CONSTRUCTING ENSEMBLES OF DISSIMILAR ACOUSTIC MODELS USING HIDDEN ATTRIBUTES OF TRAINING DATA

Takashi Fukuda[1], *Ryuki Tachibana*[1], *Upendra Chaudhari*[2], *Bhuvana Ramabhadran*[2], *and Puming Zhan*[3]

[1]IBM Research – Tokyo, IBM Japan Ltd.
[2]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
[3]Nuance Communications Inc., Boston, MA, USA
E-mail:{fukuda1, ryuki}@jp.ibm.com, {uvc, bhuvana}@us.ibm.com, Puming.Zhan@nuance.com

ABSTRACT

One of the objectives in acoustic modeling is to realize robust statistical models against the wide variety of acoustic conditions that are present in real world environments. As large amounts of training data become available, modeling subsets of the data with similar acoustic qualities can be done accurately and multiple acoustic models are jointly used as a form of system combination or model selection. In this paper, we propose a method to partition the training data for constructing ensembles of acoustic models using metadata attributes such as SNR, speaking rate, and duration via a binary tree. The metadata attribute used at each binary split in the decision tree is obtained using a metric proposed in this paper that is cosine-similarity based. The resulting multiple models are combined using voting techniques such as n-best ROVER. The proposed method improved the recognition accuracy by up to 4% relative over the state-of-the-art system on a large vocabulary continuous speech recognition voice search task.

Index Terms— Automatic speech recognition, multiple acoustic modeling, system combination, large corpora

1. INTRODUCTION

Hidden Markov Models (HMMs) have proved to be an effective basis for modeling time-varying sequences of speech data. However, in order to accurately capture the variations in real speech, it is necessary to have a large number of models and to use relatively complex output probability distributions. HMMs are typically built on cross-word context-dependent states. However, many non-phonetic features can be used to improve the resolution of the acoustic model and obtain higher performance. Examples of such features include gender, speaker or speaker group identity, speaking rate, channel and environment condition, ambient noise level, etc. One approach to model these features as tags via state clustering in a maximum likelihood framework was first proposed in [1]. An alternate approach involves explicitly partitioning the training data using many of these attributes and building separate models that capture the specific characteristics of all the data in the individual partitions. This leads to several model combination and selection strategies [2, 3]. This paper uses existing strategies for system combination and addresses how to build dissimilar acoustic models for an efficient system combination.

The rest of this paper is organized as follows. Section 2 provides a brief survey of ensemble Large Vocabulary Continuous Speech Recognition (LVCSR) systems. Section 3 describes the proposed algorithm for partitioning training data, and Sections 4 and 5 cover the experimental results. Finally, Section 6 presents our conclusions.

2. BACKGROUND

Modeling acoustic variability is a tough and well-researched problem and several techniques have been proposed in the literature. These include, universal models [4], ensemble models [5, 6, 7], gender or speaker-dependent models, adaptation and models trained on subsets of the data with similar acoustic properties. An example of ensemble methods proposed in [8] discovers factors of acoustic variability from the data in a hierarchical fashion, resulting in the modeling of the long tail of acoustic conditions present in the data. Partitioning the input data space to create this ensemble of models can occur in a supervised and unsupervised fashion with the goal of keeping these partitions maximally dissimilar from one another. Every utterance is represented by a sparse vector of Gaussian posteriors and the Kullback-Leibler distance is used to cluster these utterances iteratively to yield a tree of clusters of utterances. The analysis presented in the paper shows that there is a relation between utterance loudness, SNR, gender, pitch and perplexity and the nodes of the tree grown in this fashion [8]. While the root models are trained with manual transcriptions, augmenting them with additional untranscribed data did not matter much. However, the untranscribed data was used to augment the remaining nodes of the tree that partitions the training data space. During decoding, there is no system combination of the various models associated with the tree, however, selecting the model(s) that best represents the test utterance could be useful in minimizing the WER of a larger percentage of the test data [8]. While the paper does not present an overall relative improvement in WER, it demonstrates a potential for improving LVCSR performance by partitioning the training data based on its acoustic characteristics.

System combination is a promising way to obtain a significant error reduction in an overall WER. The gains from combining multiple outputs are increased when the models are built with maximally different acoustic properties. This tends to result in greater diversity between the hypotheses than when using systems from a single model. A successful system combination and model selection approaches require

the construction of multiple systems with complementary errors, or the combination will not outperform any of the individual systems. The system combination is, for example, realized by building systems on different features. This paper describes one approach to build maximally dissimilar acoustic models for system combinations.

3. ALGORITHMS

In this paper, we propose to do tree-structured training data splits by explicitly exploiting non-phonetic features including SNR, speaking rate, and duration unlike an unsupervised clustering. The data splits are done by representing an utterance with a vector composed of Gaussian posteriors and evaluating a set of posterior vectors with an objective function as described in the following sections.

3.1. Posterior Vector Representation

Every utterance in the training data set $\chi = \{x_1, x_2, ..., x_n, ..., x_N\}$, where x_n is an utterance and N is the data size, is represented using a D-dimensional vector of Gaussian posterior probabilities. The posterior vector is averaged in the utterance to make it a single vector representing the utterance

$$p(\boldsymbol{y}_n|g_i) = \frac{1}{T}\sum_{t=1}^{T}\frac{p(\boldsymbol{y}_{nt}|g_i)}{\sum_{k\in\mathcal{G}} p(\boldsymbol{y}_{nt}|g_k)}, \quad (1)$$

where $\boldsymbol{y}_{nt}$ is the feature vector for utterance x_n at time t, and $p(\boldsymbol{y}_{nt}|g_i)$ is the likelihood of i-th Gaussian in $\mathcal{G}$. The posterior vector $\boldsymbol{p}_n$ for the utterance x_n is composed of $\boldsymbol{p}_n = [p(\boldsymbol{y}_n|g_1), p(\boldsymbol{y}_n|g_2), \cdots, p(\boldsymbol{y}_n|g_D)]^T$. These posterior vectors are used to evaluate goodness of separation of the data.

Beaufays et. al. suggested that each utterance can be represented by a sparse supervector of Gaussian posteriors, whose intrinsic dimensionality is the total number of Gaussians in the system [8]. In contrast, the supervector consisting of Gaussian posteriors in our proposed approach were generated not from all Gaussian in the system but from G_N Gaussian mixtures obtained by clustering down one of our baseline models because the supervector is expected not to reflect contents of utterances but to represent acoustic conditions only. Reduced Gaussians created by clustering are generalized against utterances as used in a feature space discriminative training, and thus we expect that the supervector of reduced Gaussian set represents acoustic conditions more, that are less sensitive to contents of utterances. In this paper, we used G_N = 512.

3.2. Tree Construction with the Cosine Similarity-based Metric

The training data set is hierarchically divided by the most dominant attribute of non-phonetic features at each level and the posterior vectors are used to judge which attribute is the best for multiple acoustic modeling strategy. The training data is partitioned using the following procedure.

1. The training data set χ is split into two clusters χ_1 and χ_2 using an attribute of non-phonetic features. If the feature is continuous value such as SNR, the training data is split at the median of the value corresponding to the equal-sized splits.
2. A split score is computed by using posterior vectors of utterances. If a set of utterances has similar acoustic characteristics, the posterior vectors point in the similar direction. Thus the values of cosine distance between posterior vectors within the same cluster should be close to 1. In contrast, the between-class cosine similarity of χ_1 and χ_2 is expected to be small. Thus the similarity of the divided training data sets $\{\chi_1, \chi_2\}$ is represented as a cosine similarity metric consisting of within-class similarity c_W and between-class similarity c_B as
$$c_W = \frac{1}{N}\sum_{i=1}^{2}\sum_{p_n\in\chi_i}\frac{\boldsymbol{p}_n\cdot\boldsymbol{m}_i}{|\boldsymbol{p}_n||\boldsymbol{m}_i|}, \quad (2)$$
$$c_B = \frac{\boldsymbol{m}_1\cdot\boldsymbol{m}_2}{|\boldsymbol{m}_1||\boldsymbol{m}_2|}, \quad (3)$$
where $\boldsymbol{m}_i$ is the mean vector of $\boldsymbol{p}_n$ in the i-th cluster and N is the data size in the whole training data set χ. The cosine similarity-based split score is defined as
$$J_c = c_W - \alpha c_B, \quad (4)$$
where α is the scaling factor. A larger score is better for data splits because we can regard it as more dissimilar.
3. The split score is computed for the remaining non-phonetic features under consideration and the best feature is the one with the largest score and the data is split using this feature.
4. Let $\hat{\chi_1}$ and $\hat{\chi_2}$ be maximally dissimilar data sets divided using steps 1 to 3. This split process can be hierarchically repeated on the binary tree structure against resultant $\hat{\chi_1}$ and $\hat{\chi_2}$ until the desired number of splits are reached.

In this paper, we used non-phonetic features to create binary partitioned data sets, but any algorithm to partition the data such as the unsupervised clustering based one proposed by Beaufays et. al [8] can be used and the quality of the splits can be evaluated using the metric presented here.

Instead of using cosine similarity metric we addressed in the paper, we can think of using variance-based metric. But the variance-based partitioning can be unexpectedly dependent on the contents of the utterances. When the utterance is phonetically well-balanced, the posterior vector $\boldsymbol{p}_n$ can be a good representation of only the acoustic characteristics as techniques such as Cepstral Mean Subtraction (CMS) average out the channel related characteristics. In contrast, if the utterances are short or not well-balanced, the posterior vector includes phonetically biased components, and thus cosine similarity metric can be better for utterance representation than variance-based metric.

4. EXPERIMENTS

The experiments presented in this paper are all based on speaker independent models that are discriminatively trained (DT) on an LVCSR voice search task. We present results on an in-house test set for voice search in English. To date, no standardized test exists in the community to benchmark systems for the voice search task. However, similar tasks

Table 1. Baseline performance with DT models.

	WER%	
Training Data Set	Dev	Eval
Set A	24.0	25.0
Set B	22.5	23.1
Set C	20.8	21.4

have been studied in the literature [9, 10] where the baseline systems range in WERs from 16% to 25%. The acoustic models are built on data from several hundreds of speakers with the data ranging from a few seconds to few hours per speaker.

4.1. Baseline Models

The front-end acoustic features are 13-dimensional PLP features. Utterance level mean normalization, where the statistics are calculated only on the speech regions of the data, is used throughout the Maximum Likelihood (ML) and discriminative training steps. In ML training, LDA+MLLT transforms are generated by splicing 9 frames of PLP features and reducing the feature vector to the 40-dimensional feature space [12]. We build three sets of acoustic models trained from different quantities of data. Training set A comprises of about 150 hours, training set B is made up of approximately a thousand hours and training set C comprises of an order of magnitude more data i.e. over 5000 hours (Table 1). The ML models for set A contain roughly 150K Gaussians with 5000 quinphone context-dependent states while the ML models for set B contain roughly 200K Gaussians with 7K states. Models trained with set C contain roughly 600K Gaussians and 20K states. After ML training, the models are discriminatively trained using the boosted MMI criterion. Results are presented on 2 test sets: **Dev** and **Eval** with approximately 4K and 70K words.

Table 1 tabulates the performance of the baseline models for the development and evaluation test sets. Consistent with what has been reported in the literature, a five-fold increase in training data size yields a reduction in WER of about 5 to 7% relative while an order of magnitude increase in the amount of training data leads to approximately 14 to 16.5% relative improvement in the WER. Several training methodologies to derive benefits from additional training data have been presented in the literature with a consistent message that one has to carefully select the most beneficial parts and find a model that best represents the selected data.

5. RESULTS

5.1. Generated Binary Decision Tree

In this section, we discuss a binary tree generated from cosine similarity-based metric. Figure 1 shows the generated tree. A set of non-phonetic features including random criterion which means data set is randomly partitioned into two clusters, audio duration, silence duration before speech, lattice density, the average log likelihood of an utterance, SNR, speaking rate, silence duration after speech ends, usage frequency, and utterance duration are used to build the tree. In the figure, each node of the tree corresponds to a data cluster. Looking at the tree, the non-phonetic feature of SNR appeared at the root node. This means that SNR is the most important factor to build the dissimilar models from one another. Considering the second level of the tree, SNR is dominant again in N_0 nodes, but N_1 node is split using a different feature, which is speaking rate here. When we built the tree with variance-based metric in another experiment, a non-phonetic feature of duration appeared in N_1 node. The duration of an utterance relates to the bias of utterance, and hence we hypothesize that the variance-based metric is unable to properly handle the biased short utterances. In contrast, the cosine similarity metric is more robust against such bias and speaking rate is chosen as the attribute to split the data in N_1 node. Looking at the third level of the trees, SNR was selected again as a split factor in N_{00} node. This indicates that low SNR data has strong individuality and can configure the particular acoustic space effectively along multiple AM strategy. In addition to SNR, the third level includes likelihood and duration as other useful piece of metadata for partitioning the feature space. Surprisingly, only 4 attributes of the non-phonetic features are dominant in tree generation.

5.2. ASR Results using Middle Size Corpus

We tested the complementary models trained using our proposed algorithm and compared it to other criteria. We split training Set B into the partitions obtained at the third level (8 clusters) produced by the tree and built eight acoustic models of more or less the same size. Each cluster has roughly one eighth of the training data. A single system was trained with ML criterion using the entire Set B and this was used to build the eight discriminatively trained models using cluster-specific data. This process could be interpreted as an unsupervised discriminative adaptation. The model structure (topology) and dimension of the model are not affected but its parameters are re-estimated with MMI criterion on both feature and model space using the cluster-specific data. Individual models comprise 150K Gaussians with 5000 states, similar to the baseline system trained with Set A (lesser data).

The experimental results on the development test set are given in Table 2. In this table, the baseline is a single DT model trained with Set B. "ROVER with random split" refers to the cluster models built from randomly partitioned training data into clusters that are the same size as Set A. The proposed cosine similarity-based split is also used to partition Set B and the results are compared against the random splits. We used n-best ROVER [3] for system combination. In the ROVER combination, the baseline system was also included. Models built from random splits are acoustically well balanced and thus each random-cluster model had higher performance than cosine similarity based cluster models. However the dissimilarity of such models is far less. Therefore, the system combination using random-cluster models had poor performance while the proposed method based on cosine similarity metric showed the best gain of 2.7% relative from the individual baseline system trained with Set B.

5.3. ASR Results using Large Size Corpus

Similar to the previous section, to better understand the scalability of the proposed method, we split the training set C into 8 clusters of equal size. The individual cluster models have approximately 200K Gaussians with 7000 states. In the previous, we used an ML-trained single system to build eight mul-

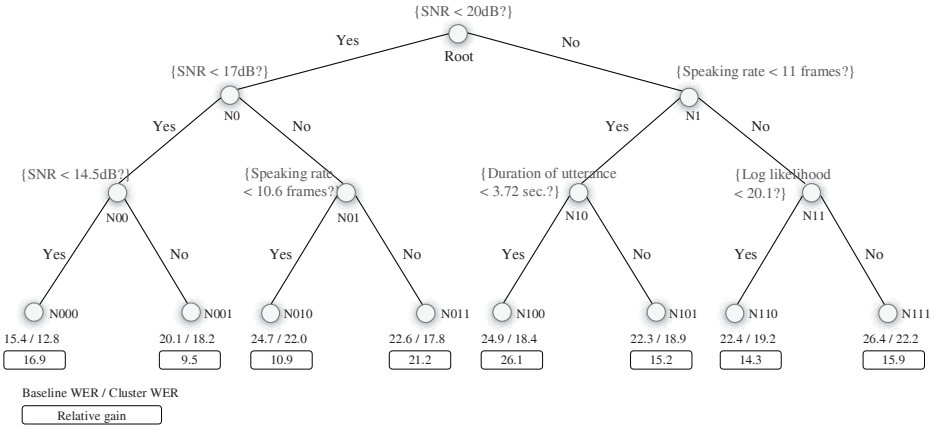

Fig. 1. Tree illustrating splits based on cosine similarity-based metric.

Table 2. Comparison of models built with different partitioning criteria on development test set

System	WER%
Set B-based baseline single model	22.5
ROVER with random split	22.3
ROVER with proposed cosine split	21.9

Table 3. Results with system combination on both development and evaluation test sets

Combination	WER%	
	Dev	Eval
Set C-based baseline single model	20.8	21.4
ROVER with proposed cosine split	20.0	20.5
Model selection (Oracle)	16.6	18.2

tiple systems using discriminative training. This means each acoustic model shared the same decision tree. In this experiments, we trained the acoustic models to have different decision trees to enforce more dissimilarity. Table 3 presents the experimental results with our proposed method on the development and evaluation test sets. It can be seen that the WER dropped from 21.4% to 20.5 (4.2% relative) on the evaluation test set.

In addition to system combination, we also considered oracle WERs when the best cluster model was selected for test utterances. Each test utterance was clustered by searching the cluster which gives lowest WER per speaker. The WERs of test utterances belonging to the cluster were filled in Figure 1 and the overall WER in model selection strategy was added to Table 3. Looking at Figure 1, it can be seen that the proposed method had siginificant gains in N_{000}, N_{011}, and N_{100} nodes. This result suggests that utterances featured by low SNR, slow utterance, and short duration are difficult to be recognized accurately, therefore cluster models trained by subset of the data with similar acoustic condition are promising to decrease an overall WER. If we could pick the best model, 15% relative improvement can be seen on the evaluation set.

6. CONCLUSION

This paper proposes a novel approach to partition the training data with a focus on effective construction of acoustic model ensembles. The data split is done by using non-phonetic features of the training data and the resulting clusters are evaluated based on the cosine-similarity-based objective function to make them maximally dissimilar. System combination using these cluster models trained with data split by our proposed method showed gains of up to 4% relative over the state-of-the-art system. In future work, we will investigate an automatic model selection algorithm using posterior vectors of utterances.

7. REFERENCES

[1] W. Reichl and W. Chou, "A unified approach of incorporating general features in decision-tree based acoustic modeling," *Proc. ICASSP*, pp. 573-576, 1999.

[2] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," *Proc. EuroSpeech*, pp. 495-498, 1999.

[3] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, J. Zheng, "The SRI March 2000 HUB-5 Conversational Speech Transcription System," *Proc. NIST Speech Transcription Workshop*, 2000.

[4] D. Povey, S. M. Chu, and B. Varadarajan, "Universal Background Model Based Speech Recognition," *Proc. ICASSP*, pp. 4561-4564, 2008.

[5] G .Cook and T. Robinson, "Boosting the performance of connectionist large vocabulary speech recognition," *Proc. ICSLP*, pp. 1305-1308, 1996.

[6] Y. Tsao, C. H. Lee, "An ensemble speaker and speaking environment modeling approach to robust speech recognition," *IEEE Trans., Audio, Speech, and Language Processing*, Vol. 17, No. 5, pp. 1025-1037, 2009.

[7] J. Xue and Y. Zhao, "Random forests of phonetic decision trees for acoustic modeling in conversational speech recognition," *IEEE Trans., Audio, Speech, and Language Processing*, Vol. 16, No. 3, pp. 519-528, 2008.

[8] F. Beaufays, V. Vanhoucke, B. Strope, "Unsupervised discovery and training of maximally dissimilar cluster models," *Proc. Interspeech*, pp. 66-69, 2010.

[9] C. Chelba, J. Schalkwyk, T. Brants, V. Ha, B. Harb, W. Neveitt, C. Parada, and P. Xu, "Query language modeling for voice search," *Proc. 2010 IEEE Workshop on Spoken Language Technology*, 2010.

[10] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition," *IEEE Trans. Audio, Speech, and Language Processing*, 2011.

[11] H. Soltau, G. Saon, B. Kingsbury, H. K. J. Kuo, L. Mangu, D. Povey, A. Emami, "Advances in Arabic speech transcription at IBM Under the DARPA GALE program," *IEEE Trans. Audio, Speech and Language processing*, Vol. 17, No. 5, pp. 884-894, 2009.

[12] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans., Speech and Audio Processing*, Vol. 7, No. 3, pp. 272-281, 1999.