AUTOMATIC TRAINING SET SEGMENTATION FOR MULTI-PASS SPEECH RECOGNITION

Mark Z. Mao[†⋆]*, Vincent Vanhoucke*[⋆]*, Brian Strope*[⋆]

[†] Department of Electrical Engineering, Stanford University, CA, USA
[⋆] Nuance Communications, Menlo Park, CA, USA
markmao@stanford.edu, vincent@nuance.com, bps@nuance.com

ABSTRACT

A common approach to automatic speech recognition uses two recognition passes to decode an utterance: the first pass limits the search to a smaller set of likely hypotheses; and the second pass rescores the limited set using more detailed acoustic models which may target gender or specific channels. A question raised by this architecture is how to define and train the second pass models. Here we describe an extensible automatic solution that requires no manual gender or channel labeling. To train the second pass models, we cluster the training data into datasets containing utterances whose acoustics are most similar across the entire utterance. The clustering is based on which regions of a more general acoustic model are activated during forced alignments. Experiments with commercial English-American digit strings show 9.3% relative error rate reductions over a gender-based two pass system with similar numbers of model parameters.

1. INTRODUCTION

A well-known limitation of the Hidden Markov Model (HMM) approach to speech recognition is its inability to model long-term statistical dependencies in the speech signal. This causes the model to over-generate, and introduces potential confusions which affect accuracy adversely. As an example, while it takes a few seconds of speech to determine with reasonable accuracy if a speaker is male or female, most recognizers would not be capable of using this information to constrain its acoustic search, even though this information is relevant to the performance of an Automatic Speech Recognition system [1, 2].

Many approaches have been proposed to address this issue, mostly in the form of a "two-pass" decoding process [3, 4]:

1. learn the characteristics of the speaker and channel in the first decoding pass,
2. use a distinct model adapted to those characteristics to re-evaluate the recognition

Other approaches have attempted to embed the constraints into the architecture of the decoder in order to constrain the HMM search synchronously as more information is learned about the speaker and the environment [5, 6].

The two-pass approach is conceptually simple and in practice very powerful. The problem amounts to extracting relevant speaker and channel information from the utterance, and applying this knowledge by either adapting the model to those characteristics using well-known adaptation techniques [7, 8], or by picking a distinct recognizer which has been trained to match those characteristics. This latter "model selection" approach is particularly attractive when a lot of training data is available, because it is not limited by the amount of adaptation data available during decoding. Second pass models can be trained offline using an arbitrary amount of data which matches the characteristics of a broad range of speakers and environments.

The main issue related to building such systems has to do with separating the training data into the different subsets with which the second pass models are trained. This classification of utterances can be done in a supervised way, based on labels such as the gender of the speaker. However, these labels are arbitrary, and do not necessarily correspond well to the objective of improving recognition.

This classification task can be approached as a clustering problem over the training data: since the class labels are auxiliary to the task of coming up with a partition of the data which maximizes the recognition accuracy, we can ignore them and instead cluster the data based on a metric which directly relates to this objective. Assuming that a standard Vector Quantization (VQ) [9] algorithm can be used to perform this clustering, the question becomes one of defining an appropriate distance measure which will maximize the utility of those clusters for the purpose of maximizing recognition accuracy.

In Section 2, we propose such distance measure for the degree of similarity of the long-term statistics of a set of utterances. In Section 3, we describe the overall clustering algorithm in detail. Finally in Section 4, we detail the architecture of the resulting recognition system, explain the experiment setup and give the experimental results demonstrating the benefits of this approach.

2. DISTANCE MEASURE

The two-pass approach addresses the limitations of the HMM model by enforcing a distinct HMM recognizer for utterances which exhibit dissimilar characteristics. We will consider this feature of the system as our main design objective: to obtain a partition of the training data into clusters which are maximally dissimilar in terms of the long-term statistics, as perceived by a recognizer. This implies that a good distance measure will have to be able to evaluate the degree of similarity between an utterance U and a cluster C using the statistics accessible to a speech recognition system.

To achieve this, the simplest method is to consider the probability model p of the recognizer, to adapt it to both the utterance and the cluster, and to measure the divergence $D(p^U \| p^C)$ between

both adapted models. Because the evaluation of the distance measure has to be performed on each utterance of the training set repeatedly during the iterations of the VQ algorithm, the sufficient statistics used to compute the divergence have to be compact and fast to compute. In addition, this distance measure has to be robust to the lexical variability of the utterance, so that each cluster gets allocated data with a rich enough phonetic coverage to be robust to any speech.

In order to satisfy the compactness constraints, the adaptation is performed exclusively on the mixture weights of the GMM of the recognizer. The underlying idea is that different Gaussians in the GMM would be active for utterances with different statistical characteristics. To illustrate this, one could imagine a gender independent system to contain Gaussians which are predominantly "females", and others predominantly "male", as was observed in [6]. By adapting the mixture weights to the current utterance, one can determine its degree of "female-ness" by comparing its statistics to the typical pattern of Gaussians active in a female context.

In order to satisfy the computational constraints, the adaptation for clustering is performed using a single iteration of the EM algorithm [10]. Let us assume that $u_1, ..., u_K$ is a subset of observations in utterance U that are drawn from distribution p. The cluster centroid $C = c_1, ..., c_q$ is the set of all observations drawn from distribution p from all utterances in a cluster. A GMM model has probability density $p(o) = \sum_{j=1}^{M} w_j \mathcal{G}_j(o)$, where $\mathcal{G}_j$ is the j^{th} Gaussian and w_j its mixture weight. The KL divergence [11] between the adapted GMM p^U and p^C can be written as:

$$D(p^U \| p^C) = \int_o [\sum_{j=1}^{M} w_j^U \mathcal{G}_j(o)] \log \frac{\sum_{l=1}^{M} w_l^U \mathcal{G}_l(o)}{\sum_{l=1}^{M} w_l^C \mathcal{G}_l(o)} \quad (1)$$

Where w_j^U can be expressed as (resp. w_j^C):

$$w_j^U = \frac{1}{K} \sum_{k=1}^{K} \frac{w_j \mathcal{G}_j(u_k)}{\sum_{l=1}^{M} w_l \mathcal{G}_l(u_k)} \quad (2)$$

The formulation of the divergence can be further simplified by making the assumption, generally valid for Gaussians in high dimensions, that in the region where a given Gaussian contribution is dominant, i.e. wherever $w_j \mathcal{G}_j(o) >> 0$, the log-likelihood of the GMM is dominated by this most significant term. This implies that:

$$w_j \mathcal{G}_j(o) \log[\sum_{l=1}^{M} w_l \mathcal{G}_l(o)] \approx w_j \mathcal{G}_j(o) \log[w_j \mathcal{G}_j(o)] \quad (3)$$

With that assumption:

$$D(p^U \| p^C) \approx \sum_{j=1}^{M} w_j^U \int_o \mathcal{G}_j(o) \log \frac{w_j^U \mathcal{G}_j(o)}{w_j^C \mathcal{G}_j(o)} \quad (4)$$

$$\approx \sum_{j=1}^{M} w_j^U \log \frac{w_j^U}{w_j^C} = D(W^U \| W^C) \quad (5)$$

where $W^U = w_1^U, ..., w_M^U$, $W^C = w_1^C, ..., w_M^C$ are the discrete densities of the Gaussian weights for the models adapted from utterance U and the cluster centroid C respectively. Using the KL divergence of the distribution of the posterior probabilities of the Gaussians corresponds well to the intuitive idea that we are discriminating between utterances which activate different areas of the acoustic space. This formulation is also very tractable computationally since the weights w_j^U can be accumulated once, and used to derive the w_j^C by averaging them over the whole cluster.

The last issue to address is the lexical independence: in practice, an acoustic model comprises several GMMs $p_i(o), i = 1 \ldots N$. Not all these GMMs will be observed by a given utterance, depending on its phonetic content. A normalized distance measure can be achieved by adapting each GMM independently, and only measuring the distance for those N_u GMMs which the utterance U actually observes. The total distance measure is thus:

$$D_{Total}(p^U \| p^C) = \frac{1}{N_u} \sum_i D(p_i^U \| p_i^C) \quad (6)$$

Since the clusters typically comprise a good fraction of the available training data, all the GMMs can safely be assumed to be covered by the data in the cluster. In addition, the recognizer considered uses a strongly tied acoustic model with only a few hundred tied GMMs [12], which reduces the sparsity of the adaptation sufficient statistics collected from each utterance.

3. CLUSTERING ALGORITHM

The algorithm used to segment the training data into clusters is described below in more details.

3.1. Obtaining Observation Statistics

Given a transcribed utterance $U = u_1, ..., u_p$, the information regarding which GMM each observation u_i is drawn from can be obtained based on its transcription through a Viterbi search, which finds the best phonetic sequence that matches the transcription while maximizing the likelihood of observations $u_1, ..., u_p$. The Gaussian weights of the adapted utterance model then can be computed from Equation 2.

3.2. Cluster Initialization through PCA

To obtain initial clusters, a simple method is to divide training utterances randomly, with the expectation that, through the Lloyd iterations, the clusters will converge to a good local minimum of the total divergence. However, a more principled approach can be used: Equation 5 shows that the distance between two adapted models can be approximated by the KL divergence between the discrete distributions of the Gaussian weights. For utterance U, the adapted model p^U has N GMMs and M Gaussian weights per GMM. By representing it as a $M \times N$ dimensional vector in the Euclidean space, Principal Component Analysis (PCA) [13] is then applied on a subset of training utterances to separate them into two clusters along the largest variance direction. Even though the Euclidean distance is different from KL divergence, the PCA split of the data produces reasonably good initial clusters.

3.3. The Lloyd Iterations and Tree Structured VQ

Once the initial clusters are obtained, the VQ algorithm alternates between two steps:

- The nearest neighbor of every utterance is determined using Equation 5,
- The centroid of each cluster is then re-estimated by accumulating the statistics of all utterances in the cluster, and applying Equation 2.

To separate data into R clusters, a Tree Structured VQ [9] approach is used. The training set is split into two clusters through PCA and Lloyd iterations, and the algorithm is applied recursively to the resulting clusters, until the desired number of clusters is reached. Between each step, new models are trained from the sub-clusters and the utterance models are re-estimated from the GMMs in the new models. This provides a finer resolution for the Gaussian weight distribution, since the newly trained cluster model covers a smaller region in the acoustic space.

4. EXPERIMENTS

4.1. Recognition System

The resulting recognition system uses one first pass which is trained from all training data, and R second pass models trained from scratch from the data lists for each cluster. During decoding, the first pass recognizer generates a set of hypotheses, a second pass selector picks which second pass model decodes the utterance best based on the maximum likelihood match of the acoustics, and the second pass recognizer uses that model to find the best answer from the set of hypotheses.

4.2. Experimental Setup

The experiments were run on a speaker-independent American English digits recognition system trained from 600k digits utterances. The utterances contain digits and a small number of filler words such as *call*, *dial*, *please* and *thank you*. The utterances are of different gender, channel conditions and noise level. The front end features are 27 dimensional, including MFCC, Δ and $\Delta\Delta$. The test-set is a collection of 16000 American English digits utterances. The length of the utterances varies from one to ten digits. The accuracy is evaluated at the sentence level, ignoring all filler words. The recognition engine used is a context-dependent HMM system. Each state cluster shares a common set of Gaussians called Genone [12], while the mixture weights are state-dependent. The first pass recognizer is the same across all experiments. It is trained from all 600k utterances, using 500 Genones and 32 Gaussians per Genone. The first pass recognizer has an error rate of 10.28%.

4.3. Baseline Systems

We considered two baseline systems. The first one is a two pass system that uses only one second pass model which is a generic Gender Independent model. This system has slightly lower error rate compared to a simple one pass system with the same Gender Independent model because the re-scoring pass is allowed to use statistics collected over the whole utterance, as opposed to the estimates computed in real time in the initial pass. The second baseline is a two pass system based on supervised male/female split of the training set. To make the comparison fair, all systems in the experiments have an identical first pass model, and the same number of Gaussians in the second pass.

4.4. Experimental Results

Table 1. *Systems using 32k Gaussians in the 2nd pass*

2nd Pass Models	# 2nd p. × Genones × Gauss	Err. Rate
Gender Ind.	1 × 1000 × 32	9.33%
Gender Ind.	1 × 500 × 64	9.83%
Male/Female	2 × 500 × 32	8.18%
Clustering	2 × 500 × 32	8.08%

Table 1 shows the recognition error rates for systems with 32k second pass Gaussians. The first two rows show results of the first baseline system where there is only one second-pass model. The first system has more Gaussian mixtures in the model, while the second has more Gaussians per mixture. Both results are presented to demonstrate that the parameters are efficiently utilized. The corresponding one pass system (not shown in the table) with 1000 Genones and 32 Gaussians per Genone has an error rate of 10.25%. The third row shows the results of the second baseline system, with male and female second pass models based on a supervised split of the training data. The fourth row shows the results of the new system, where two second pass models are trained from clusters resulting from the automatic segmentation algorithm proposed in this paper. The new system shows a 13.4% relative error rate reduction against the best result with a single second pass, and a 1.2% relative error rate reduction against the gender split baseline system. Inspection of the training data shows that the clusters obtained from the automatic segmentation algorithm are very similar to those generated by the supervised gender split. This demonstrates that the unsupervised clustering was able to pick out the feature which we know to be most relevant to multi-pass recognition.

Table 2. *Systems using 64k Gaussians in the 2nd pass*

2nd Pass Models	# 2nd p. × Genone × Gauss	Err. Rate
Gender Ind.	1 × 2000 × 32	9.14%
Gender Ind.	1 × 500 × 128	9.36%
Male/Female	2 × 1000 × 32	8.22%
Male/Female	2 × 500 × 64	8.69%
Clustering	4 × 500 × 32	7.62%

Table 2 shows the main benefit of the proposed approach, namely that we can perform more splits of the data and improve the system further. Row 1-2 show the results of the single second pass system using 64k Gaussians in the second pass. The corresponding one pass system (not shown in the table) with 2000 Genones and 32 Gaussians per Genone has an error rate of 10.19%. Row 3-4 show the results of the male/female two pass system. Row 5 shows the new system, with the training set segmented into 4 clusters which provide data sets to train 4 second pass models. The new system has a 16.6% relative error rate reduction compared to the best result of the single second pass system and a 7.3% relative error rate reduction compared to the best result of the male/female two pass system.

Table 3 shows that the number of clusters can be increased further and lead to further error rate reduction. Row 1-4 show the results of the two pass systems using 128k Gaussians in the second pass. The corresponding one pass system (not shown in

Table 3. *Systems using 128k Gaussians in the 2nd pass*

2nd Pass Models	# 2nd p. × Genone × Gauss	Err. Rate
Gender Ind.	1 × 4000 × 32	9.05%
Gender Ind.	1 × 500 × 256	8.89%
Male/Female	2 × 2000 × 32	7.95%
Male/Female	2 × 500 × 128	8.16%
Clustering	8 × 500 × 32	7.21%

the table) with 4000 Genones and 32 Gaussians per Genone has an error rate of 10.08%. The new system, with 8 second pass models, as shown in row 5, achieves a 18.9% relative error rate reduction compared to the best result of the single second pass system, and a 9.3% relative error rate reduction compared to the best result of the male/female two pass system.

These results show that using a large number of second pass models can lead to better accuracy with little overhead cost. In terms of decoding speed, there is little difference between the speed of a system with 8 second passes and a system with a single second pass, provided that the cost of second pass selection is negligible. The memory footprint of the system grows with the number of second passes, although sharing of the parameters across passes [14] can significantly reduce the size and complexity of the system while maintaining accuracy.

5. CONCLUSION

We have introduced a new approach for designing two pass speech recognition system to address the well know limitation of the GMM/HMM model – its inability to model long range dependencies in the statistics of the speech. We proposed an algorithm to segment training utterances into clusters based on their similarities in the acoustic space, so that more focused second pass models can be trained. The segmentation process is fully automatic and requires no manual labeling of gender or channel. We showed that this approach improved significantly upon single pass systems as well as state of the art two pass systems which were built based on a supervised gender split. We also showed that increasing the number of clusters improved the system even further, while having a negligible effect on the complexity of the overall system.

6. ACKNOWLEDGMENTS

The authors would like to thank R.M. Gray, A. Sankar, C.J. Leggetter and M.M. Hochberg for their input. This research was supported in part by NSF grant CCR-0309701 and by Nuance Communications.

7. REFERENCES

[1] R. Vergin, A. Farhat, D. O'Shaughnessy, "Robust Gender-dependent Acoustic-phonetic Modelling in Continuous Speech Recognition Based on a New Automatic Male/Female Classification," *Proceedings of ICSLP 96*, vol. 2 pp. 1081-1084, 1996.

[2] W.H. Abdulla and N.K. Kasabov, "Improving speech recognition performance through gender separation," *Proceedings of the Fifth Biannual Conference on Artificial Neural Networks and Expert System*, p. 218-222, 2001.

[3] P.C. Woodland, "Speaker Adaptation: Techniques and Challenges" *Proceedings of IEEE ASRU Workshop*, pp. 85-90, Keystone, Colorado, USA, December 1999.

[4] M.J.F. Gales, "Acoustic Factorisation," *Proceedings of IEEE ASRU Workshop*, Madonna Di Campiglio, Trento, Italy, December 2001.

[5] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, 4(5):360-378, 1996.

[6] R. Teunen, "Acoustic Modeling For Automatic Speech Recognition: Deriving Discriminative Gaussian Networks," *PhD thesis, Department of Electrical Engineering, Stanford University*, August 2002.

[7] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, Vol. 9, pp. 171-185, 1995.

[8] L.F. Uebel and P.C. Woodland, "An Investigation into Vocal Tract Length Normalisation," *Proceedings of Eurospeech 99*, pp. 2519-2522, Budapest, 1999.

[9] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. of Royal Statistical Soc., Ser. B., vol. 39, no.1, pp. 1-38, 1977.

[11] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[12] V. Digalakis, P. Monaco, and H. Murveit, "Genones: Generalized mixture tying in continuous hidden Markov model-based speech recognizers," *IEEE Transactions on Speech and Audio Processing*, 4(4):281-289, 1996.

[13] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, 2001.

[14] M.Z. Mao, V. Vanhoucke, "Design of Compact Acoustic Models through Clustering of Tied-Covariance Gaussians," *To appear in Proceedings of Interspeech 04*, Jeju, Korea, October 2004.