A Back-off Discriminative Acoustic Model for Automatic Speech Recognition

Hung-An Chang, James R. Glass

MIT Computer Science and Artificial Intelligence Laboratory,
32 Vassar Street, Cambridge, MA 02139, USA
{hung_an, glass}@csail.mit.edu

Abstract

In this paper we propose a back-off discriminative acoustic model for Automatic Speech Recognition (ASR). We use a set of broad phonetic classes to divide the classification problem originating from context-dependent modeling into a set of sub-problems. By appropriately combining the scores from classifiers designed for the sub-problems, we can guarantee that the back-off acoustic score for different context-dependent units will be different. The back-off model can be combined with discriminative training algorithms to further improve the performance. Experimental results on a large vocabulary lecture transcription task show that the proposed back-off discriminative acoustic model has more than a 2.0% absolute word error rate reduction compared to clustering-based acoustic model.

Index Terms: context-dependent acoustic modeling, back-off acoustic models, discriminative training,

1. Introduction

Over the years, context-dependent acoustic modeling has been shown to be effective for Large Vocabulary Continuous Speech Recognition (LVCSR) tasks. Considering nearby contexts of phonetic units can provide acoustic level constraints for the speech recognizer, and thus can potentially increase the recognition accuracy. However, considering acoustic contexts can also exponentially increase the size of possible labels of the acoustic model. For example, given a set of 60 basic phonetic units, a diphone model that considers the left (or right) context of a basic phonetic unit can have $60^2 = 3,600$ possible labels; a triphone model that considers both left and right context can have $60^3 = 216,000$ possible labels. Because the size of the label set can grow very large, many context-dependent units may be of very limited occurrence (or even unseen) in the training data, resulting in significant data sparsity. To have a robust parameter estimation for context-dependent models, this data sparsity problem must be dealt with appropriately.

One common way to deal with the data sparsity problem embedded in context-dependent modeling is clustering. By clustering a sufficient amount of similar context-dependent acoustic units together, the resulting cluster can have enough training data for robust parameter estimation. Such clustering can be applied to phone-level or state-level acoustic units [1], and the clustering is generally guided by a decision tree constructed using acoustic phonetic knowledge [2]. Although the clustering approach solves the data sparsity problem, it has an intrinsic disadvantage; that is, if two acoustic units are clustered together, the scores returned by the acoustic model will always be the same. As a result, the clustered units become acoustically identical to the speech recognizer, and it has to rely on other constraints such as lexicon or language models to identify the clustered acoustic units.

Another way of dealing with the data sparsity problem is to reduce the problem of modeling a long context-dependent acoustic unit into a problem of modeling a composition of a set of shorter units. One example of this reduction-based approach is the quasi-triphone modeling proposed in [3], where a Hidden Markov Model (HMM) for a triphone is decomposed into a left context sensitive diphone state at the beginning, several context independent states in the middle, and a right context sensitive diphone state at the end. Another way of decomposing the triphones is using the Bayesian approach proposed in [4]. In the Bayesian approach, the left context of a triphone is assumed to be independent of the right context given the center phonetic unit of the triphone. Under such an assumption, the probability of a triphone can be represented by a product of two diphone probabilities divided by the monophone probability of the center unit. By using an appropriate reduction, the data sparsity problem can be solved while keeping the context-dependent units acoustically distinguishable to the recognizer. However, a pure reduction-based approach does not consider the fact that if a context-dependent unit u has a sufficient amount of occurrences in the training data, it would be beneficial to incorporate a sub-model that directly models the occurrences of u into the acoustic model.

In this paper, we propose a back-off acoustic modeling that utilizes a set of broad phonetic classes derived from acoustic-phonetic knowledge. By using the broad phonetic classes, we can divide the classification problem originating from the context-dependent acoustic modeling into a set of sub-problems. The classification scores of the sub-problems can be combined with the score of the original problem to form a back-off acoustic score for each context-dependent unit. By using an appropriate combination, the back-off scores can be guaranteed to be distinguishable and robust against data sparsity. The back-off models can be easily combined with discriminative training methods such as Minimum Classification Error (MCE) training [5] to further improve the performance. The proposed modeling scheme is evaluated by a large vocabulary lecture transcription task on the MIT Lecture Corpus [6].

The organization of the paper is as follows. Section 2 introduces the formulation of the proposed back-off modeling scheme. Section 3 gives a brief overview of discriminative training and illustrates how to combine the model with discriminative training. Experimental results of the proposed model on the MIT Lecture Corpus are reported in Section 4, followed by some concluding remarks in Section 5.

2. Back-off Acoustic Models

This section introduces the formulation of the proposed back-off acoustic model. We first present how to construct back-off models for diphones and then we show how to generalize the idea to triphones or higher order context-dependent units.

2.1. Back-off diphone models

Given an acoustic feature vector $\mathbf{x}$, the goal of an acoustic model is to return an acoustic score $a_{\boldsymbol{\lambda}}(\mathbf{x}, p)$ for each possible label p considered by the model. One common choice for the acoustic score $a_{\boldsymbol{\lambda}}(\mathbf{x}, p)$ is the log-likelihood of a Gaussian Mixture Model (GMM). For each p, the log-likelihood can be computed by

$$l_{\boldsymbol{\lambda}}(\mathbf{x}, p) = \log(\sum_{m=1}^{M^p} w_m^p \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_m^p, \boldsymbol{\sigma}_m^p)), \quad (1)$$

where m is the index of mixture components, M^p is the total number of Gaussian mixture components of p, w_m^p is the mixture weight of the m^{th} component, and $\mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_m^p, \boldsymbol{\sigma}_m^p)$ is the multivariate Gaussian density function of $\mathbf{x}$ with respect to mean vector $\boldsymbol{\mu}_m^p$ and standard deviation $\boldsymbol{\sigma}_m^p$.

For a diphone acoustic model, a label p in (1) can generally be denoted by $\langle s_l | s_r \rangle$, where s_l and s_r are strings that represent the basic phonetic units in the left context and in the right context respectively. For example, let s_l be 'k' and s_r be 'oy', and the label $\langle$k|oy$\rangle$ can represent the acoustic context occurring in the word "*coin*". From a classification point of view, the log-likelihood $l_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_r \rangle)$ can be thought of as the model's confidence on answering "yes" to the problem "Does $\mathbf{x}$ represent a context-dependent acoustic unit with left context being s_l and right context being s_r?" However, such confidence may not be always reliable due to the potential data sparsity problem intrinsic in context-dependent modeling.

Instead of asking the above question directly, we can consider examining a pair of sub-problems that are related to broad phonetic class identifications of the left and the right context. For example, instead of directly asking the question "Does $\mathbf{x}$ represent a context-dependent unit with left context being 'k' and right context being 'oy'?", we can ask the following pair of sub-problems "Does $\mathbf{x}$ represent a context-dependent unit with left context being 'k' and right context being a vowel?" and "Does $\mathbf{x}$ represent a context-dependent unit with left context being a stop consonant and right context being 'oy'?" If the model answers "yes" to both of the sub-problems, we can still identify the context-dependent unit in question as $\langle$k|oy$\rangle$. In this way, we can reduce the original classification problem into a pair of sub-problems with less context resolution. As a result, we can construct a back-off modeling scheme based on broad phonetic classes as follows.

Let $B(s)$ denote the broad phonetic class of basic phonetic unit s, where the mapping function $B(\cdot)$ can be constructed according to some acoustic phonetic properties, such as manner of pronunciation or articulation place. Given the broad phonetic class assignments, the acoustic score $a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_r \rangle)$ of a back-off model can be computed as:

$$\begin{aligned} a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_r \rangle) = \; & \omega_0 l_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_r \rangle) + \omega_l l_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | B(s_r) \rangle) \\ & + \omega_r l_{\boldsymbol{\lambda}}(\mathbf{x}, \langle B(s_l) | s_r \rangle), \end{aligned} \quad (2)$$

where $l_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | B(s_r) \rangle)$ denotes the log-likelihood of a GMM trained for all context-dependent units with the left context being s_l and the right context belonging to class $B(s_r)$, and $l_{\boldsymbol{\lambda}}(\mathbf{x}, \langle B(s_l) | s_r \rangle)$ denotes the log-likelihood of another GMM trained for all context-dependent units with the left context belonging to class $B(s_l)$ and the right context being s_r. Note that the combination weights in (2) should satisfy some constraints. For example, the three weights w_0, w_1, and w_r should sum to 1 to make the acoustic score a convex combination of log-likelihood. Also, in this work, we assume the left and the right context are contexts are equally important, so we set $w_l = w_r$.

Adjusting the weights appropriately can ensure the acoustic score in (2) is robust against data sparsity. For example, if the number of occurrences of $\langle s_l | s_r \rangle$ in the training data is smaller than a threshold, we can set the weight ω_0 to 0. Also, because any two different diphones differ at least in one of the left or right contexts, there will be at least one different term in (2) for the two diphones, making the back-off acoustic score always distinguishable to the recognizer.

2.2. Back-off triphone models

Given a feature vector $\mathbf{x}$ and a triphone label $\langle s_l | s_c | s_r \rangle$, the acoustic score $a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_c | s_r \rangle)$ can be computed by:

$$\begin{aligned} a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_c | s_r \rangle) = \; & \omega_0 l_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_c | s_r \rangle) \\ & + \omega_l a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_c | B(s_r) \rangle) \\ & + \omega_r a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle B(s_l) | s_c | s_r \rangle), \end{aligned} \quad (3)$$

where the two back-off scores $a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_c | B(s_r) \rangle)$ and $a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle B(s_l) | s_c | s_r \rangle)$ can be further decomposed similarly, as in (2). For example, $a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_c | B(s_r) \rangle)$ can be decomposed into a linear combination of three log-likelihoods $l_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_c | B(s_r) \rangle)$, $l_{\boldsymbol{\lambda}}(\mathbf{x}, \langle B(s_l) | s_c | B(s_r) \rangle)$, and $l_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | B(s_c) | B(s_r) \rangle)$.

Depending on the tasks, sometimes we may want to drop some context resolution to reduce the number of log-likelihoods needing to be evaluated. In this case we can drop the context dependencies retained by the broad phonetic classes in (3); that is, replace $a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_c | B(s_r) \rangle)$ with $a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_c | \cdot \rangle)$ and $a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle B(s_l) | s_c | s_r \rangle)$ with $a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle \cdot | s_c | s_r \rangle)$.

The back-off modeling scheme can be generalized beyond triphones. For example, $\langle s_{ll} | s_l | s_r | s_{rr} \rangle$ can be reduced by $\langle s_{ll} | s_l | B(s_r) | B(s_{rr}) \rangle$, $\langle B(s_{ll}) | B(s_l) | s_r) | s_{rr} \rangle$, and $\langle B(s_{ll}) | s_l | s_r | B(s_{rr}) \rangle$; for each of the three reduced ones, the diphone back-off scheme can be applied.

3. Discriminative back-off models

Here we show how to combine the back-off models with discriminative training. We first introduce a brief overview of discriminative training, and then we show how we can combine the back-off modeling with discriminative training algorithms.

3.1. Discriminative training

Discriminative training methods seek model parameters that can reduce the confusions in the training data made by the model. In general, an objective function that reflects the degree of confusion is constructed, and the optimal parameter setting is obtained by minimizing the objective function over the parameter space. One commonly used discriminative training method is Minimum Classification Error (MCE) training [5].

The goal of MCE training is to seek model parameters $\boldsymbol{\lambda}$ that can minimize the number of incorrectly classified utterances. Given the acoustic feature vectors $\mathbf{X}_n$ of the n^{th} training utterance and a label sequence $\mathbf{S}$, let $L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S})$ be the joint log-likelihood of $\mathbf{X}_n$ and $\mathbf{S}$ computed by the recognizer. In general, $L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S})$ can be computed by summing all the acoustic, pronunciation, and language model scores related to the sequence $\mathbf{S}$. Given the correct label sequence $\mathbf{Y}_n$ for each utterance, the MCE loss function can be expressed by

$$\mathcal{N}_{err} = \sum_{n=1}^{N} \text{sign}[-L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{Y_n})) + \max_{\mathbf{S} \neq \mathbf{Y}_n} L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S}))], \quad (4)$$

where the function sign$[d]$ equals 1 for $d > 0$ and equals 0 for $d < 0$. However, because the sign function in (4) is not differentiable, a differentiable sigmoid function $\ell(d) = \frac{1}{1+\exp(-\zeta d)}$ with a positive ζ is often utilized to smooth the loss function.

In our experiments, an N-Best version of MCE training is implemented; that is, given a training set of N utterances, the loss function can be expressed by

$$\mathcal{L} = \sum_{n=1}^{N} \ell[-L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{Y}_\mathbf{n}) + \log([\tfrac{1}{C}\sum_{\mathbf{S}\in\boldsymbol{\mathcal{S}}_n} \exp(\eta L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S}))]^{\frac{1}{\eta}})], \quad (5)$$

where $\boldsymbol{\mathcal{S}}_n$ is the set of best C competing hypotheses and η is a positive constant to adjust the importance weight of the hypotheses in $\boldsymbol{\mathcal{S}}_n$. In our experiments, we set η to 1.0, which enables the hypotheses to contribute equally to the loss function. Given the loss function in (5), the model parameters can also be updated by gradient-based optimization methods such as the Quickprop algorithm [5].

3.2. Discriminative training of back-off models

To optimize a loss function $\mathcal{L}$ over the parameter space, it often requires the computation of the gradient of $\mathcal{L}$ with respect to the acoustic model parameter vector $\boldsymbol{\lambda}$. Once the gradient $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}}$ is computed, various optimization methods can be applied to find the optimal parameters. In general, $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}}$ is computed by first taking the partial derivative of $\mathcal{L}$ with respect to each acoustic score and then summing up the contribution of gradient with respect to each acoustic score:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = \sum_{n=1}^{N} \sum_{\mathbf{x}\in\mathbf{X}_n} \sum_{p} \frac{\partial \mathcal{L}}{\partial a_{\boldsymbol{\lambda}}(\mathbf{x}, p)} \frac{\partial a_{\boldsymbol{\lambda}}(\mathbf{x}, p)}{\partial \boldsymbol{\lambda}}, \quad (6)$$

where $a_{\boldsymbol{\lambda}}(\mathbf{x}, p)$ denotes the acoustic model score for feature $\mathbf{x}$ and context-dependent label p. For a dihphone back-off model, p is of the form $\langle s_l | s_r \rangle$ and $a_{\boldsymbol{\lambda}}(\mathbf{x}, p)$ can be expressed by (2). As a result, the gradient $\frac{\partial a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_r \rangle)}{\partial \boldsymbol{\lambda}}$ can be decomposed into

$$\frac{\partial a_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_r \rangle)}{\partial \boldsymbol{\lambda}} = \omega_0 \frac{\partial l_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | s_r \rangle)}{\partial \boldsymbol{\lambda}} + \omega_l \frac{\partial l_{\boldsymbol{\lambda}}(\mathbf{x}, \langle s_l | B(s_r) \rangle)}{\partial \boldsymbol{\lambda}} + \omega_r \frac{\partial l_{\boldsymbol{\lambda}}(\mathbf{x}, \langle B(s_l) | s_r \rangle)}{\partial \boldsymbol{\lambda}}. \quad (7)$$

Note that the above decomposition can be applied to triphone or higher-order context-dependent back-off models. Therefore, the discriminative training of a back-off model can be reduced to first computing the gradients with respect to all acoustic scores, as would be done in the training of clustering-based models, and then distributing the contribution of the gradient into different levels of back-off according to the back-off weights.

While maximum-likelihood training will train the two back-off terms in (2) independently, discriminative training can couple the two terms through the objective function. As a result, the two back-off terms may gradually learn to complement each other via discriminative training, and can potentially further improve the model performance.

4. Lecture transcription experiments

In this section we evaluate the performance of the proposed back-off acoustic modeling scheme on a large vocabulary lecture transcription task. We first introduce the MIT Lecture Corpus and the SUMMIT landmark-based speech recognizer used in the experiments. Then, we compare the Word Error Rate (WER) of the proposed back-off model and that of clustered-based acoustic models on the test lectures.

4.1. MIT Lecture Corpus

The MIT Lecture Corpus contains audio recordings and manual transcriptions for approximately 300 hours of MIT lectures from eight different courses and nearly 100 MITWorld seminars given on a variety of topics [6]. The audio data was recorded with omni-directional lapel microphones and was generally recorded in a classroom environment. The recordings were manually transcribed in a way that, in addition to spoken words, disfluencies such as filled pauses, false starts, and partial words are labeled. The lecture corpus is a difficult data set for ASR systems because it contains many disfluencies, poorly organized or ungrammatical sentences, and lecture specific key words.

Among the lectures in the corpus, a 119-hour training set that includes 7 lectures from 4 courses and 99 lectures from 4 years of MITWorld lectures covering a variety of topics is selected for the acoustic model training. Two held-out MITWorld lectures (about 2 hours) are used for model development such as deciding when to stop the discriminative training. The test lectures are composed of 8 lectures from 4 different class subjects with roughly 8 hours of audio data and 7.2K words. There is no speaker overlap between the three sets of lectures.

4.2. SUMMIT Landmark-Based Speech Recognizer

Instead of extracting feature vectors at a constant frame-rate as in conventional Hidden Markov Model (HMM) speech recognizers, the SUMMIT landmark-based speech recognizer [7] first computes a set of perceptually important time points as landmarks based on an acoustic difference measure, and extracts a feature vector around each landmark. The landmark features are computed by concatenating the average values of 14 Mel-Frequency Cepstrum Coefficients in 8 telescoping regions around each landmark (total 112 dimensions), and are reduced (and whitened) to 50 dimensions by Principal Component Analysis. For each landmark, because it can either be a transition point between two adjacent phonetic units or a time point within a phonetic unit, it is natural to represent the landmarks by a set of diphone labels to model the left and right contexts of the landmarks.

For recognition, the set of diphones are modeled by a set of GMM parameters. All the other constraints used for the recognition, including pronunciation rules, lexicon, and language models are represented by a Finite-State Transducers (FST), and speech recognition is conducted by performing path search in the FST [8].

4.3. Clustering-based Models

As in a conventional setup for LVCSR tasks, diagonal GMMs were used for acoustic modeling. In order to model background noise and speech artifacts in the lectures, we extend the size of basic phone set into 74, resulting in 5,549 diphone labels (5,476 for transition and 73 for internal). The 5,549 diphone labels used by SUMMIT were clustered into 1,871 diphone classes using a top-down decision tree clustering algorithm. For each diphone class, the model is allowed to add one Gaussian mixture component per 50 training exemplars until the maximum number of mixture components is achieved. To investigate how the number of parameters affects the WER, we trained 3 sets of initial models, ML-C_{30}, ML-C_{60}, and ML-C_{100}, using Maximum-Likelihood (ML) criterion with the maximum number of mixture components set to 30, 60, and 100, respectively. The MCE training criterion in (5) was applied to the ML models, and three

Models	#Mixtures	WER Dev.	WER Test.
ML-C_{30}	31,873	45.7%	38.2%
ML-C_{60}	49,470	44.2%	36.8%
ML-C_{100}	65,340	43.8%	36.5%
MCE-C_{30}	31,873	41.2%	32.8%
MCE-C_{60}	49,470	41.5%	32.7%
MCE-C_{100}	65,340	41.5%	33.3%

Table 1: Size of the clustering-based models and their word error rates of the development and test lectures.

discriminative models, MCE-C_{30}, MCE-C_{60}, and MCE-C_{100} were generated. During the MCE training, the number of competing hypotheses is set to 20 for each utterance in the training.

Table 1 summarizes the WERs of the clustering-based models. Comparing the WERs of ML models with that of MCE models, we can see that the MCE training provides significant WER reductions. However, while the performance of ML models improves as the number of mixture components increases, the discriminatively trained models suffer some degradation in WERs. This fact suggests that the discriminatively trained models may start over-fitting the training data during the training and thus the gain resulted from the reduction of the loss function does not translate to unseen data.

4.4. Back-off models

To construct a back-off acoustic model, we first trained a ML-A_{30} (setting maximum number of mixture components to 30) diphone model without clustering the diphones. Among the 5,549 diphones, 3,653 diphones are unseen in the training lectures. For each basic phonetic unit used by SUMMIT, we classify it into one of the following broad phonetic classes: {high vowel}, {low vowel}, {retroflex}, {l}, {stop}, {fricative}, {closure}, {silence}, and {noise}. In addition to ML-A_{30}, we also trained a ML-L_{30} with all the right contexts of the diphones reduced to the broad phonetic classes, and a ML-R_{30} with all the left contexts reduced to the broad phonetic classes. Combining the three models {ML-A_{30}, ML-L_{30}, ML-R_{30}}, we can form a back-off model ML-$B_{30,30,30}$. The combination weights $(\omega_0, \omega_l, \omega_r)$ in (2) are set to $(0.5, 0.25, 0.25)$ for the diphones of more than 20 occurrences in the training data, and are set to $(0, 0.5, 0.5)$ otherwise. To investigate how the size of the back-off model affect the performance, we also set the maximum number of mixture components of the back-off to 60 and construct a ML-$B_{30,60,60}$. For both of the two back-off models, we also applied MCE training and generated two discriminative model MCE-$B_{30,60,60}$ and MCE-$B_{30,60,60}$.

Table 2 compares the performance of the back-off models with that of clustering-based models. As shown in the table, the discriminative back-off model MCE-$B_{30,30,30}$ has over 2.0% absolute WER reduction over MCE-C_{60}. The McNemar significance test [10] shows that the improvement of the back-off model over the clustering-based model is statistically significant ($p < 0.001$). Although MCE-$B_{30,60,60}$ has much more mixture components than MCE-$B_{30,30,30}$, it still yields a marginal improvement, showing that the back-off model is less sensitive to over-fitting. Also, the back-off models gain more benefit through MCE training than the clustering-based models, supporting the hypothesis that via discriminative training, the sub-models in (2) becoming more complementary to each other.

5. Conclusion and Future Work

In this paper, we have proposed a back-off discriminative acoustic modeling scheme based on broad phonetic classes. The proposed scheme works well for the diphone-based system on lecture transcription tasks. In the future, we plan to apply the back-off scheme to higher-order context-dependent models. Also, in the current setup, we only use one set of broad phonetic classes to perform the back-off. However, based on different acoustic-phonetic aspects, we can construct different sets of broad phonetic classes. For example, the phone 't' can belong to {stop consonant} based on the manner of pronunciation, but it can also belong to {alveolar voice} based on place of articulation. By incorporating different aspect of acoustic-phonetic knowledge, we can add multiple pairs of back-off terms in (2). Using multiple back-off pairs can potentially increase the difference between different context-dependent units and can thus further improve the performance of the back-off models.

Models	#Mixtures	WER Dev.	WER Test.
ML-C_{60}	49,470	44.2%	36.8%
ML-$B_{30,30,30}$	58,741	45.5%	37.8%
ML-$B_{30,60,60}$	79,116	44.0%	36.5%
MCE-C_{60}	49,470	41.5%	32.7%
MCE-$B_{30,30,30}$	58,741	40.0%	30.3%
MCE-$B_{30,60,60}$	79,116	39.3%	30.1%

Table 2: WER comparison of the clustering-based and back-off models.

6. Acknowledgement

This work is supported by Taiwan Merit Scholarship (Number NSC-095-SAF-I-564-040-TMS), by the T-Party Project, a joint research program between MIT and Quanta Computer Inc., Taiwan, and by Nippon Telegraph and Telephone Corp., Japan.

7. References

[1] Young, S. J., Odell, J. J. and Woodland P.C. "Tree-based state tying for high accuracy acoustic modeling", Proc. Human Language Technology, 307–312, 1994.

[2] Hwang, M., Huang, X. and Alleva, F.,"Predicting unseen triphones with senones", IEEE Trans. Speech and Audio Proc., 4(6): 412–419, 1996.

[3] Ljolje, "High accuracy phone recognition using context clustering and quasitriphonec models", Computer Speech Language, 8: 129–151, 1994.

[4] Ming, J., O'Boyle P., Owens, M. and Smith, F. J., "A Bayesian approach for building triphone models for continuous speech recognition", IEEE Trans. Speech and Audio Proc., 7(6): 678–683, 1999.

[5] McDermott, E., Hazen, T. J., Roux J. L., Nakamura A. and Katagiri S., "Discriminative training for large-vocabulary speech recognition using minimum classification error", IEEE Trans. Audio, Speech and Language Proc., 15(1): 1–21, 2007.

[6] Park, A., Hazen, T. J. and Glass, J. R., "Automatic processing of ausio lectures for information retrieval: vocabulary selection and language modeling", Proc. ICASSP, 497–500, 2005.

[7] Glass, J. R., "A probanilistic framework for segment-based speech recognition", Computer Speech and Language, 17: 137–152, 2003.

[8] Hetherington, L., "MIT finite-state transducer toolkit for speech and language processing", Proc. ICSLP, 2609–2612, 2004.

[9] Chang, H.-A. and Glass, J. R., "Discriminative training of hierarchical acoust models for large vocabulary continuous speech recognition", Proc. ICASSP 2009.

[10] Gillick, L. and Cox, S. J., "Some statistical issues in the comparison of speech recognition algorithms," Proc. ICASSP, 532–535, 1989.