ICASSP 2003, Hong Kong

SPLIT-LEXICON BASED HIERARCHICAL RECOGNITION OF SPEECH USING SYLLABLE AND WORD LEVEL ACOUSTIC UNITS

Abhinav Sethy, Shrikanth Narayanan

Speech Analysis and Interpretation Lab
Integrated Media Systems Center
Department of Electrical Engineering-Systems
University of Southern California
http://sail.usc.edu
[sethy,shri] @sipi.usc.edu

ABSTRACT

Most speech recognition systems, especially LVCSR, use context dependent phones as the basic acoustic unit for recognition. The primary motive for this is the relative ease with which phone based systems can be trained robustly with small amounts of data. However as recent research indicates, significant improvements in recognition accuracy can be gained by using acoustic units of longer duration such as syllables. Syllable and other longer length units provide an efficient way for modeling long term temporal dependencies in speech which are difficult to cover in a phoneme based recognition framework. But these longer duration units suffer from training data sparsity problem since a large number of units in the lexicon will have little or no acoustic training data. In this paper we present a two step approach to address the training data sparsity problem. First we use CD phones to initialize the higher level units in a manner which minimizes the impact of training data sparsity. Subsequently we present methods to split the lexicon into units of different acoustic length based on a analysis of the training data. We present results which show that a 25-30% improvement in terms of word error rate can be acheived by using CD phone initialization and variable length unit selection on a LVCSR task.

1. INTRODUCTION

Automatic speech recognition (ASR) systems typically focus on short-time information distributed over periods of 10-20ms. A speech signal is partitioned into overlapping frames of 20-30 ms for the purpose of feature extraction. Using this feature space representation and search constraints from the language model, a decoding process finds out the best matching word or phone sequence for the given speech signal. This short term representation of speech has proved very successfully in a wide range of recognition tasks. However there are good indications [1][2] that information distributed over longer period's of time, such as syllabic or word level time span can lead to substantial gains in recognition accuracy.

The number of different acoustic units required for a given recognition task is a function of the vocabulary size and the nature of the underlying acoustic units. For phonemes the number of basic models (without context modeling) is fixed, but if we decide to use syllable or word size units the number increases in general with the vocabulary size. Many of these units, corresponding to words which are not used frequently will have poor coverage in the training data. This sparsity of training data has been the main hindering block in using larger units for large vocabulary speech recognition tasks such as LVCSR or spoken name recognition. For small vocabulary tasks such as alphabet or digit recognition, larger units(typically word level units) are used more commonly.

In this paper we present techniques for initialization of larger units using CD phones which ensure that the system performs nicely even with minimal or no acoustic training data for training the larger units. Also, we provide a comparative evaluation of syllable/word and phoneme based systems which show the impact of suprasegmental properties on recognition accuracy. Our results (Sec.5) indicate that there is not much to gain beyond the syllabic time span. This would indicate that most temporal correlations in speech are limited to syllabic duration. We also present different criteria to split the lexicon such that we use the appropriate units for representing the vocabulary words.

The next section will discuss the motivation for hierarchical speech recognition in more detail. In section 3 we describe the design of our recognition systems. Training strategies and corpora are described in section 4. Comparative performance evaluations and our findings are discussed in section 5. In the concluding section, we provide a brief summary of our work, the major findings and an outline for future research.

2. HIERARCHICAL SPEECH RECOGNITION

The use of an acoustic unit with a longer duration facilitates exploitation of temporal and spectral variations simultaneously. Parameter trajectories and multi path HMMs[3][4] are examples of techniques that can exploit the longer acoustic context, and yet have had marginal impact on phone-based systems. Longer units of syllabic duration or more are much more effective in using the cross phone correlations and temporal dependencies.

In this paper we present recognition systems which use word and syllable units. Word level units represent the longest units possible in a typical LVCSR system which uses bigram word grammar. Word level units are also used extensively in isolated digit and alphadigit recognition tasks. The motivation for using syllables comes from recent research on syllable based recognition[1][2] as well as studies of human perception [5][6] which demonstrate the central role that the syllable plays in human perception and generation of speech. One important factor that supports the use of

syllables as the acoustic unit for recognition, is the relative insulation of syllable from pronunciation variations arising from addition and deletions of phonemes as well as coarticulation. In studies of the Switchboard corpus [7] it has been shown that syllables had a deletion rate of 1% whereas the deletion rate for phonemes was 12%.

The major challenge in using syllables and word level units for recognition is the training data sparsity problem. In [2] this problem is partially resolved by using only those syllables which have good coverage in the acoustic data. However syllable being a larger unit requires more training data than phone sized units and hence proper training of syllable level models using flat initialization strategies, as described in [2] is difficult. In addition, we need to estimate the advantage in terms of recognition accuracy which we can gain by moving from phoneme representation to syllable or whole word representation. This is important to minimize the increase in system complexity that arises from the use of larger units. In general the achievable improvement depends heavily on the training data, since that would decide how well higher acoustic units can be trained. Thus depending on the acoustic training data available to us we need to find out the proper representation for every word in the lexicon. The next section will describe in more detail our initialization and lexicon splitting strategy.

3. HIERARCHICAL RECOGNIZER DESIGN

At the first stage we build three separate recognizers corresponding to the different acoustic units of interest i.e. phoneme, syllable and word. The design of the phone based recognizer follows the standard flat start Baum Welch reestimation strategy with decision tree based triphone creation and clustering[8].

3.1. Syllable Recognizer

The first step in designing a syllable based recognition system is to prepare the syllabic lexicon. We represent syllables in terms of the underlying phone sequence. Thus given a phonetic transcription of the speech in a standardized format like Worldbet or IPA we can write a syllable representation by coming up with a set of syllable symbols from the phonemes comprising the syllable for e.g.. *Junior* with the phonetic transcription jh uw n y er can be represented in syllabic terms as 'jh_uw' 'n_y_er'.

The next stage in designing a syllable lexicon is to identify the phone clustering, which corresponds to the correct syllabic representation. The process of clustering phones to get a syllable representation is called syllabification. Syllabification principles are described in [9] as a set of rules which define permitted syllable-initial consonant clusters, syllable-final consonant clusters and prohibited onsets. Syllabification software available from NIST [10] implements these rules and comes up with a set of alternative possible syllable clusters given a phoneme sequence which are used to generate the syllabic lexicon.

The phone level HMM models have the same basic topology with equal number of states, for different phonemes. However syllable models require different number of states depending on their size. A syllable comprising four phonemes such as 's_w_eh_l' requires more number of states than single phonemes or other shorter syllables such as 't_eh_n'. To account for this the number of states was chosen to be three per phoneme comprising the syllable.

To initialize the models for the syllable recognizer we use pretrained CD phone models. The number of states in a given syllable model is the sum of the number of states of the constituent phoneme models. Moving from the leftmost phoneme, we pick the initial state parameters from the corresponding CD phone models. As an illustration consider the syllable m_uw_v. Assuming 3 states per phoneme, states 1-3 in the syllable model will be initialized using the CD phone m+uw, states 4-6 from the CD phone m-uw+v and states 7-9 from the CD phone uw-v (Fig 1). Thus we need to first build a CD phone recognition system for seeding the syllable recognizer.

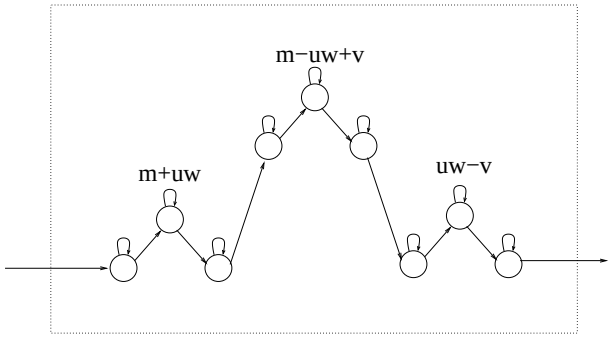

Fig. 1. Initialization of the 9 state syllable m_uw_v

3.2. Word Recognizer

The implementation of the word recognizer is similar to the syllable recognizer. The only difference being that the pronunciation phone sequence for every word in the lexicon is used as a separate word level unit. Thus we have acoustic units corresponding to all the different words in the lexicon. Homophones were given the same lexical representation. Model topology and initialization strategies are identical to the syllable recognizer.

Initialization from phoneme level models in this manner ensures that the syllable and word level models have performance identical to or only slightly lesser than the corresponding phoneme recognizer even without further acoustic training. Training on acoustic data leads to substantial improvement in accuracy as the temporal and spectral correlation information gets embedded in the longer length units. However the achievable gain results depends on the coverage of the unit in the training data as well as the linguistic nature of the unit. Thus a word or syllable unit with no training data will not lead to improvement in accuracy as compared to the corresponding phonemic representation. Thus we need to identify the proper lexical representation or choice of units to represent the words in the lexicon. We tried two different strategies for addressing this problem. The first uses a simple threshold based on the number of training units available in the acoustic data. The second uses the difference in recognition accuracy achieved on the training data by using syllable, phoneme or word level units. Once the decision is taken on what units are to be used, we split the word lexical entries appropriately. Certain commonly occurring words such as 'the' will have word level models. Other words will have either a pure syllable representation or a mixed syllable and phoneme representation. As an illustrative example, consider the phone level representation 'ae k t x r z' for 'Actors'. The pure syllabic representation would be ae_k t_x_r_z. However if the syllable t_x_r_z is not included in the list of acoustic units based on

the selection criteria mentioned above, we split the lexicon entry as 'ae_k t+x t-x+r x-r+z r-z'.

4. TRAINING : CORPORA AND IMPLEMENTATION

The system was trained and evaluated on the TIMIT corpus using the train and test sets provided. In the first stage we built a phoneme based recognizer. The feature space comprised of 26 mel frequency cepstral coefficients extracted at a frame rate of 10ms using a 16ms Hamming window. First and second order differentials plus an energy component were also included. For the baseline phone based recognizer, 46 three-state left-to-right phoneme models were initialized and trained on hand labeled data provided in the TIMIT corpus. These were then cloned to yield context triphone models, which underwent reestimation. Tree based clustering was used for state tying to ensure proper training of the models. Output distributions were approximated by eight Gaussians.

The syllable and word models were initialized using the CD phone models as described in section 3. The model parameters were then reestimated using the training acoustic data. This reestimation phase allowed the incorporation of cross phone correlations in the corresponding syllable and word models. For performance comparison purposes we also build syllable and word level systems using the standard flat start and embedded training of the acoustic models[2].

5. RESULTS

After training, recognition experiments were conducted on TIMIT test set. The language model used was a word level bigram network with a vocabulary consisting of around 4000 words.

5.1. Performance improvements through larger acoustic units

We compared the performance of recognizers using syllable and word level units with phoneme based system. Using CD phone system based initialization guarantees that the syllable and word level systems perform at-least as good as the baseline phoneme case even without reestimation. As can be seen in Table 1 which shows the achieved accuracy at the different stages of parameter reestimation, the initial accuracy is identical to the CD phone system for the word case and is slightly lower for syllable models which can be attributed to the lack of context modeling across syllable boundaries. Subsequent stages of reestimation embed the long term correlations in the syllable and word level units and the accuracy improves.

We compared the recognition accuracy achieved with syllable and word systems trained using the standard flat start strategy. As can be seen that the choice for the initialization strategy makes a significant difference in performance. Assuming that a typical syllable or word level model will have three or more phonemes, the number of parameters to be estimated for the model is around 3 times that for the phoneme models. Thus a large number of units in the flat start method are poorly trained. An analysis of the recognition errors confirmed that the performance difference between the flat start method and CD phone based initialization can be attributed to units which do not occur frequently in the training data.

Recognizer Type (ms)	First Reestimation	Third Reestimation
Context Free Syllable	72%	85%
Context Free Word	74%	87%
Context Dependent Phoneme	74%	74%

Table 1. Recognition accuracy of syllable and word level units at different stages of reestimation after CD phone initialization compared to baseline phoneme recognizer.

Recognizer Type (ms)	Flat initialization	CD phone based initialization
Context Free Syllable	80%	85%
Context Free Word	81%	87%

Table 2. Recognition accuracy after reestimation of syllable and word level units with and without CD phone based initialization

Recognizer Type (ms)	Accuracy	Number of model states in recognizers
Context Free Word	87%	43380
Context Free Syllable	85%	24460
Mixed Unit Recognizer	90%	13450

Table 3. Recognition accuracy and complexity in number of states of syllable, word and mixed lexicon recognizers

5.2. Mixed lexical unit recognizer

As described in section 3, we built mixed unit recognizers which combined syllable and word level units with context dependent phones. Two different criteria were used to determine the lexical split for every word in the vocabulary. The first is based on a analysis of the number of instances of syllable and word units in the training data. The second scheme does an analysis of recognition performance on training data itself to decided which representation is best. Currently the second scheme chooses between pure syllabic, word or phoneme representations only.

The complexity of the recognizers was evaluated in terms of the total number of states the models required (see Table 3). Each state has a Gaussian mixture model comprising of 8 Gaussians. As expected the mixed recognizer has a substantially lower complexity as compared to the syllable or word recognizers. Interestingly the mixed unit recognizers also had a a slightly higher accuracy. We can attribute this to the unit selection process which ensures that the mixed recognizers include only those units which are robustly trained. It is important to note that the complexity in terms of the physical model states is much lower for the context dependent phoneme recognizer. The CD phone recognizer had around 4000 distinct physical states and about 18000 logical states. Currently no form of state tying is used for word and syllable models.

6. CONCLUSIONS AND FUTURE WORK

In this paper we presented an ASR system which uses syllable and word level units in conjunction with CD phone system. To address the problem of training data scarcity for word and syllable units we used CD phone based initialization which guarantees that the higher level units will give equivalent performance to CD phone recognizer even in the absence of any acoustic training data. Our results comparing the performance of CD phone based system with syllable and word level systems shows that substantial performance gains can be achieved by exploiting the long term correlations present in speech by using longer duration units. We also describe methods to split the lexicon into units of different acoustic length based on a analysis of the training data so that we reduce the system complexity and ensure robust training. Currently we have tested our system on the TIMIT database which provides good phonetic coverage, but is lacking in terms of language coverage. The utterances are non-conversational and the language vocabulary as well as grammar are very restricted. Given the nature of our acoustic modeling using syllable and words combined with expected performance based variable unit length selection, we expect our results to generalize well across different corpora and language models.

Our results show that there is not much gain in recognition performance when word level units are used instead of syllable level units. This seems to indicate that most of the long term acoustic correlations are limited to syllable duration. Word and syllable distributions are different in natural speech from the TIMIT corpus. Experiments on corpora wchich allow for more natural language models such as WSJ are required to confirm this. We would also like to investigate this further by using information theoretic techniques to measure the correlations in speech which lie beyond syllable durations.

The unit selection or lexical split process tries to select only those syllable and word units which are expected to improve the recognition accuracy. We plan to extend this framework to include cost functions which will account for the complexity increase when using a larger unit and also the relative importance of different words. For example certain keywords might be of more importance to a ASR system and even if there is a small gain in accuracy when we switch from smaller units to larger units it might have a larger impact on the overall system performance. We are also investigating ways to incorporate state tying in a mixed recognizer so that we can reduce the complexity of the hybrid recognizer further.

We believe that using larger units will help in solving the spoken name pronunciation generation problem. Names have varied pronunciations and in tasks like directory assistance the name lists may have more than 100K names, which makes it impossible to have manual generation or verification of the pronunciation dictionaries. Extending the current phoneme based techniques for pronunciation generation such that they use variable length units(demisyllables/syllables/words) would open the possibility of compensating phonemic representation ambiguity in the acoustic model itself.

7. ACKNOWLEDGMENTS

This research was funded in parts by the Integrated Media Systems Center, an NSF ERC, under cooperative agreement No. EEC-9529152 and by the Department of the Army under contract number DAAD 19-99-D-0046.

8. REFERENCES

[1] Abhinav Sethy, Shrikanth Narayanan and S. Parthasarthy, "A syllable based approach for improved recognition of spoken names", *Proceedings of the ISCA Pronunciation Modeling Workshop,* Estes Park, Colorado, September 2002

[2] Ganapathiraju, J. Hamaker, M. Ordowski, G. Doddington and J. Picone, "Syllable-Based Large Vocabulary Continuous Speech Recognition", *IEEE Transactions on Speech and Audio Processing,* vol. 9, no. 4, pp. 358-366, May 2001.

[3] H. Gish and K.Ng, "Parameter trajectory models for speech recognition", *Proceedings of ICSLP, Philadelphia. PA, 1996,* pp 466-469.

[4] F.Kormazskiy, "Generalized mixture of HMM's for continuous speech recognition" , *Proceedings of ICASSP, Munich, Germany, 1997,* pp 1443-1446.

[5] Kirchhoff, K., "Syllable-level desynchronisation of phonetic features for speech recognition", *International Conference of Spoken Language Processing 1996,* pp 2274-2276.

[6] Su-Lin Wu, Brian Kingsbury, Nelson Morgan, and Steven Greenberg, "Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition", *ICASSP-98,* Seattle, pp. 721-724.

[7] S. Greenberg, "Speaking in Shorthand - A Syllable-Centric Perspective for Understanding Pronunciation Variation", *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition,* Kerkrade, The Netherlands, May 3-6, 1998.

[8] Odell J, Ollason D, Woodland P, Young S , Jansen J, "The HTK Book for HTK V2.0", Cambridge University Press, Cambridge, UK, 1995.

[9] D. Kahn, "Syllable-Based Generalizations in English Phonology", Indiana University Linguistics Club, Bloomington, Indiana, USA, 1976.

[10] W.M. Fisher, "Syllabification Software", *http://www.itl.nist.gov/div894/894.01/slp.htm,* The Spoken Natural Language Processing Group, National Institute of Standards and Technology, Gaithersburg, Maryland, U.S.A., June 1997.