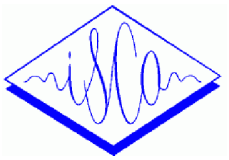

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

A Study on Using Word-Level HMMs to Improve ASR Performance over State-of-the-Art Phone-Level Acoustic Modeling for LVCSR

I-Fan Chen and Chin-Hui Lee

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
ichen8@gatech.edu, chl@ece.gatech.edu

10.21437/Interspeech.2012-487

Abstract

In this paper, we propose word-level hidden Markov models (HMMs) to supplement state-of-the-art phone-based acoustic modeling in order to enhance the performance of automatic speech recognition (ASR) system. Each word in a vocabulary is initially modeled by well-trained triphone models. Maximum a posteriori adaptation is then applied to generate models for words with a large number of occurrences in the training set so that the acoustic distribution of the words can be modeled more precisely. Experimental results show that the proposed word-based systems outperform phone-based systems on the TIMIT task with a small training corpus. While in tasks with plenty of training data, word-based systems still show improvements over phone-based systems, such as the WSJ task. Furthermore the word-based systems have a better discriminating ability on short words and homophones. They are also more robust to language model weight variation than conventional phone-based systems.
Index Terms: word-level HMM, automatic speech recognition, detection-based ASR, language model weight, homophone

1. Introduction

Phone-level hidden Markov models (HMMs) have been adopted as fundamental speech units in most state-of-the-art large vocabulary continuous speech recognition (LVCSR) systems [1] mostly due to the fact that word-based units do not appear too often in a training set to train good acoustic models even in a context-independent scenario. Furthermore phone units provide a better flexibility than words so that a modification of the phone models is not required when the system vocabulary changes.

However, using phone-level units for acoustic modeling often implies that a phone or triphone is characterized exactly the same way in different words. This is sometimes not entirely true due to the differences in prosody and usages of these words [2]. Previous research on speech unit selection in Mandarin LVCSR systems had shown a benefit of using high-level units, such as syllables, for acoustic modeling [3]. Meanwhile, since word-level HMMs are theoretically able to capture detailed pronunciation variation inside a word, they are expected to perform better than conventional phone-level HMMs. This is important in detection-based ASR, in which acoustic evidence plays an important role in the recognition process [4].

Despite the advantages of using word units for acoustic modeling, two major issues about training data sparsity and computational cost need to be addressed. We propose the use of word HMMs for words appearing frequently in the training set, and construct these models through maximum a posteriori (MAP) adaptation [5]. Experiments and detailed analyses comparing word-based and phone-based systems on two LVCSR tasks, WSJ and TIMIT, are performed and the results show that word-based acoustic models always outperform conventional phone-based models especially for short words with fewer phones exhibiting more pronunciation variability. Word-based systems are especially more effective than phone-based systems when the language models used are not as strong, or when the language model multiplication factor is not exactly known.

2. Word-Based System Configurations

Training data sparsity is one of the major concerns for using words as acoustic units in LVCSR. By adopting MAP adaptation, we expect that word models can be reliably established if such words appears enough times in a training set; while for uncommon words the models can fall back to the triphone-based word description. Two types of word models, which have different complexity and acoustic modeling abilities, are proposed. They are described in the following.

2.1. Monoword model

In monoword systems, each entry of the word-to-phone-sequence dictionary in the LVCSR systems corresponds to a word model. The monoword models are built by concatenating triphone models according to the dictionary. Though these models are easy to construct, a drawback is the lack of context modeling ability at the word boundaries. To preserve context information, states at the beginning and the end of each monoword model are merged from all possible triphones at the word boundaries. Therefore the mixture sizes of the HMM states at the beginning and the end of a monoword will be greater than those in the states at the center of the model.

Since monoword models do not increase the branching factor in the search phase too much, they can be used in first-pass decoding. The decoding time is usually similar to the conventional phone-based systems.

2.2. Triword model

It is known that monoword models can only capture part of the context information at word boundaries. To better handle context information, implementation of a triword system is inevitable. However, building a triword system means the number of acoustic models would blow up exponentially as the vocabulary size increases. This means most of the acoustic models would have no adaptation data; and the computational cost would be dramatically larger than the phone-based systems. To limit the

model number from exponentially growing, first it is assumed that the context effects only take place on the adjacent phones. Then based on this assumption, triword models whose boundary triphones are the same can be tied together (Figure 1 (a).) Also, for each word, the center states of its triword models are tied together as well (Figure 1 (b).)

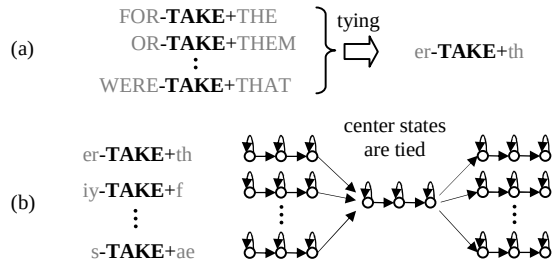

Figure 1: (a) word level tying for triword models
(b) state level tying for triword models

Due to the high branching factor, it is impractical to do first-pass decoding using triword models. A second-pass decoding is therefore used for triword systems in this paper. The adapted triword models will be used to rescore word lattices generated by the baseline phone-based system.

3. LVCSR Experiments

Two sets of LVCSR experiments are carried out to test the performance of the proposed word-based systems in different application scenarios. The first one is conducted on the TIMIT acoustic-phonetic corpus, which is a small set (about 3 hours of training data) designed mainly for phone model training and testing. The second series of experiments is executed on the WSJ0 corpus, which is a much larger English set with about 15 hours of training data for LVCSR system training and testing.

3.1. TIMIT Corpus

The TIMIT database is divided into three parts: a training set (3296 utterances, 2.79 hours), a development set (400 utterances), and a test set (1344 utterances). The training and development sets are subsets of the standard TIMIT training set, while the test set is the standard TIMIT test set. Dialect utterances (SA1 and SA2) are not used in the experiment.

The baseline system is a crossword triphone system using the CMU/MIT phone set. Each triphone is modeled by a strictly left-to-right, 3-state continuous-density hidden Markov model (CDHMM), with a Gaussian mixture density of 8 components per state, and there are a total of 415 tied states. The phone error rate for free phone decoding, which uses an unweighted phone loop as the grammar, of the baseline phone-based system is 34.41%. For word-based systems, we test the monoword system in the TIMIT task. There are 6231 entries in the TIMIT dictionary. For each entry we construct a monoword HMM using the method described in Section 2.1. Parameters are tuned using the development set; and the HTK toolkit is used for model training, adaptation and decoding [6].

Since the TIMIT corpus does not provide a language model, we train two language models using the TIMIT transcriptions by ourselves. The first one was a bigram language model trained by the transcriptions of the 3296 training utterances with Good-Turing smoothing (min 3, max 7) [7]. While the second one was a bigram language model trained by transcriptions of the 1344 test utterances without smoothing. The second language model was for an oracle test since we noticed that there is no overlap in text materials for the TIMIT training and test data; and it can be expected that the perplexity [8] of the training-data-trained language model over test data would be very high, which is different from the usual case in conventional LVCSR tasks. To simulate the usual LVCSR scenario where language model plays an important role, an oracle experiment which assumes the best language model for the TIMIT test data is available is conducted so that we can observe the performance of word-based systems in situations when a good language model is used in decoding. All language models are trained by SRILM toolkit [9]. The resulted language model perplexities on the test data are 1454 and 6.46, respectively.

Table 1 shows the experiment results on the TIMIT test set. It is clear that the word-based system significantly outperforms the baseline phone-based system in both free word decoding (6.65% absolute improvement) and LVCSR (5% absolute improvement). Even in the oracle language model case, the word model can still provide useful acoustic information to improve the word accuracy from 96.08% to 97.99% (1.91% absolute)

Table 1. Word accuracies of phone and word based systems in three different testing conditions

Language Model		Real: No LM (word loop)	Real: LM trained with the training set	Oracle: LM trained with the testing set
Phone-based System		26.33	40.61	96.08
Monoword System	No Adp	28.69	40.95	96.89
	Adapted	**32.98**	**45.61**	**97.99**

3.2. WSJ0 Corpus

In the WSJ0 task, all evaluation experiments were carried out on the speaker independent set. The standard SI-84 training set (7130 utterances, ~15 hours) was used to train the baseline triphone system and to adapt the word models of the proposed word-based systems. The standard Nov92 5K non-verbalized punctuation test set (330 utterances) and a subset of the WSJ0 5K development set (utterances in si_dt_05 that do not contain words outside the 5K vocabulary, 450 utterances) were used for performance evaluation.

The baseline phone-based system is a maximum-likelihood trained crossword triphone system (strictly left-to-right 3-state CDHMM, 8 components per state, and totally 2831 tied states.) For word-based systems, both monoword and triword systems are evaluated in the experiment. The monoword system runs first-pass decoding; while the triword system decodes by rescoring the word lattices generated by the baseline system. The non-verbalized punctuation bigram language model provided by the WSJ0 corpus was used for all decoding. Language model weight is set to 15.

Experimental results of the phone and word based systems are listed in Table 2. Unlike in the TIMIT case, the performance of the initial monoword model degrades from the baseline phone-based system significantly. Though some improvement was observed after adaptation, the performance is still worse than the baseline phone-based system. On the other hand, the initial triword system has about the same performance with the

baseline phone-based system; and the adapted triword system consistently outperforms the baseline phone-based system on both the Nov92 test set and the si_dt_05 subset.

Table 2. Word accuracy of phone and word based systems on WSJ0 5K test set and development subset.

		Nov92	si_dt_05 subset
Phone-based System		92.60	91.36
Monoword System	No Adapt	90.70	89.02
	Adapted	91.33	89.96
Triword System	No Adapt	92.60	91.33
	Adapted	**92.98**	**92.00**

The performance difference of the monoword systems in the TIMIT and WSJ0 tasks might be explained by the difference of state tying factors in the baseline phone-based systems of these two experiment sets. In the TIMIT task, the phone-based system tied most of the state in order to have better distribution estimation on the small training data. The monoword system was able to release this heavy constrain from the baseline phone-based system and thus obtained better performance on the TIMIT corpus. However in the WSJ0 task, since the state tying factor is much less than the TIMIT case the drawback of not being able to model well context information at word boundaries makes the monoword system being worse than the phone-based system.

4. Comparison between Phone and Word Based LVCSR Systems

The LVCSR experiments so far showed that the word-based systems significantly outperformed the phone-based systems under limited training data scenarios. While in the WSJ0 task, though the improvement becomes smaller, the triword system still has better performance than the baseline phone-based system. In this section, further comparative experiments are conducted for the phone and triword based systems on the WSJ0 corpus from three aspects: word length, homophones and acoustic model.

4.1. Word length

It is well known that short words are more difficult to be recognized than long words in LVCSR tasks since fewer acoustic cues are available; and because the word-level HMMs theoretically have better acoustic modeling capabilities, the triword system should show better discrimination on short words.

There are about five thousand words in the WSJ0 5K test vocabulary. The word length ranges from 1 to 17 phones while over 92% of words in the vocabulary are within the range of 1 to 9 phones; and about 90% of the word occurrences in the WSJ0 corpus are words with lengths from 1 to 6 phones. Table 3 and Table 4 show the precision and recall of words with lengths from 1 to 6 phones in the WSJ0 Nov92 test set and the si_dt_05 subset, respectively.

It is clear that the precision and recall are positively correlated with the word length. Namely the longer the word is, the more accurate the LVCSR system can recognize the word. It can also be observed that the triword system indeed has a much better precision on short words (length 1~3 phones) than the conventional phone-based system. Especially for words with one phone, such as "a" and "I", the precision has about an absolute 2% to 3% improvement.

Table 3. Word length effect analysis (Nov92): precision and recall of the phone and word based systems for words with length from 1 to 6 phones. Only words with enough adaptation data are considered here.

Nov92	wLen	1	2	3	4	5	6
Precision	Phone Sys	63.9	75.4	84.5	**85.4**	**91.8**	89.2
	Word Sys	**66.9**	**76.0**	**85.5**	85.1	91.0	**89.5**
Recall	Phone Sys	**71.1**	78.2	84.4	86.8	**91.0**	**89.5**
	Word Sys	70.5	**78.8**	**85.0**	**87.1**	90.9	89.1

Table 4. Word length effect analysis (si_dt_05 subset): precision and recall of the phone and word based systems for words with length from 1 to 6 phones. Only words with enough adaptation data are considered here

si_dt_05 sub	wLen	1	2	3	4	5	6
Precision	Phone Sys	63.9	74.8	84.8	89.2	89.8	93.3
	Word Sys	**65.8**	**75.5**	**85.7**	**89.6**	**90.2**	**93.6**
Recall	Phone Sys	64.3	76.3	85.0	**88.9**	89.3	92.5
	Word Sys	64.3	**77.1**	**86.0**	88.5	**90.5**	**92.6**

4.2. Homophones

One advantage for word-based systems is that by using different models for different words, these systems are able to distinguish words having the same pronunciations from the acoustic aspect. In traditional phone-based systems this can only be done with the help of a language model. Although homophones' pronunciations are labeled the same in a dictionary, the real pronunciations of the words may vary according to their meaning and usage differences. Word-level HMMs can capture such slight differences and help the system with a better determination of these words. For example, in the Nov92 test set, the utterance 440c0207 (… CONTRACT **IN TWO TO** THREE WEEKS …) is always recognized by the baseline phone-based system as "… INTO TWO…" since the phone sequences for these two utterances are exactly the same, and the language model is not strong enough to correct the error; however, the utterance can be correctly recognized by the triword system using the same language model.

Table 5. Precision and recall of homophones for phone and triword systems on the WSJ0 Nov92 Test set.

	Precision	Recall
Phone-based System	81.96	84.75
Triword System	83.93	85.29

In the WSJ0 5K-word dictionary we find there are 738 words having the same pronunciation with some other words. While among these 738 words, 177 words have enough training data and dominate over 95% of the homophone occurrences in the corpus. In the Nov92 test data, these 117 words are present 1292 times. Table 5 shows the precision and recall of these homophones for both the phone and word based systems on the Nov92 test data. While the phone-based system can only rely on the language model to handle the homophone problem, the word-based system, by providing extra acoustic information, shows a better ability in dealing with homophones. The triword system has a 1.97% absolute improvement over the phone-based system on homophone detection precision, while 0.54% improvement on recall.

4.3. Acoustic Modeling

In the WSJ0 task, the language model weight is fine tuned for the phone-based system. Yet it is well known that a best language model weight for one LVCSR task might not be the best for another task; and sometimes a good language model might not be available as well. In such situations, acoustic models play more important roles in the LVCSR systems. In this section, we investigate the effect of language model weight on both phone and word based systems. Free word decoding experiments were also conducted to compare the phone and word based systems at the acoustic level directly.

4.3.1. Robustness against LM weight variation

Weight of the language model is always a parameter required to be fine tuned in a LVCSR system. Theoretically word-based HMMs have a better acoustic modeling ability, the performance of word-based systems should therefore be better than phone-based systems when the weight of the language model is tuned down. In other words, the word-based system should have better robustness against language model weight variation.

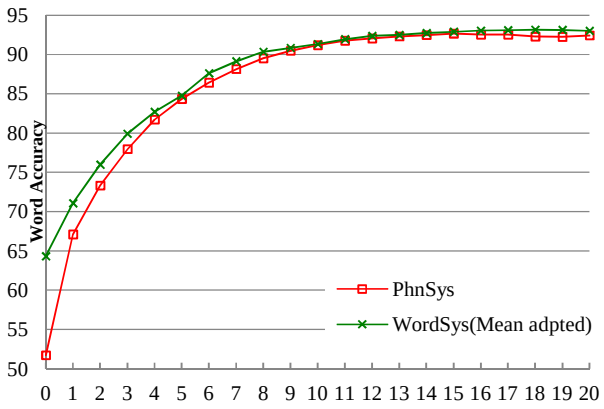

Figure 2: Effect of language model weight on lattice rescoring for both phone and word based systems.

Figure 2 shows word lattice rescoring results with different language model weights on the WSJ0 Nov92 test data. Word lattices were generated by the baseline phone-based system with the corpus-provided bigram language model. In Figure 2, the green curve with cross marks is the performance of the proposed triword system, while the red curve with square marks indicates the performance of the baseline phone-based system. It is obvious that word accuracies of both systems degrade when language model weight decreases. However when compared to the significant performance drop in the red curve, the proposed word-based system shows a smoother curve than the phone-based system. When the language model weight is set to 1, the word-based system outperforms the phone-based system by 4% absolute; while when no language model is used in the lattice rescoring, the word-based system has about 13% absolute improvement over the phone-based system. It is also worth noting that in Figure 2 the word-based system's performance is always above the phone-based system, which shows the word-based system has better robustness against language model weight variation than the phone-based system.

4.3.2. Free word decoding

Free word decoding provides another way to examine acoustic models. Table 6 shows the free word decoding performance of the phone and word based systems on the two WSJ0 data sets. It is clear that since word models are able to capture more pronunciation variation for each word than the phone models, the word accuracy of the word-based system is significantly better than the phone-based system (3.85% on the Nov92 data set and 5.98% on the si_dt_05 subset). The results also imply that, for detection-based ASR systems, using the proposed word-level HMMs for word detection has a great advantage over using phone-level HMMs.

Table 6. Free-word-decoding word accuracies of the phone and word based systems.

	Nov92	si_dt_05 subset
Phone-based System	30.82	34.70
Triword System	34.67	40.68

5. Conclusion

In this paper, we propose two methods for realizing word-based LVCSR systems. Experimental results on the TIMIT and WSJ0 corpora demonstrate that word-based systems have a performance improvement over phone-based systems. The proposed word-based systems also show a good performance on short words and are capable of disambiguating homophone words. Furthermore, they are more robust to the variation of language model weights.

The acoustic features used in this paper are conventional MFCC with its first and second derivatives. However, it is known that prosodic variation in words is also an important cue for word determination. In future studies, prosodic features, such as pitch and intensity, will be included as acoustic features in the word-based systems.

6. Reference

[1] K.-F. Lee, "On large-vocabulary speaker-independent continuous speech recognition," *Speech Communication* 7(4): 375–379 (1988)

[2] S. Greenberg, "Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation," *Speech Communication* 29(2-4): 159-176, 1999

[3] X. Liu, M. J. F. Gales, J. L. Hieronymus, and P. C. Woodland, "Investigation of Acoustic Units for LVCSR Systems," in *Proc. ICASSP* 2011

[4] S. M. Siniscalchi and C.-H. Lee, "A Study on Integrating Acoustic-Phonetic Information into Lattice Rescoring for Automatic Speech Recognition," *Speech Communication* 51(11): 1139-1153 (2009)

[5] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing* 2(2): 291-298, 1994

[6] http://htk.eng.cam.ac.uk

[7] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika* 40(3-4): 237-264, 1953

[8] C. D. Manning and H. Schuetze, *Foundations of statistical natural language processing*, The MIT Press, Cambridge, 1999

[9] http://www.speech.sri.com/projects/srilm