of IEEE ASRU Workshop, pp. 347-352, Santa Barbara, 1997.

[7] J. Huang and M. Padmanabhan, "A study of adaptation techniques on a voicemail transcription task", elsewhere in the Proceedings.

[8] M. Finke and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition", Proceedings of Eurospeech '97, Rhodes, Greece, 1997.

[9] S. Basu, C.A. Micchelli and P. Olsen, "Maximum likelihood estimates for exponential type density families", Proceedings of ICASSP, Phoeniz, 1999.

[10] H. Ney, A. Noll, Phoneme modelling using continuous mixture densities, Proceedings of ICASSP, pp. 437-440, 1988.

[11] E. Gòmez, M. A. Gòmez–Villegas, J. M. Marin, A multivariate generalization of the power exponential family of distributions, Comm. Stat. — Theory Meth. 17(3), pp.589–600, 1998.

[12] G. Zweig and M. Padmanabhan, "Dependency Modeling with Bayesian Networks in a Voicemail Transcription System", elsewhere in the proceedings.