Combining multiple-sized sub-word units in a speech recognition system using baseform selection

T. Nagarajan, P. Vijayalakshmi, and Douglas O'Shaughnessy

INRS-EMT, University of Quebec
Montreal, Canada
raju@emt.inrs.ca

Abstract

A Longer-sized sub-word unit is known to be a better candidate in the development of a continuous speech recognition system. However, the basic problem with such units is the data sparsity. To overcome this problem, researchers have tried to combine longer-sized sub-word unit models with phoneme models. In this paper, we have considered only frequently occurring syllables and VC (Vowel + Consonant) units, and phone-sized units (monophones and triphones) for the development of a continuous speech recognition system. In such a case, even for a single pronunciation of a word, there can be multiple representational baseforms in the lexicon, each with different-sized units. We show that a considerable improvement in recognition performance can be achieved if the baseforms are selected properly. Out of all possible baseforms for a given word in the lexicon, the baseform that maximizes the acoustic likelihood, for possible sub-word unit concatenations to make a word, alone is considered. In the baseline systems' word-lexicon, like pure monophone or triphone-based systems, since only the acoustically weaker baseforms are replaced by baseforms with longer-sized units, the resultant performance is guaranteed to be better than that of baseline systems. The preliminary experiments carried out on the TIMIT speech corpus show a considerable improvement in the recognition performance over a pure monophone/triphone-based systems when the larger-sized units are combined using proper selection of baseforms.

Index Terms: speech recognition, baseform selection, syllable.

1. Introduction

In the past few decades, different types of sub-word units, like phoneme, diphone, demisyllable, and syllable, have been studied by researchers for developing automatic speech recognition (ASR) systems. Each of these sub-word units has its own advantages and disadvantages. However, in general, the larger the unit, the greater is the amount of phonological phenomena contained within the unit and also the greater is the capability of resolving and aligning events in the input speech with events in the sequence of concatenated units [1]. The major constraint with considering a larger-sized sub-word unit is that the inventory required to generate representative acoustic models for all the units is very high. To overcome the problem of data sparsity, different types of sub-word units (larger-sized with smaller-sized) can be combined.

The syllable was proposed as a unit for ASR as early as 1975 [2], in which irregularities in phonetic manifestations of phonemes were discussed. It was argued that the syllable will serve as the effective minimal unit in the time-domain. Since then, several syllable-based ASR systems have been developed for different languages.

Considering the fact that only few hundreds of syllables occur frequently in conversational speech, models can be generated only for those syllables, as described in [3]. In a word lexicon, the rest of the syllables can be replaced by their corresponding phoneme sequences. If context-independent phone models are used, replacing a phone sequence by a syllable sequence guarantees a better performance since the syllables capture the co-articulation effects well. However, if context-dependent phone models are used, replacing a phone sequence by a syllable sequence should be carefully done. For some words in the lexicon, a sequence of triphones may be a better choice than replacing it by a sequence of syllables or a combination of syllables and triphones.

Several results have been reported ([4] [5] [6]) in the literature on automatically deriving a baseform from a given speech signal, mainly for handling pronunciation variations and new words. In [4], a best baseform is automatically deduced for a new word by utilizing the actual utterances of the new word in conjunction with a set of automatically derived spelling-to-sound rules. In [5] [6], automatic baseform determination techniques are presented in which no prior information concerning the pronunciation of the words is used.

In the current study, we try to combine longer-sized units (syllable and VC) with phones (monophone and triphone). In the speech corpus considered for the present work, since the number of examples for many of the syllables was found to be very small, we have considered VC units also. We assume that the pronunciation variations of each word in the vocabulary are known *a priori*. When the different types of sub-word units are combined, each word in the lexicon can be expanded in multiple ways, according to the inventory of the sub-word units, even for a single pronunciation. In the present study, we use a simple Viterbi-decoding method that uses acoustic likelihood alone, for selecting an optimal baseform among multiple choices each with different types of sub-word units. In the baseline systems' word-lexicon, like pure monophone or triphone-based systems, since only the acoustically weaker baseforms are replaced by baseforms with longer-sized units, the resultant performance is guaranteed to be better than that of the baseline systems.

The rest of the paper is organized as follows. In the next Section, the experimental setup is described in detail. In Section 3, the possibilities of having multiple baseforms for a single pronunciation of a word and the technique used to select an optimal baseform is described. The performances of different systems developed are analyzed in Section 4.

2. Experimental setup

For the present work, the TIMIT corpus is considered for both training and testing. The number of unique words in the train data is $\approx$5000 and the number of unique words in the test data is $\approx$2500. Out of the 2500 test words, $\approx$50% of the test words are not available in the train data. The word-lexicon is created only with the test words. For the words that are in common in train and test data, pronunciation variations are taken from the transcriptions provided with the corpus and for the rest of the test words only one transcription is considered.

For the development of all the systems described here, the HTK is used. For each type of unit, a separate set of label files is created using the considered units alone. The rest of the labels are replaced by their corresponding phonemes. Only the word-internal syllables and VC units are considered. Features (13 static MFCC + 13 delta + 13 acceleration) are extracted with a frame size of 20 ms.

For the monophone-based system, 46 left-to-right (3 state, 1 mixture/state) models (including a model for silence) are initialized and trained on hand-labeled data provided in the corpus. For the re-estimation of model parameters, a standard Viterbi alignment procedure is used. The number of mixtures per state for each model is then increased to 32 in steps, by a conventional mixture splitting procedure. Each time, the model parameters are re-estimated twice.

Syllabification software [7] available from NIST was used to extract the syllables from the phonetic segmentation boundaries given in the TIMIT corpus. For all the phonetic transcriptions available with the TIMIT corpus, the first-best results given by the syllabification algorithm are considered. Even though the number of unique syllables in the training corpus is $\approx$ 5000, most of the syllables have very few examples. In our work, we have considered 200 syllables that have more than 50 examples. The rest of the syllables are replaced by their corresponding phonemes in the transcriptions. Since the number of resultant syllable models is very small, we have decided to combine the VC units also into the system. When compared with the syllabic unit, the VC units are more vulnerable to co-articulation effects. However, it is better than the monophones, in the sense that the co-articulation between the corresponding vowel and consonant is captured in the VC unit. Further, it covers most of the words in the corpus. For VC units also, the more frequently occurring (above 50 examples) units alone are considered, which is 172. The syllable and VC models are trained separately, in a similar fashion as monophone models. However, the number of mixtures for the syllables and VC units is restricted to only 8. The number of states is varied depending on the number of phones in these units.

For the triphone-based system, the initialized monophone models are cloned and re-estimated to yield around 12000 triphone models. A tree-based clustering algorithm is used for state tying. For the final models, the number of mixtures is increased to 8, in steps, using the conventional mixture splitting algorithm. For the out-of-vocabulary words, the required triphone models are synthesized.

For all the experiments, no optimization is carried out other than word-insertion penalty optimization. Further, in order to clearly see the effect of longer-sized units in combination with phone-sized units during Viterbi-alignment, no word level n-gram statistics is used. Testing is carried out on the whole test corpus of TIMIT.

3. Lexicon selection

As mentioned earlier, in the current study, we try to combine the phones, syllables, and VC units. When different types of units are combined to build a speech recognition system, each word in the lexicon can be represented in multiple ways. For example, for the word 'anything', the possible baseforms using different types of sub-word units can be

- monophones → eh n iy th ih ng
- syllables → eh niy thihng
- VC units → ehn iy th ihng
- triphones → eh+n eh-n+iy n-iy+th iy-th+ih th-ih+ng ih-ng
- triphones + syllables → eh+n niy thihng
- triphones + VC units → ehn n-iy+th iy-th+ih ihng

Here, based on the type of the basic unit (monophone or triphone), the baseforms are grouped into two categories. Depending upon the type of the unit and its appropriateness in a given word, the power of each baseform may be different. Considering all the baseforms for each pronunciation of a given word will increase the computation time considerably and at the same time, the performance may also degrade. Out of all the possible baseforms, the optimal one, in the acoustic likelihood sense, can be selected as described below.

Let W_k^i be the kth training example of the ith word, and L_j^i be the jth baseform of the ith word in the lexicon. For a given set of acoustic models $\lambda = \lambda_1, \lambda_2, ..., \lambda_N$, the optimal baseform L^i for the ith word in the lexicon can be selected as:

$$L^i = \arg\max_j \frac{1}{K^i} \sum_{k=1}^{K^i} \log\ p(W_k^i | L_j^i, \lambda). \quad (1)$$

In Equation 1, the K^i is the number of training examples available for the ith word. As mentioned in Section 2, the number of words that are common with test and train data is only 50%. When different systems are combined, the baseform selection procedure is adopted only for these common words. For the rest of the words, the original systems' baseforms are used.

Separate sets of experiments are carried out for monophone and triphone-based systems. For this study, we have not tried to combine monophones and triphones. The syllables and VC units are combined with either monophones or triphones.

4. Performance analysis

For performance analysis, we have considered both monophone-based and triphone-based systems as the baseline systems. Experiments have been carried out by combining phone models with the larger-sized units in a conventional way[1], and using the baseform selection procedure. The performance of all the systems are grouped into two categories based on the type of baseline models used, i.e., monophone-based (refer to Table 1) and triphone-based systems (refer to Table 2).

As expected, when the syllables or VC units are separately combined with the monophones, the word error rates (WER) of these systems are found to be 3% less than that of the pure monophone-based system (refer to Table 1). Even though the syllables are supposed to capture the co-articulation effects better than the VC units, the word error rates of syllable-based and VC-based systems are found to be nearly the same. The main reason for this behavior is that, since the number of syllabic units is small, the coverage of these units in the word-lexicon is less than that of VC units. Combining syllables and VC units is carried out in two different methods. In the first method, each word in the word-lexicon is transcribed as a combination of syllables, VC units, and monophones, wherever possible. In the second method, based on the selection criterion discussed in Section 3, the words are transcribed as either (a) syllables and monophones, or (b) VC units and monophones. We observed that, when the baseforms are properly selected, there is a considerable reduction in WER.

Table 1: Word error rates of monophone-based systems

System	Word error rate (in %)
monophone only	46.94
syllable + monophone	44.18
VC unit + monophone	44.15
syllable + VC unit + monophone	45.98
syllable + VC unit + monophone (baseform selection)	42.80

Similarly, another set of experiments was carried out using triphones instead of monophones and we observed a similar behavior in performance (refer to Table 2). Since many of the triphone models are much stronger, when we tried to include the syllables or VC units, a considerable increase in WER is observed when compared to pure-triphone system. However, when the baseforms are properly selected, the performances of the systems are improved over the system with triphones only.

When triphones are combined with syllables, the number of triphone-baseforms replaced by the syllable-baseforms is found to be only 7% of the common words. This is mainly due to fewer training examples available with the corpus. If the number of syllables in the system is increased, further improvement can be achieved.

Table 2: Word error rates of triphone-based systems

System	Word error rate (in %)
triphone only	39.17
syllable + triphone	40.80
VC unit + triphone	41.05
syllable + triphone (baseform selection)	38.56
VC unit + triphone (baseform selection)	38.51
syllable + VC unit + triphone (baseform selection)	37.88

5. Conclusions

In this paper, we have made an attempt to combine multiple-sized units together for continuous speech recognition. When the baseforms are selected in an optimal sense, the performance of the systems was found to be consistently better. From the performance analysis, it is noted that even with fewer longer-sized units in the system, the improvement in the performance is considerable. If more such units are used, we believe that the performance of the monophone-based system combined with longer-sized units can reach the performance of a triphone-based system. Further, we observe that in the absence of proper syllabic units, if a corresponding VC unit is used, the performance of the system can be improved.

6. References

[1] A. E. Rosenberg, L. R. Rabiner, J. G. Wilpon, and D. Kahn, "Demisyllable-based isolated word recognition system," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 31, no. 3, pp. 713–726, June 1983.

[2] O. Fujimura, "Syllable as a unit of speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, no. 1, pp. 82–87, February 1975.

[3] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. R. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE Trans. Speech, and Audio Processing*, vol. 9, no. 4, pp. 358–366, May 2001.

[4] L. R. Bahl, S. Das, P. V. deSouza, M. Epstein, R. L. Mercer, B. Meralso, D. Nahamoo, M. A. Picheny, and J. Powell, "Automatic phonetic baseform determination," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Apr. 1991, vol. 1, pp. 173–176.

[5] R. C. Rose and E. Lleida, "Speech recognition using automatically derived acoustic baseforms," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Apr. 1997, vol. 2, pp. 1271–1274.

[6] B. Ramabhadran, L. R. Bahl, P. V. deSouza, and M. Padmanabhan, "Acoustic-only based automatic phonetic baseform generation," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, May 1998, vol. 1, pp. 309–312.

[7] B. Fisher, "tsylb2-1.1 - syllabification software," http://www/nist.gov/speech/tools, August 1996.

[1] In the conventional method, for a given word, if a baseform with longer-sized units is available, irrespective of its effectiveness, the original baseform will be replaced by the new one.