Vocabulary and Environment Adaptation in Vocabulary-Independent Speech Recognition

Hsiao-Wuen Hon

School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Kai-Fu Lee

Speech & Language Group
Apple Computer, Inc.
Cupertino, CA 95014

1 Abstract

In this paper, we are looking into the adaptation issues of vocabulary-independent (VI) systems. Just as with speaker-adaptation in speaker-independent system, two vocabulary adaptation algorithms [5] are implemented in order to tailor the VI subword models to the target vocabulary. The first algorithm is to generate *vocabulary-adapted clustering decision trees* by focusing on relevant allophones during tree generation and reduces the VI error rate by 9%. The second algorithm, *vocabulary-bias training*, is to give the relevant allophones more prominence by assign more weight to them during Baum-Welch training of the generalized allophonic models and reduces the VI error rate by 15%. Finally, in order to overcome the degradation caused by the different acoustic environments used for VI training and testing, CDCN and ISDCN originally designed for microphone adaptation are incorporated into our VI system and both reduce the degradation of VI cross-environment recognition by 50%.

2 Introduction

In 89' and 91' DARPA Speech and Natural Language Workshops [8, 7], we have shown that accurate vocabulary-independent (VI) speech recognition is possible. However, there are many anatomical differences between tasks (vocabularies), such as the size of the vocabulary and the frequency of confusable words., which might affect the acoustic modeling techniques to achieve optimal performance in vocabulary-dependent (VD) systems. For example, whole-word models are often used in small-vocabulary tasks, while subword models must be used in large-vocabulary tasks. Moreover, within a limited vocabulary, it is possible to design some special features to separate the confusable models. Therefore, discriminative training techniques, such as neural networks [10], and maximum mutual information estimator (MMIE) [4], have so much success in small-vocabulary tasks.

Just as with *speaker adaptation* in speaker-independent systems, it is desirable to implement vocabulary adaptation to make the VI system tailored to the target vocabulary (task). Our first vocabulary adaptation algorithm is to build vocabulary-adapted allophonic clustering decision trees for the target vocabulary based on only the relevant allophones. The adapted trees would only focus on the relevant contexts to separate the relevant allophones, thus give the resulting allophonic clusters more discriminative power for the target vocabulary. In an experiment of adapting allophone clustering tree for the Resource Management task, this algorithm achieved an 9% error reduction.

Our second vocabulary adaptation algorithm is to focus on the relevant allophones during training of generalized allophonic models, instead of focusing on them during generation of allophonic clustering decision trees. To achieve that, we give the relevant allophones more prominence by assigning more weight to the relevant allophones during Baum-Welch training of generalized allophonic models. With *vocabulary-bias training* we are able to reduce the VI error rate by 15% for the Resource Management task.

We have found that different recording environments between training and testing (CMU vs. TI) will degrade the performance significantly [6], even when the same microphone is used in either case. Based on the framework of semi-continuous HMMs, we proposed to update codebook prototypes in discrete HMMs in order to fit speech vectors from new environments [5]. Moreover, *codebook-dependent cepstral normalization* (CDCN) and *interpolated SNR-dependent cepstral normalization* (ISDCN) proposed by Acero et al. [2] for microphone adaptation are incorporated into the our VI system to achieve environmental robustness. CDCN uses the speech knowledge represented in a codebook to estimate the noise and spectral equalization correction vectors for environmental normalization. In ISDCN, the SNR-dependent correction vectors are obtained via EM algorithm to minimize the VQ distortion. Both algorithms reduced the degradation of VI cross-environment recognition by 50%.

In this paper, we first describe our two vocabulary adaptation algorithms , *vocabulary-adapted decision trees* and *vocabulary-bias training*. Then we describe the codebook adaptation algorithm and two cepstral normalization techniques, CDCN and ISDCN for environmental robustness. We will also present results with these vocabulary and environment adaptation algorithms. Finally, we will close with some concluding remark about this work and future work.

3 Vocabulary Adaptation

Unlike most speaker adaptation techniques, our vocabulary adaptation algorithms only take advantage of analyzing the target vocabulary and thus do not require any additional vocabulary-specific data. Two terminologies which play an essential role in our algorithms are defined as follows.

relevant allophones Those allophones which occur in the target vocabulary (task).

irrelevant allophones Those allophone which occur in the VI training set, but not in the target vocabulary (task).

In 91' DARPA Speech and Natural Language Workshop [7], we have shown the decision-tree based generalized allophone is a adequate VI subword model. Figure 1 is an example of our VI subword unit, generalized allophone, which is actually an allophonic cluster. The allophones in the white area are relevant allophones and the rest are irrelevant ones.

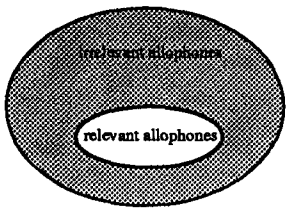

Figure 1: A generalized allophone (allophonic cluster)

3.1 Vocabulary-Adapted Decision Tree

Our first vocabulary adaptation algorithm is to change the allophone clustering (the decision trees) so that the brand new set of subword models would have a more discriminative power for the target vocabulary. Since the clustering decision tree was built on the entire VI training set, the existence of the enormous irrelevant allophones might result in sub-optimally clustering of allophones for the target vocabulary.

To reveal such facts, let's look at the following scenario. Figure 2 is a split in the original decision tree for phone /k/ generated from vocabulary-independent training set and the associated question for this split is *"Is the left context a vowel"*. Suppose all the left contexts for phone /k/ in the target vocabulary are vowels. Thus, the question for this split is totally unsuitable for the target vocabulary because the split assigns all the allophones for /k/ in the target vocabulary to one branch and discrimination among those allophones becomes impossible.

On the other hand, if only the relevant allophones are considered for this split, the associated split question would turns out to be the one of relevant questions which separates the relevant allophones appropriately and therefore possesses the greatest discriminative ability among the relevant allophones. Figure 3 just shows such optimal split for relevant allophones. The generation of the clustering decision trees are recursive. The existence of enormous irrelevant allophones would prevent the generation of the decision trees from concentrating on those relevant allophones and relevant questions, and results in sub-optimal trees for those relevant allophones.

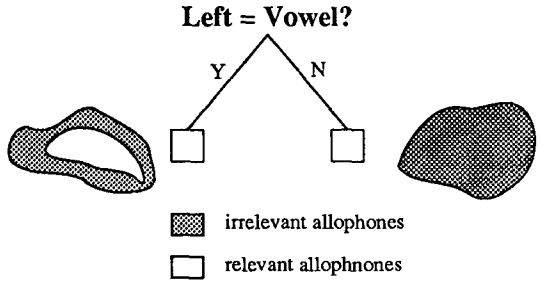

Figure 2: An split(question) in the original decision tree for phone /k/

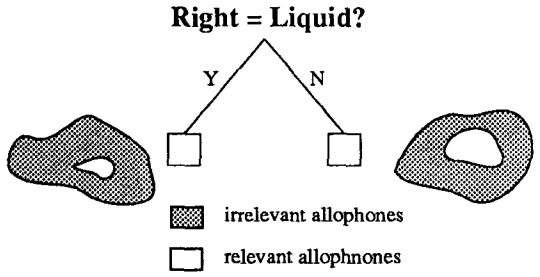

Figure 3: the correspondent optimal split(question) for relevant allophones of phone /k/

Based on the analysis, our first adaptation algorithm is to build vocabulary-adapted (VA) decision trees by using only relevant allophones during the generation of decision trees. The adapted trees would not only be automatically generated, but also focus on the relevant questions to separate the relevant allophones, therefore give the resulting allophonic clusters more discriminative power for the target vocabulary.

Three potential problems are brought up when one examining the algorithm closely. First of all, some relevant allophones might not occur in the VI training set since we can't expect 100% allophone coverage for every task, especially for large-vocabulary task. Nevertheless, it is essential to have all the models for relevant allophones ready before generating the VA decision trees because we need the entropy information of models for each split. It is trivial for those relevant allophones which also occur in VI training set. The correspondent allophonic models trained from the training data can be

used directly. Because of the nature of decision trees, every allophone could find its closest generalized allophonic cluster by traversing the decision trees. Therefore, the correspondent generalized allophonic models could be used as the models for those relevant allophones not occurring in the VI training set during the generation of the VA clustering trees.

Secondly, if only the part of VI training set which contains the relevant allophones is used to train new generalized allophonic models, the new adapted generalized allophonic models would be under-trained and less robust. Fortunately, we can retain the entire training set because of the the nature of decision trees. All the allophones could find their generalized allophonic clusters by traversing the new VA decision trees, so the entire VI training set could actually contribute to the training of new adapted generalized allophonic models and make them well-trained and robust.

The entropy criterion for splitting during the generation of decision trees is weighted by the counts (frequencies) of allophones [6]. By preferring to split nodes with large counts (allophones appearing frequently), the counts of the allophonic cluster will become more balanced and the final generalized allophonic models will be equally trainable. Since the VA decision tress are generated from the set of relevant allophones which is not the same as the set of allophones to train the generalized allophonic models. The balance feature of those models will be no longer valid. Some generalized allophonic models might only have few (or even none) examples in the VI training set and thus cannot be well-trained. Fortunately, we can enhance the trainability of VA subword models through gross validation with the entire VI training set. The gross validation for VA decision trees is somehow different than the conventional cross validation which uses one part of the data to grow the trees and the other part of independent data to prune the trees in order to predict new contexts. Since relevant allophones is already only a small portion of the entire VI training set, further dividing it will prevent the learning algorithm from generating reliable VA decision trees. Instead, we grow the VA decision trees very deeply; replace the entropy reduction information of each split by traversing through the trees with all the allophones (including irrelevant ones); and finally prune the trees based on the new entropy information. This will prune out those splits of nodes without enough training support (too few examples) even though they might be relevant to the target vocabulary. Therefore the resulting generalized allophonic models will become more trainable.

The vocabulary-adapted decision tree learning algorithm, emphasizing the relevant allophones during growing of the decision trees and using the gross validation with the entire VI training set provides an ideal mean for finding the equilibrium between adaptability for the target vocabulary and trainability with the VI training database.

3.2 Vocabulary-Bias Training

While the above adaptation algorithm tailors the subword units to the target vocabulary by focusing on the relevant allophones during the generation of clustering decision trees, it treated relevant and other irrelevant allophones equally in the final training of generalized allophonic models. Our next adaptation algorithm is to give the relevant allophones more prominence during the training of generalized allophonic models.

Since the VI training database is supposed to be very large, it is reasonable to assume that the irrelevant allophones are the majority of almost every cluster. Thus, the resulting allophonic cluster will more likely represent the acoustic behavior of the set of irrelevant allophones, instead of the set of relevant allophones.

In order to make relevant allophones become the majority of the allophonic cluster without incorporating new vocabulary-specific data, we must impose a bias toward the relevant allophones during training. Since our VI system is based on HMM approach, it is trivial to give the relevant allophones more prominence by assigning more weight to them during Baum-Welch training. The simplest way is to multiply a prominent weight to the parametric re-estimation equations for relevant allophones.

The prominent weight can be a pre-defined constant, like 2.0 or 3.0, or a function of some variables. However, it is better for the prominent weight to reflect the reliability of the relevant allophones toward which we imposed a bias. If a relevant allophone occur rarely in the training set, we shouldn't assign a large weight to it because the statistics of it is not reliable. On the other hand, we could assign larger weights to those relevant allophones with enough examples in the training data. In our experiments, we use a simple function based on the frequencies of relevant allophones. All the irrelevant allophones have the weight 1.0 and the weight for relevant allophones is given by the following function:

$1 + log_a(x)$ where x is the frequency of relevant allophones

a is chosen to be the minimum number of training examples to train a reasonable model in our configuration.

Imposing a bias toward the relevant allophones is similar to duplicating the training data of relevant allophones. For example, using a prominent weight of 2.0 for an training example in the Baum-Welch re-estimation is like observing the same training example twice. Therefore, our vocabulary-bias training algorithm is identical to duplicating the training examples of relevant allophones according to the weight function. Based on the same principle, this adaptation algorithm can be applied to other non-HMM systems by duplicating the training data of relevant allophones to make relevant allophones

become the majority of the training data during training. The resulting models will then be tailored to those relevant allophones.

4 Environment Adaptation

It is well known that when a system is trained and tested under different environments, the performance of recognition drops moderately [8] However, it is very likely for training and testing taking place under different environments in VI systems because the VI models can be used for any task which could happen anywhere. Even if the recording hardware remains unchanged, e.g., microphones, A/D converters, pre-amplifiers, etc, the other environmental factors, e.g. the room size, background noise, positions of microphones, reverberation from surface reflections, etc, are all out of the control realm. For example, when comparing the recording environment of Texas Instruments (TI) and Carnegie Mellon University (CMU), a few differences were observed although both used the same close-talk microphone (Sennheiser HMD-414).

- **Recording equipment** - TI and CMU used different A/D devices, filters and pre-amplifiers which might change the overall transfer function and thus generate different spectral tilts on speech signals.
- **Room** - The TI recording took place in a sound-proof room, while the CMU recording took place in a big laboratory with much background noise (mostly paper rustle, keyboard noise, and other conversations). Therefore, CMU's data tends to contain more additive noise than TI's.
- **Input level** - The CMU recording process always adjusted the amplifier's gain control for different speakers to compensate the varied sound volume of speakers. Since the sound volume of TI's female speakers tends to be much lower, TI probably didn't adjust the gain control like CMU did. Therefore, the dynamic range of CMU's data tends to be larger.

4.1 Codebook Adaptation

The speech signal processing of our VI system is based on a characterization of speech in a codebook of prototypical models [7]. Typically the performance of systems based on a codebook degrade over time as the speech signal drifts through environmental changes due to the increased distortion between the speech and the codebook.

Therefore, two possible adaptation strategies include:

1. continuously updating the codebook prototypes to fit the testing speech spectral vectors $\mathbf{x}_t$.
2. continuously transforming the testing speech spectral vectors $\mathbf{x}_t$ into normalized vectors $\mathbf{y}_i$, so that the distribution of the y_i is close to that of the training data described by the codebook prototypes.

Our first environment adaptation algorithm belongs to the first strategy, while two cepstral normalization algorithms which will be described in Section 4.2 belongs to the second strategy.

Semi-continuous HMMs (SCHMMs) or tied mixture continuous HMMs [9, 3] has been proposed to extend the discrete HMMs by replacing discrete output distributions with a combination of the original discrete output probability distributions and continuous pdf's of codebooks. SCHMMs can jointly re-estimate both the codebooks and HMM parameters to achieve an optimal codebook/model combination according to a maximum likelihood criterion during training. They have been applied to several recognition systems with improved performance over discrete HMMs [9, 3].

The codebooks of our vocabulary-independent system can be modified to optimize the probability of generating data from new environment by the vocabulary-independent HMMs according to the SCHMM framework. Let μ_i denote the mean vector of codebook index i in the original codebook, then the new vector $\overline{\mu_i}$ can be obtained from the following equation

$$\overline{\mu_i} = \frac{\sum_m (\sum_{t=1}^{T} \gamma_i^m(t) \mathbf{x}_t)}{\sum_m (\sum_{t=1}^{T} \gamma_i^m(t))} \quad (1)$$

where $\gamma_i^m(t)$ denotes the posterior probability observed the codeword i at time t using HMM m for speech vector $\mathbf{x}_t$.

Note that we did not use continuous Gassian pdf's to represent the codebooks in the Equation 1. Each mean vector of the new codebook is computed from acoustic vector $\mathbf{x}_t$ associated with corresponding posterior probability in the discrete forward-backward algorithm without involving continuous pdf computation. The new data from different environment, $\mathbf{x}_t$, can be automatically aligned with corresponding codeword in the forward-backward training procedure. If the alignment is not closely associated with the corresponding codeword in the HMM training procedure, reestimation of the corresponding codeword will then be de-weighted by the posterior probability $\gamma_i^m(t)$ accordingly in order to adjust the new codebook to fit the new data.

4.2 Cepstral Normalization

The types of environmental factors which differ in TI's and CMU's recording environments can roughly be classified into two complementary categories :

1. additive noise - noise from different sources, like paper rustle, keyboard noise, other conversations, etc.

2. spectral equalization - distortions from the convolution of the speech signal with an unknown channel, like positions of microphones, reverberation from surface reflections, etc.

Acero *at al.* [1, 2] proposed a series of environment normalization algorithms based on joint compensation for additive noise and equalization. They has been implemented successfully on SPHINX to achieve robustness to different microphones. Among those algorithms, *codeword-dependent cepstral normalization* (CDCN), is the most accurate one, while *interpolated SNR-dependent cepstral normalization* (ISDCN) is the most efficient one[1]. In this study, we incorporate these two algorithms to make our vocabulary-independent system more robust to environmental variations.

$$\mathbf{x} = \mathbf{z} - \mathbf{w}(\mathbf{q}, \mathbf{n}) \quad (2)$$

Equation 2 is the environmental compensation model, where $\mathbf{x}$, $\mathbf{z}$, $\mathbf{w}$, $\mathbf{q}$ and $\mathbf{n}$ represent respectively the normalized vector, observed vector, correction vector, spectral equalization vector and noise vector. The CDCN algorithm attempts to determine $\mathbf{q}$ and $\mathbf{n}$ that provide an ensemble of compensated vectors $\mathbf{x}$ being collectively closest to the set of locations of legitimate VQ codewords. The correction vector $\mathbf{w}$ will be obtained using MMSE estimator based on $\mathbf{q}$, $\mathbf{n}$ and the codebook. In ISDCN, $\mathbf{q}$ and $\mathbf{n}$ were determined by an EM algorithm aiming at minimizing VQ distortion. The final correction vector $\mathbf{w}$ also depends on the instantaneous SNR of the current input frame using a sigmoid function.

5 Experiments and Results

All the experiments are evaluated on the speaker-independent DARPA resource management task. This task is a 991-word continuous task and a standard word-pair grammar with perplexity 60 was used throughout. The test set, **TI-TEST**, consists of 320 sentences from 32 speakers (a random selection from June 1988, February 1989 and October 1990 DARPA evaluation sets).

In order to isolate the influence of cross-environment recognition, another identical same test set, **CMU-TEST**, from 32 speakers (different from TI speakers) was collected at CMU. Our baseline is using 4-codebook discrete SPHINX and decision-tree based generalized allophones as the VI subword units[7]. Table 1 shows that about 9% error reduction is achieved by adapting the decision trees for Resource Management task, while about 15% error reduction is achieved by using vocabulary-bias training for the same task. Nevertheless, when we try to combine these two adaptation algorithms to further tailor the vocabulary-independent models to the Resource Management task, no compound improvement was produced. It might be because either both algorithms are learning the similar characteristics of the target task, or the combination of these two algorithms already reaches the limitation of adaptation capability within our modeling technique without the help of vocabulary-specific data.

[1]The reader is referred to [1] for detailed CDCN and ISDCN algorithms

Condition	Error Rate	Error Reduction
Baseline	5.4%	N/A%
+VA decision trees	4.9%	9.3%
+VB training	4.6%	14.8%
+VA trees & VB training	4.6%	14.8%

Table 1: The results for Resource Management using vocabulary-adapted decision trees and vocabulary-bias training algorithms

Adaptation Sentence	CMU-TEST	TI-TEST
Baseline	5.4%	7.4%
100	N/A	7.1%
300	N/A	7.0%
1000	N/A	7.0%
2000	N/A	6.9%

Table 2: The vocabulary-independent results on TI-TEST by adapting the codebooks for TI's data

In codebook adaptation experiments, the 4 codebooks used in our HMM-based system are updated according Equation 1. We randomly select 100, 300, 1000, 2000 sentences from TIRM database to form different adaptation sets. Two iteration were carried out for each adaptation sets to estimated the new codebooks for TI's data, while the HMM parameters are fixed. Table 2 shows the adaptation recognition result on TI testing set. It is indicated that only marginal improvement by adapting codebook for new environment even with lots of adaptation data. The result suggested that the adaptation of codebook alone fail to produce adequate adaptation because the HMM statistics used by recognizer have not been updated.

Table 3 shows the recognition error rate on two test sets for VI systems incorporated with CDCN and ISDCN. Be aware that our VI training set was recorded at CMU. The degradation of cross-environment recognition with **TI-TEST** is roughly reduced by 50%. Like most environment normalization algorithms, there is also a minor performance degradation for same-environment recognition when gaining robustness to other environments.

Test Set	CMU-TEST	TI-TEST
Baseline	5.4%	7.4%
CDCN	5.6%	6.4%
ISDCN	5.7%	6.5%

Table 3: The results for environment normalization using CDCN & ISDCN

6 Conclusions

In this paper, we have presented two vocabulary adaptation algorithms, including *vocabulary-adapted decision trees* and *vocabulary-bias training*, that improve the performance of the vocabulary-independent system on the target task by tailoring the VI subword models to he target vocabulary. In 91' DARPA Speech and Natural Language Workshop [7], we have shown that our VI system is already slightly better than our VD system. With these two adaptation algorithms which led to 9% and 15% error reduction respectively on Resource Management task, the resulting VI system is far more accurate than our VD system. In [8], we have demonstrated improved vocabulary-independent results with vocabulary-specific adaptation data. In the future, we plan to extend our adaptation algorithms with the help of vocabulary-specific data to achieve further adaptation with the target vocabulary (task).

CDCN and ISDCN have been successfully incorporated to the vocabulary-independent system and reduce the degradation of VI cross-environment recognition by 50%. In the future, we will keep investigating new environment normalization techniques to further reduce the degradation and ultimately achieve the full environmental robustness across different acoustic environments. Moreover, environment adaptation with environment-specific data will also be explored for adapting the VI system to the new environment once we have more knowledge about it.

To make the speech recognition system more robust for new vocabularies and new environments is essential to make the speech recognition application feasible. Our results have shown that plentiful training data, careful subword modeling (decision-tree based generalized allophones) and suitable environment normalization have compensated for the lack of vocabulary and environment specific training. With the additional help of vocabulary adaptation, the vocabulary-independent system can be further tailored to any task quickly and cheaply, and therefore facilitates speech applications tremendously.

Acknowledgements

This research was sponsored by the Defense Advanced Research Projects Agency (DOD), Arpa Order No. 5167, under contract number N00039-85-C-0163. The authors would like to express their gratitude to Professor Raj Reddy and CMU speech research group for their support.

References

[1] Acero, A. *Acoustical and Environmental Robustness in Automatic Speech Recognition.* Department of Electrical Engineering, Carnegie-Mellon University, September 1990.

[2] Acero, A. and Stern, R. *Environmental Robustness in Automatic Speech Recognition.* in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1990, pp. 849–852.

[3] Bellegarda, J. and Nahamoo, D. *Tied Mixture Continuous Parameter Models for Large Vocabulary Isolated Speech Recognition.* in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1989, pp. 13–16.

[4] Brown, P. *The Acoustic-Modeling Problem in Automatic Speech Recognition.* Computer Science Department, Carnegie Mellon University, May 1987.

[5] Hon, H. *Vocabulary-Independent Speech Recognition: : The VOCIND System.* School of Computer Science, Carnegie Mellon University, February 1992.

[6] Hon, H. and Lee, K. *CMU Robust Vocabulary-Independent Speech Recognition System.* in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. Toronto, Ontario, CANADA, 1991, pp. 889–892.

[7] Hon, H. and Lee, K. *Recent Progress in Robust Vocabulary-Independent Speech Recognition.* in: **DARPA Speech and Language Workshop**. Morgan Kaufmann Publishers, Asilomar, CA, 1991.

[8] Hon, H. and Lee, K. *Towards Speech Recognition Without Vocabulary-Specific Training.* in: **DARPA Speech and Language Workshop**. Morgan Kaufmann Publishers, Cape Cod, MA, 1989.

[9] Huang, X., Lee, K., and Hon, H. *On Semi-Continuous Hidden Markov Modeling.* in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. Albuquerque, NM, 1990, pp. 689–692.

[10] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. *Phoneme Recognition using Time-Delay Neural Networks.* **IEEE Transactions on Acoustics, Speech, and Signal Processing**, vol. ASSP-28 (1989), pp. 357–366.