TOWARDS SPEECH RECOGNITION WITHOUT VOCABULARY-SPECIFIC TRAINING

Hsiao-Wuen Hon, Kai-Fu Lee, Robert Weide
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

With the emergence of high-performance *speaker-independent* systems, a great barrier to man-machine interface has been overcome. This work describes our next step to improve the usability of speech recognizers—the use of *vocabulary-independent* (VI) models. If successful, VI models are trained once and for all. They will completely eliminate task-specific training, and will enable rapid configuration of speech recognizers for new vocabularies. Our initial results using *generalized triphones* as VI models show that with more training data and more detailed modeling, the error rate of VI models can be reduced substantially. For example, the error rates for VI models with 5,000, 10,000 and 15,000 training sentences are 23.9%, 15.2% and 13.3% respectively. Moreover, if task-specific training data were available, we can interpolate them with VI models. Our prelimenary results show that this interpolation can lead to an 18% error rate reduction over task-specific models.

1. Introduction

One of the most exciting and promising areas of speech research is large-vocabulary continuous speech recognition. A myriad of applications await a good speech recognizer. However, while many reasonable recognizers exist today, they are impractical and inflexible due to the tedious process of configuring a recognizer. This tedium is typically embodied in one of the following forms:

- **Speaker-specific training**: each speaker must speak for about an hour to *train* the system.
- **Vocabulary-specific training**: with each new vocabulary comes the dilemma of tedious retraining for optimal performance, or tolerating substantially higher error rate.
- **Training time**: with each new speaker or vocabulary, many hours are needed to process the speech and train the system.

Recent work at Carnegie Mellon [1, 2] and several other laboratories has shown that highly accurate speaker-independent speech recognition is possible, thus alleviating the need for speaker-dependent training. However, these *speaker-independent* systems still need *vocabulary-dependent* training on a large population of speakers for each vocabulary, which requires a very large amount of time for data collection (weeks to months), dictionary generation (days to weeks), and processing (hours to days).

As speech recognition flourishes and new applications emerge, the demand for vocabulary-specific training will become the bottleneck in building speech recognizers. In this paper, we will discuss our initial work to alleviate the tedious vocabulary-specific training process.

Our work thus far has involved collecting and processing a large *general English* database, and evaluating the *generalized triphone* [3, 2] as a vocabulary independent unit. We collected and trained generalized triphone models on up to 15,000 training sentences, and compared our results to that from vocabulary-dependent models. We found that as we increased VI training data, VI generalized triphones improved from 109% more errors than vocabulary-dependent training to only 16% more errors. In another vocabulary-adaptation experiment, we found that interpolating vocabulary-dependent models with vocabulary-independent models reduces the error rate by 18%.

Based on the enouraging results of this preliminary study, we conjecture that generalized triphones are a reasonable starting point in our search for a more vocabulary-independent subword unit. In the future, we hope to further increase our training database. With increased training data will come the ability to train more detailed subword models. We expect that this combination will enable us to further improve our results.

In this paper, we will first discuss the issue of VI models. Next, we will briefly describe generalized triphones. Then, we will describe our databases and experimental results. Finally, we will close with some concluding remarks about this and future work.

2. Vocabulary-Independent Subword Modeling

Subword modeling has become an increasingly more important issue because as the vocabulary capacity of recognizers increases, it becomes difficult, if not impossible, to train whole-word models. Many subword modeling techniques have been proposed (see [4] for a survey on these techniques). However, most subword models were evaluated using the same vocabulary for training and testing. An important question that has often been ignored is: *how well will these subword models perform under vocabulary-independent conditions?* In other words, if we train on one vocabulary and test on another, will the performance degrade considerably? If so, it will then be necessary to retrain for each new vocabulary, which is time-consuming, tedious, and costly.

Why should performance degrade across vocabularies? There are two main causes: the lack of coverage and the inconsistency of the models. The coverage problem is caused

by the fact that the phonetic events in the testing vocabulary are not covered by the training vocabulary. This lack of coverage makes it impossible to train the models needed for the testing vocabulary. Instead, we must improvise with a more general model. For example, if the phone /t/ in the triphone context of /s t r/ occurs in testing but not in training, it will be necessary to use a more general model, such as /t/ in the context of /s t/, or /t r/, or even a context-independent /t/.

The problem of improvising with general models is that they may become *inconsistent*. That is, the same model may generate many different realizations. For example, if context-independent phone models are used, the same model for /t/ must capture various events, such as flapping, unreleased stop, and realizations in /t s/ and /t r/. Then, if /t s/ is the only context in which /t/ occurs in the training, while /t r/ is the only context in the testing, the model used will be highly inappropriate.

To ensure that the models are consistent and that new contexts are covered, it is necessary to account for all causes of phonetic variability. However, the enumeration of all the causes* will lead to an astronomical number of models. This makes the models *untrainable*, which renders them powerless.

In view of the above analysis, we believe that a successful approach to vocabulary-independent subword modeling must use models that are consistent, trainable, and generalizable. Consistency means the variabilities within a model should be minimized; trainability means there should be sufficient training data for each model; and generalizability means reasonable models for the testing vocabulary can be used in spite of the lack of precise coverage in the training.

3. Generalized Triphones

In this section, we describe the basis of our current work — generalized triphone models, which are based on triphone models [5]. Triphones account for the left and right phonetic contexts by creating a different model for each possible context pair. Since the left and right phonetic contexts are the most important factors that affect the realization of a phone, triphone models are a powerful technique and have led to good results. However, there are a very large number of triphones, which can only be sparsely trained. Moreover, they do not take into account the similarity of certain phones in their effect on other phones (such as /b/ and /p/ on vowels).

In view of this, we introduce *generalized triphone models* [3]. Generalized triphones are created from triphone models using a clustering procedure that combines triphone HMMs according to a maximum likelihood criterion. In other words, we want to cluster triphones into a set of generalized triphones that will have as high as possible a probability of generating the training data. This is consistent with the maximum-likelihood criterion used in the forward-backward algorithm.

Context generalization provides the ideal means for finding the equilibrium between trainability and consistency. Given a fixed amount of training data, it is possible to find the largest number of trainable models that are consistent. Moreover, it is easy to incorporate other causes of variability such as stress, syllable position, and word position.

One flaw with bottom-up clustering approaches to context generalization is that there is no easy way of generalizing to contexts that have not been observed before. Indeed, in a pilot experiment, we found that generalized triphones trained on the resource management task performed poorly on a new voice calculator vocabulary. We believe this was mainly due to the fact that 36.3% of the triphones in the testing vocabulary were not covered, and context-independent phones had to be used.

In order to overcome these problems, we need a much larger database that has a better coverage of the triphones that are more vocabulary-independent. To that end, we are currently collecting a *general English database*. Our first step is to use this database to train triphone and generalized triphone models, and then evaluate them on new vocabularies. As this database grows, more triphone-based models can be adequately trained. Eventually, we will be able to model other acoustic-phonetic detail such as stress, syllable position, between-word phenomena, and units larger than triphones.

4. Databases

Training : The General English Database

In order to train VI models, we need a very large training database that covers all English phonetic variations. Fortunately, because our focus is speaker-*independent* recognition, this database can be collected incrementally without creating an unreasonable burden on any speaker. Initially, this database is a combination of four sub-databases, which we will describe below. Two of the databases were recorded at Texas Instruments in a soundproof booth, and the other two were collected at Carnegie Mellon in an office environment. The same microphone and processing were used for all four sub-databases. The ratio of male to female speakers is about two to one in all four sub-databases.

The first database is the 991-word **resource management** database [6], which was designed for inquiry of naval resources. For this study, we used a total of 4690 sentences from the 80 training and the 40 development test speakers.

The **TIMIT** database [7] consists of 630 speakers, each saying 10 sentences. We used a subset of this database, including a total of 420 speakers and 3300 sentences. There are total of 4900 different words.

The **Harvard** database consists of 108 speakers each say-

*A partial list might include: phonetic contexts, articulator position, stress, semantics, prosody, intonation, dialect, accent, loudness, speaking-rate, speaker anatomy, etc.

ing 20 sentences for a total of 2160 sentences. There are 1900 different words.

The **General English** database consists of 250 speakers each saying 20 sentences for a total 5000 sentences. It covers about 10000 different words.

Testing: The Voice Calculator Database

An independent task and vocabulary was created to test the VI models. This task deals with the operation of a calculator by voice. There are 122 words, including the alphabet and numbers, which are highly confusable. We used 1000 sentences from 100 speakers to train vocabulary-dependent models and 90 sentences from 10 speakers to test various systems under a word-pair grammar with perplexity 53.

5. Experiments and Results

We used a version of SPHINX for the experiments on our VI models. Since SPHINX is described elsewhere in these proceedings [8], we will not be repetitive here. We note, however, that between-word triphone models [9] and corrective training [10] were not used in this study. More detailed descriptions of SPHINX can be found in [1, 2].

We used 90 sentences from 10 speakers from the voice calculator task for evaluation. The following training sets were used:

VI-5000 Approximately 5000 sentences from resource management. The triphone coverage on the voice calculator task is 63.7%, and word coverage is 44.3%.

HARV-TIMIT
Approximately 5000 sentences from Harvard and TIMIT database. Triphone coverage is 91.9% and word coverage is 53.3%

GENENG Approximately 5000 sentences from general English database. Triphone coverage is 96.9% and word coverage is 65.6%

VI-10000 Approximately 10,000 sentences from resource management, TIMIT, and Harvard. Triphone coverage is 95.3%, and word coverage is 63.9%.

VI-15000 Approximately 15,000 sentences from resource management, TIMIT, Harvard, and general English. Triphone coverage is 99.2%, and word coverage is 70.5%.

VD-1000 Approximately 1000 sentences from voice calculator training. Triphone coverage is 100%, and word coverage is 100%.

Our first experiment used 48 phonetic models, trained from each of the above four training sets, and tested them on the voice calculator task. Table 1 shows the accuracy of phone models. Although phones are well-covered in each of the three VI databases, the VD results are still much better than the VI results. This is due to the fact that the voice calculator has a small vocabulary, and the VD phone models were able to model the few contexts in this vocabulary well.

Training Set	Recognition Accuracy
VI-5000	31.1%
HARV-TIMIT	25.4%
GENENG	22.9%
VI-10000	22.8%
VI-15000	21.5%
VD-1000	16.4%

Table 1: Recognition results using phonetic models, with vocabulary-independent (VI) and vocabulary-dependent (VD) training.

Next, we trained generalized triphone models on the four training databases. For each VI training set, we chose an appropriate number of generalized triphones to train from the training corpus. Then, for each phone in the voice calculator task, if the triphone context was covered, we mapped it to a generalized triphone. Otherwise, we used the corresponding context-independent phone. For vocabulary-dependent training, we felt that sufficient training was available for all triphones, so no generalization was performed, and we used VD triphone models. In all four cases, the trained model parameters were interpolated with context-independent phone models to avoid insufficient training. The results of these models are shown in Table 2. Also shown in Table 2 are the triphone and word coverage statistics using the above four training databases. Note that as training data is increased, triphone coverage improves more rapidly than word coverage. With 15,000 training sentences, almost all triphones are covered and the result is close to that from VD training with 1000 training sentences. Moreover, the result of GENENG which only conatins 5000 sentences is almost the same as that of VI-10000 which contains 10000 sentences, because the triphone coverage of GENENG is better. Therefore, to cover as many as triphone contexts is crucial for Vocabulary-Independent training.

Training Set	Word Coverage	Triphone Coverage	Recognition Accuracy
VI-5000	44.3%	63.7%	23.9%
HARV-TIMIT	53.3%	91.9%	16.3%
GENENG	65.6%	96.6%	15.1%
VI-10000	63.9%	95.3%	15.2%
VI-15000	70.5%	99.2%	13.3%
VD-1000	100%	100%	11.4%

Table 2: Recognition results using generalized triphones with vocabulary-independent (VI) and vocabulary-dependent (VD) training.

The final experiment involves the combination of the VI

and the VD models. Assuming that we have a set of VI models trained from a large training database, and a vocabulary-dependent training set, we use the following algorithm to utilize both training sets:

1. **Initialization** - Use the VI models to initialize VD training. As before, for each phone in the voice calculator task, if the triphone is covered, then it is used to initialize that triphone. Otherwise, the corresponding context-independent phone is used.
2. **Training** - Run the forward-backward algorithm on the VD sentences to train a set of VD models.
3. **Interpolation** - Use deleted interpolation [11, 1] to combine the appropriate task-specific VD models with the robust task-independent VI models.

Table 3 shows results using the above interpolation algorithm. We found that the combination of the VI models from 15,000 sentences and the VD models from 1000 sentences can reduce the error rate by 18% over VD training alone. This algorithm can be used to improve any task-dependent recognizer given a set of VI models. Also, these results show that vocabulary-adaptation is promising.

Training Set	Recognition Accuracy	Error Reduction
VD-1000	11.4%	-----
VI-5000 & VD-1000	10.3%	9.7%
VI-10000 & VD-1000	9.5%	16.7%
VI-15000 & VD-1000	9.3%	18.4%

Table 3: Recognition results of vocabulary-dependent models interpolated with vocabulary-independent models.

We have begun to experiment without grammar; however, at the time of this writing, the results with VI models are not as good relative to the VD models.

6. Conclusion and Future Work

This paper addressed the issue of vocabulary-independent subword modeling. Vocabulary independence is important because the overhead of vocabulary-dependent training is very high. Yet, vocabulary-independent subword models must be consistent, trainable, and generalizable. We believe this requires a large training database and a set of flexible subword units. To this end, we have collected a large multi-speaker database, from which we trained generalized triphone models. We found that with sufficient training, over 99% triphone coverage of the testing vocabulary can be attained. We report a vocabulary-dependent word accuracy of 88.6%, while the best vocabulary-independent models led to 86.7%. In another experiment, we found that it is possible to reduce the vocabulary-dependent error rate by 18% (to 90.7%) by interpolating the vocabulary-dependent models with the vocabulary-independent ones.

These results are very encouraging. In the future, we hope to further enlarge our multi-speaker database. As this database grows, we hope to model other acoustic-phonetic detail such as stress, syllable position, between-word phenomena, and units larger than triphones. To reduce the large number of resultant models, we will first use phonetic knowledge to identify the relevant ones, and then apply the clustering technique used in generalized triphones to reduce these detailed phonetic units into a set of *generalized allophones*. We will also experiment with top-down clustering of allophones. While the top-down approach may lead to less "optimal" clusters, it has more potential for generalization in spite of poor coverage.

The choice of speaker-independence gives us the luxury of plentiful training. We believe that the combination of knowledge and subword clustering will lead to subword models that are consistent, trainable, and generalizable. We hope that plentiful training, careful selection of contexts, and automatic clustering can compensate for the lack of vocabulary-specific training.

Acknowledgments

The authors wish to thank the members of the Carnegie Mellon Speech Group for their contributions. We would also like to acknowledge US West and DARPA for their support.

References

1. Lee, K.F., *Automatic Speech Recognition: The Development of the SPHINX System,* Kluwer Academic Publishers, Boston, 1989.
2. Lee, K.F., Hon. H.W., Reddy, R., "An Overview of the SPHINX Speech Recognition System", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, January 1990.
3. Lee. K.F., Hon, H.W., Hwang, M.Y., Mahajan, S., Reddy, R., "The SPHINX Speech Recognition System", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1989.
4. Lee, K.F., "Context-Dependent Phonetic Hidden Markov Models for Continuous Speech Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, April 1990.
5. Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M., Makhoul, J., "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1985.
6. Price, P.J., Fisher, W., Bernstein, J., Pallett, D., "A Database for Continuous Speech Recognition in a 1000-Word Domain", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1988.

7. Fisher, W.M., Zue, V., Bernstein, J., Pallett, D., "An Acoustic-Phonetic Data Base", *113th Meeting of the Acoustical Society of America*, May 1987.

8. Lee, K.F., "Hidden Markov Models : Past, Present, and Future", *Proceedings of Eurospeech*, September 1989.

9. Hwang, M.Y., Hon, H.W., Lee, K.F., "Modeling Between-Word Coarticulation in Continuous Speech Recognition", *Proceedings of Eurospeech*, September 1989.

10. Lee, K.F., Mahajan, S., "Corrective and Reinforcement Learning for Speaker-Independent Continuous Speech Recognition", *Proceedings of Eurospeech*, September 1989.

11. Jelinek, F., Mercer, R.L., "Interpolated Estimation of Markov Source Parameters from Sparse Data", in *Pattern Recognition in Practice,* E.S. Gelsema and L.N. Kanal, ed., North-Holland Publishing Company, Amsterdam, the Netherlands, 1980, pp. 381-397.