SYLLABLE - A PROMISING RECOGNITION UNIT FOR LVCSR

Aravind Ganapathiraju, Vaibhava Goel, Joseph Picone, Andres Corrada, George Doddington, Katrin Kirchhoff, Mark Ordowski, Barbara Wheatley

The Syllable-Based Speech Processing Team
Summer Research Workshop on Innovative Techniques for LVCSR
Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD-21218

Abstract - We present an attempt to model syllable-level acoustic information as a viable alternative to the conventional phone-level acoustic unit for large vocabulary continuous speech recognition. The motivation for this work were the inherent limitations in the phone-based approach, primarily the decompositional nature and lack of larger scale temporal dependencies. In this paper we present preliminary but encouraging results on a syllable-based recognition system which exceeds the performance of a comparable triphone system both in terms of word error rate (WER) and complexity. The WER of the best syllable system reported here was 49.1% on a standard SWITCHBOARD evaluation.

1 INTRODUCTION

For at least a decade now the triphone has been the dominant method of modeling speech acoustics for speech recognition. However, triphones are a relatively inefficient decompositional unit due to the large number of frequently occurring patterns. Moreover, since a triphone unit spans an extremely short time-interval, such a unit is not suitable for integration of spectral and temporal dependencies. For applications such as SWITCHBOARD (SWB), where performance of phone-based approaches is unsatisfactory, the focus has shifted to a larger acoustic context. The syllable is one such acoustic unit. Its appeal lies in its close connection to articulation, its integration of some co-articulation phenomena, and the potential for a relatively compact representation of conversational speech.

We also conjecture that using a syllable as the fundamental acoustic unit obviates the need for explicit pronunciation modeling, since it can model many of the common variations in pronunciation based on a longer context window. Also, an analysis of the hand-transcribed data from the SWB corpus [1] revealed that the deletion rate for syllables was below 1%. Not surprisingly, the comparable rate for phone deletions was an order of magnitude higher — 12%. This is a clear indication of the stability of a syllable-sized acoustic unit.

The use of an acoustic unit with a longer duration also makes it possible to simultaneously exploit temporal and spectral variations. Parameter trajectories [2] and multi-path HMMs [3] are examples of techniques that can exploit the longer acoustic context, but have had marginal impact on triphone-based systems. Recent research on stochastic segment modeling of phones [4] demonstrates that recognition performance can be improved by exploiting correlations in spectral and temporal structure. However, these experiments were limited to phone-based systems — their viability on larger units is yet to be proven. We believe that applying these ideas to a syllable-sized unit, which has a longer contextual window, will result in significant improvements.

2 BASELINE SYSTEMS

In this research, we present two baseline systems: a context-independent monophone system and a word-internal triphone system. Both of these were carefully constructed to provide state-of-the-art performance on a standard SWB task within the constraints of the technology used for implementation. All systems described in this paper were based on a standard LVCSR system developed from a commercially available package — HTK [5]. We decided not to incorporate cross-word context for the syllable system, since this adds significant complexity to the decoder and only provides a marginal improvement in performance (a reduction in the WER from 49% to 45%). We also restricted our experiments to a bigram language model which could be efficiently processed in a lattice rescoring framework. Our recognition experiments were based upon rescoring lattices generated from a more sophisticated recognition system [6]. The lattices had an error rate of a little under 10%.

2.1 Phone-based Baseline Systems

Since the syllable system was constructed as a context-independent system (CI-SYL), a comparable context-independent phone (CI-PHN), or monophone system was constructed as a baseline. This system used a phone inventory consisting of 42 phones and a silence model (in addition, a word-level silence model was used as well). All phone models were standard 3-state left-to-right models without skip states. These models were seeded with a single Gaussian observation distribution. The number of Gaussians was increased to 32 per state during reestimation using a segmental K-means approach.

A context-dependent phone system (CD-PHN) was then bootstrapped from the CI-PHN system. The triphone models were initialized with a subset of the SWB data consisting of 10 hours of data chosen to span the variation in the corpus. The single-Gaussian monophone models from the **CI-PHN** system

were clustered and used to seed the triphone models. Four passes of Baum-Welch reestimation were used to generate single-component mixture distributions for the triphone models. These models were then increased to eight Gaussians per state using a standard divide-by-2 clustering algorithm. The resulting system had 81314 virtual triphones, 11344 real triphones, 34042 states and 8 Gaussians per mixture. The final count for the number of Gaussians is, however, reduced by tying states in the triphones.

Several features common in state-of-the-art SWB LVCSR systems were deliberately not included in this baseline system since the main goal of this work was to study the feasibility of syllables as an acoustic unit. In fact, it is hoped that some of these features will not be needed in a syllable system due to the inherent advantages of the syllable. The most prominent missing features were the use of a crossword decoder, a trigram language model, vocal tract length normalization, and speaker adaptation.

2.2 Syllable-based Baseline System

Perhaps the most critical issue in a syllable-based approach is the number of syllables required to cover the application. The number of lexical syllables in English is estimated to be on the order of 10000. This makes building a context-dependent syllable system a challenge, if not impossible. The first step in developing such a system was to represent each entry in the lexicon, previously defined in terms of phones, as a sequence of syllables — a process known as syllabification of the lexicon. We used a syllabified lexicon developed at Workshop'96 [7] for this stage. This lexicon consisted of over 70000 word entries for SWB and required 9023 syllables for complete coverage of the 60+ hour training data [1].

The model topology for the syllable models was kept similar to the **CD-PHN** system. However, each syllable model was allowed to have a unique number of states. The number of states was selected to be equal to one half the average duration of the syllable, measured in 10 msec. frames. The duration information for each syllable was measured from a forced alignment based on a state-of-the-art triphone system. Syllable models were trained in a manner analogous to the **CD-PHN** system, minus the clustering stage. The resulting models had 8 Gaussians per state.

3 HYBRID SYSTEMS

Given the limited syllable coverage achievable in the baseline system, it was imperative that a system comprising a mixture of phones and syllables be developed to handle words not covered by the syllabary. For computational efficiency reasons, this system was trained using a subset of the syllabary

consisting of all syllables that occurred at least 20 times in the training database. This resulted in a set of 2419 syllables. We refer to this approach of mixing acoustic units as a hybrid system (e.g., **CI-SYL-HY**).

Since the hybrid system had both syllables and phones, each unique word in the training database could be classified into one of three categories — syllable-only (**SO**) words have one or more syllables in their lexical representation but do not have any phones, phones-only (**PO**) words have only phones in their lexical representation and mixed (**MX**) words are represented in terms of both phones and syllables. Table 1 shows a comparison of the errors for a baseline syllable and triphone system. 'miss' are incorrectly recognized or deleted reference words. It is evident from the comparison that a syllable system's performance degrades on **MX** and **PO** words.

Data set	**# words**	**% miss**	
		CI-SYL	**CD-PHN**
All Words	18069	53	47
SO	15676	51	46
MX	1186	58	46
PO	1207	71	60

Table 1: Error Analysis of CI-SYL & CD-PHN

It was observed that many models in the above system were poorly trained. Due to time constraints, we circumvented this problem by building a system consisting of the 800 most frequent syllables and word-internal context dependent phones. It is interesting to note here that these 800 syllables covered almost 90% of the training data. The remaining 10% were replaced by its underlying phone representation. Several important issues, such as ambisyllabicity and resyllabification were ignored in this process. For example, if a syllable with an ambisyllabic marker was to be replaced by its phone representation, we ignored the marker all together. For instance,

sh_ey_d# _#d_ih_ng => sh ey _#d_ih_ng

The following example shows how the context for a sequence of phones was obtained from the adjoining syllables:

_ah_n k => _ah_n n-k
p _t_ih_ng => p+t _t_ih_ng

Syllable models from the above system and triphone models from the baseline triphone system were combined and reestimated using 4 passes of Baum-Welch over the entire training database.

4 FINITE DURATION MODELING

As mentioned before, we expect the syllable to be durationally more stable than the phone. However, when we looked at the forced alignments using our baseline system, we noticed very long tails in the duration histograms for many syllables. We also observed a very high deletion rate. This suggested a need for some additional durational constraints on our models.

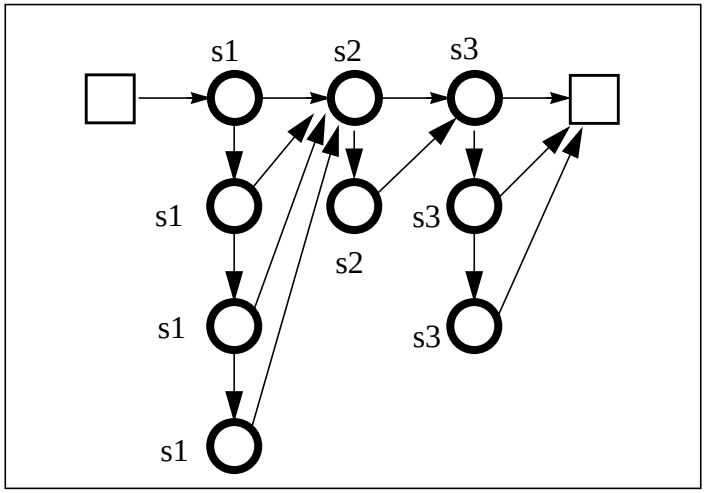

Figure 1: Finite duration HMM topology

To explore the importance of durational models, we decided to evaluate a finite duration topology. A model was created by using the corresponding infinite duration model as a seed, and replicated each state in the finite duration model P times, where P is obtained using the following equation:

$$P = E[S] + 2.stddev(S) = f(p) \quad (1)$$

where S is the number of frames that have been mapped to that state for a given syllable token. Note that this computation is a function of p, the self-loop probability. The observations of each replicated state are tied to the

observations of the entry state so that we maintain a manageable number of free variables for a model, and that there is sufficient training data per parameter. This topology is illustrated in Figure 1. To achieve a quick turnaround time we decided not to do a complete training of the models. Rather, we did a 4 pass reestimation of the seed models from the baseline syllable and triphone systems.

5 MONOSYLLABIC WORD MODELING

In the systems described thus far, syllables were represented in a context-independent manner. This, however, may not be a good assumption for some or all the syllables. Syllables that exist as a monosyllabic word, and also appear as part of the pronunciation of another word (an **MX** or **SO** word) demonstrate a much greater variation in pronunciation. We implemented a small number of monosyllabic word models as an attempt to capture some of this context dependency in syllables. Also, monosyllabic words constitute about 80% of the word tokens in SWB. The error rate on these words is about the same as the overall error rate. However, as a percentage of the total errors, monosyllabic words are clearly dominant. Hence, an experiment was conducted to create a separate models for 200 most frequent monosyllabic words. These words cover 71% of the word tokens in the training database.

The training data for this system was aligned using the **CI-SYL-HY** system. The alignments were relabeled to reflect the 200 monosyllabic words. A new syllable inventory was defined in which a syllable was included based on the number of remaining training tokens after removing the monosyllabic words. The durations of syllables and words were then reestimated. The final system had 200 monosyllabic words, and 632 syllables and word-internal triphones. It is referred to as **CI-SYL-MW**.

System	WER	Sub	Ins	Del
CI-PHN	62.3	41.4	2.5	18.7
CD-PHN	49.8	32.2	2.9	14.8
CI-SYL	55.1	35.7	2.5	16.9
CI-SYL-HY	51.7	33.9	3.5	14.3
CI-SYL-MW	49.3	31.8	3.1	14.4
CI-SYL-MW (fin. dur.)	49.1	32.2	3.6	13.3

Table 2: Summary of results

6 CONCLUSIONS AND FUTURE WORK

Table 2 summarizes the performance of various baseline and syllable systems. The major innovation of this system is the smooth integration of a mixture of acoustic models ranging from monosyllabic words to context-dependent phones, and including a large number of syllable models. The large performance improvement (a 2.4% absolute reduction in WER) with monosyllabic words and syllables can be attributed to the combination of multiple pronunciations in monosyllabic words into one acoustic model and separation of different monosyllabic words with the same baseform (e.g. _n_ow: KNOW, NO).

The system presented here is clearly deficient in a number of areas, including the representation of ambisyllabics in the lexicon, and the integration of syllable and phone models in a mixed word entry. We do believe, however, that the current system provides the proper framework to simultaneously exploit the temporal and spectral characteristics of the syllable by clustering or trajectory modeling. Preliminary results in this direction are promising. In a recent experiment we performed to validate the effect word models had on reducing the error rate, word models were used in conjunction with triphones and that gave only a marginal improvement in performance. This seems to indicate that mixing models of significantly different contexts may not be very useful. Another important area of research is the introduction of context-dependent syllables in a constrained way to keep the number of free variables in the system manageable. Note that the syllable systems presented here do not use any form of state-tying across models or states, yet contain fewer parameters than their comparable context-dependent phone systems. Hence, we believe that additional syllable models can be introduced without a significant increase in the overall system complexity.

REFERENCES

[1] Greenberg, S., "The Switchboard Transcription Project", *1996 LVCSR Summer Research Workshop*, Research Notes 24, CLSP, Johns Hopkins University, April 1997.

[2] Gish, H. and Ng, K., "Parameter Trajectory Models for Speech Recognition", *Proceedings of ICSLP '96*, pp. 466-469, Philadelphia, USA, April 1997.

[3] Korkmazskiy, F., et. al., "Generalized Mixture of HMMs for Continuous Speech Recognition", *Proceedings of the IEEE ICASSP* '97, pp. 1443-1446, Munich, Germany, April 1997.

[4] Ostendorf, M., and Roukos, S., "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition", *IEEE Transactions on*

Acoustics, Speech and Signal Processing, Vol. 37, No. 12, pp. 1857-1869, December 1989.

[5] Woodland, P., et. al., "HTK Version 1.5: User, Reference and Programmer Manuals", *Cambridge University Engineering Department & Entropic Research Laboratories Inc.*, 1995.

[6] Odell, J., "The Use of Context in Large Vocabulary Speech Recognition", Ph.D. Thesis, University of Cambridge, March 1995.

[7] Ostendorf, M. et.al., "Modeling Systematic Variations in Pronunciations via a Language-Dependent Hidden Speaking Mode", *1996 LVCSR Summer Research Workshop*, Research Notes 24, CLSP, Johns Hopkins University, April 1997.