SYLLABLE-BASED LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Aravind Ganapathiraju, Jonathan Hamaker, Joseph Picone

Institute for Signal and Information Processing
Mississippi State University
Mississippi State, MS 39762

Mark Ordowski, George R. Doddington[1]

Department of Defense
9800 Savage Road
Ft. George G. Meade, MD 20755

ABSTRACT

Most large vocabulary continuous speech recognition (LVCSR) systems in the past decade have used a context-dependent phone as the fundamental acoustic unit. In this paper, we present one of the first robust LVCSR systems that uses a syllable-level acoustic unit for LVCSR on telephone-bandwidth speech. This effort is motivated by the inherent limitations in phone-based approaches — namely the lack of an easy and efficient way for modeling long-term temporal dependencies. A syllable unit spans a longer time frame, typically three phones, thereby offering a more parsimonious framework for modeling pronunciation variation in spontaneous speech.

We present encouraging results which show that a syllable-based system 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 is 49.1% on a standard SWITCHBOARD evaluation, a small improvement over the triphone system. We also report results on a much smaller recognition task, OGI Alphadigits, which was used to validate some of the benefits syllables offer over triphones. The syllable-based system exceeds the performance of the triphone system by nearly 20%, an impressive accomplishment since the alphadigits application consists mostly of phone-level minimal pair distinctions.

EDICS: SA 1.6.2

CORRESPONDENCE: Aravind Ganapathiraju
Institute for Signal and Information Processing
Department of Electrical and Computer Engineering, PO Box 9571
Mississippi State University, Mississippi State, MS 39762
Phone: (601) 325-8335 Fax: (601) 325-3149
Email: ganapath@isip.msstate.edu

1. G. R. Doddington is now with the Information Technology Laboratory of the National Institute of Standards and Technology.

1. INTRODUCTION

For at least a decade the triphone has been the dominant method of modeling acoustics in speech recognition. However, triphones are a relatively inefficient decompositional unit due to the large number of triphone patterns with a nonzero probability of occurrence, leading to systems that require vast amounts of memory for model storage, and numerous models with poorly trained parameters. Moreover, since a triphone unit spans an extremely short time interval, integration of spectral and temporal dependencies is not easy. For applications such as the SWITCHBOARD (SWB) Corpus [1], where performance of phone-based approaches has stagnated over the past few years, we have shifted our focus to a larger acoustic context [2]. The syllable is one such acoustic unit whose appeal lies in its close connection to human speech perception and articulation, its integration of some coarticulation phenomena, and the potential for a relatively compact representation of conversational speech.

For example, consider the phrase “Did you get much back?” shown in Figure 1. This example[1] has been excised from a conversation in SWB. The first two words, “Did you,” are merged into the word “get” resulting in a pronunciation “jh y uw g eh”. Previous approaches to model such behavior involved the use of context-dependent (CD) phones [3]. However, since phones are deleted (or often extremely reduced), what is needed is higher-level information predicting the deletion of the phone. Modeling words as sequences of phones, though logical, is not justified when we try to derive a one-to-one mapping between the acoustics and the phone. Recent attempts at pronunciation modeling [4,5] have demonstrated limited success at modeling such phenomena. For example, in SWB, as many as 80 different pronunciations of the word “and” have been labelled [6]. The example in Figure 1, though an extreme case, demonstrates the challenges for explicit phone-based pronunciation modeling in conversational speech.

A syllable, on the other hand, seems to be an intuitive unit for representation of speech sounds. Without much difficulty listeners identify the number of syllables in a word [6] and, with a high degree of agreement, even the syllable boundaries. It is our conjecture in this paper that such behavior makes the syllable a more stable acoustic unit for speech recognition. The stability and robustness of syllables in the

1. This example is available at *http://www.isip.msstate.edu/projects/switchboard/faq/data/example_023*.

English language is further supported by comparing the phone and syllable deletion rates in conversational speech. In the analysis of data from a transcription project on SWB [6,7], it was estimated that the deletion rate for phones was 12%, compared to a deletion rate for syllables of less than1%. The example in Figure 1 supports this observation. This suggests that explicit pronunciation modeling becomes a necessary feature of phone-based systems to accommodate the high phone deletion rate in conversational speech, and can perhaps be circumvented by a syllable-based system.

The use of an acoustic unit with a longer duration facilitates exploitation of temporal and spectral variations [8] simultaneously. Parameter trajectories and multi-path HMMs [9] are examples of techniques that can exploit the longer acoustic context, and yet have had marginal impact on phone-based systems. Recent research on stochastic segment modeling of phones [10] demonstrates that recognition performance can be improved by exploiting correlations in spectral and temporal structure. We believe that applying these ideas to a syllable-sized unit, which has a longer acoustic context, will result in significant improvements in speech recognition performance while maintaining a manageable level of complexity.

In this paper, we describe a syllable-based system and compare its performance with word-internal and cross-word triphone systems on two publicly available databases: SWITCHBOARD (SWB) [11] and Alphadigits (AD) [12]. Note that these evaluations span the spectrum of telephone-based continuous speech recognition applications from spoken letters and digits to conversational speech. We focus on aspects of the syllable system which are significantly different from triphone systems, such as the model inventory and lexicon. We demonstrate some improvements that were achieved using monosyllabic word models and finite duration models.

2. BASELINE SYSTEMS

There are many ways to incorporate syllable information into a speech recognition system. Below, we describe several approaches that integrate syllable acoustic models with existing phone-based systems. We also describe the phone-based systems used as baselines since performance on these demanding applications depends significantly on the complexity of the technology used.

2.1. DESIGN OF A SYLLABLE-BASED LEXICON

By definition, a syllable spans a longer period of time compared to a phonemic unit [13]. In English, we can categorize different types of syllables using their consonant (C) and vowel (V) content. For example a syllable labeled CVC consists of a sequence of a consonant, vowel and consonant. An example of a CVC syllable is "_t_eh_n", pronounced "ten." Though other forms of syllables exist (for example, VVC and CCVVC), the CVC and CV syllables cover close to three-quarters of SWB [6]. The syllable is defined based on human perception and speech production phenomenon typically assisted by stress patterns within a word. For example, the word "atlas" intuitively appears to consist of two distinct segments. It is these segments that are called syllables. For ease of representation, syllables are typically represented as a concatenation of the baseform phones comprising the segment (_ae_t l_ax_s). It does not however mean that the acoustic segments always contain the phone sequence in its entirety.

Research conducted during the 1996 LVCSR Summer Workshop at the Center for Language and Speech Processing at Johns Hopkins University demonstrated that the use of syllable-level information and stress markings can reduce word error rate in triphone-based systems [14]. A by-product of this work was a high quality dictionary annotated with stress and syllable information. The annotated dictionary was developed from a baseline dictionary of about 90,000 words. Stress markings were obtained from the Pronlex dictionary [15]. Pronunciations were marked for primary stress and secondary stress. Syllabifications were introduced automatically using software provided by NIST [16]. This software implements syllabification principles [13] based on permitted syllable-initial consonant clusters, syllable-final consonant clusters and prohibited onsets. In order to keep the number of pronunciations and syllable units manageable, only one syllabification per word was used.

One complication in using syllables is the existence of ambisyllabic consonants: consonants at syllable boundaries that belong, in part, to both the preceding and the following syllable. Examples of this phenomenon, as they appear in the syllable lexicon described above, are:

ABLE → _ey_b# _#b_ax_l
GRAMMAR → _g_r_ae_m# _#m_er

The “#” denotes ambisyllabicity and is used as a tag for the phone which is the most plausible boundary. Though no clear syllable boundary exists, it makes sense to assume that the above examples consist of two syllables each. The most commonly occurring variant for each word was chosen and the *ambisyllabic* markings were retained. Hence, some syllables appear in our syllabary multiple times, with the additional entries containing the ambisyllabic maker “#” (which can appear at the beginning or end of the syllable).

For the systems discussed in this paper, stress was ignored for two reasons. First, our goal was to keep the baseline system as simple as possible and to prevent an abundance of undertrained acoustic parameters. Second, the value of lexical stress information seemed questionable. Even though stress plays an important prosodic role in English, the use of stress marks would increase the number of syllables by an order of magnitude and would induce a combinatorial explosion of parameters. Our syllabified lexicon for SWB consisted of about 70,000 word entries and required 9,023 syllables for complete coverage of the 60+ hour training data [17]. As explained in Figure 2, 3% of the total number of syllables appearing in SWB, which is approximately 275 syllables, cover over 80% of the syllable tokens in the training data. Of the 70,000 words represented in the lexicon, approximately 40% have at least one ambisyllabic representation.

2.2. BASELINE TRIPHONE SYSTEM

All systems described in this paper are based on a standard LVCSR system developed from a commercially available package, HTK [18]. This system, though extremely powerful and flexible, did not support cross-word decoding for an application as large as SWB. Considering the exploratory nature of this work, we decided not to use context dependency modeling in our syllable systems. Context dependent syllable models would introduce a few million models into the system, and this, in turn, would necessitate the use of clustering and/or state-tying. Many other features standard in state-of-the-art LVCSR systems [19], such as speaker adaptation and vocal tract length normalization, were similarly excluded from our study. Such things provide well-calibrated improvements in recognition performance, yet add to the overall system complexity, complicate the system optimization process, and require a deeper understanding of a baseline system’s performance before they can be successfully introduced. Therefore, our baseline syllable system is a word-internal context-independent system, while our baseline phone systems are word-internal context-dependent systems.

The phone-based system follows a fairly generic strategy for triphone training. The training procedure is

essentially a four-stage process consisting of:

1. *Flat-start monophone training*: Generation of monophone seed models with nominal values, and reestimation of these models using reference transcriptions.
2. *Baum-Welch training of monophones:* Adjustment of the silence model, reestimation of single-Gaussian monophones using the standard Viterbi alignment process.
3. *Triphone creation:* Creation of triphone transcriptions from monophone transcriptions, initial triphone training, triphone clustering, state tying, training of state-tied triphones.
4. *Mixture generation:* Split single Gaussian distributions into mixture distributions using an iterative divide-by-two clustering algorithm; reestimation of triphone models with mixture distributions.

The first two stages of training produce a context-independent monophone system. This system uses 42 monophones, a silence model and a word-level silence model (short pause). All phone models are 3-state left-to-right HMMs without skip states, and use a single Gaussian observation distribution. Ten hours of data was used for the flat-start process. This data was chosen to span variability in the corpus. A new silence model was created at this stage which had additional transitions intended to create a more robust model capable of absorbing impulsive noises (common in the training data). In addition, a Viterbi alignment of the monophone transcriptions was performed based on the fully-trained monophone models. The monophone models were reestimated using these Viterbi alignments.

A context-dependent (CD) phone system was then bootstrapped from the context-independent (CI) system. The single-Gaussian monophone models generated from the CI system were clustered and used to seed the triphone models. Four passes of Baum-Welch reestimation were used to reestimate the triphone model parameters. The number of Gaussians was, however, reduced by tying states [20]. Finally, these models were increased to eight Gaussians per state using a standard divide-by-two clustering algorithm. The resulting system had 81,314 virtual triphones (i.e. all triphones possible using an inventory of 44 phones and the lexicon), 11,344 real triphones, 6,125 states and 8 Gaussian mixture components per state.

2.3. PRELIMINARY SYLLABLE SYSTEM

The preliminary syllable system consisted of 9,023 syllable models. A standard left-to-right model topology with no skip states was used. The number of states in each model was set to one-half the median duration of the syllables measured using a 10 msec frame duration. The duration information for each syllable was obtained from a Viterbi alignment based on a state-of-the-art triphone system. Syllable models were trained in a manner analogous to the baseline triphone system, excluding the triphone clustering stage (unnecessary for a context-independent system). The resulting models, similar to the baseline phone system, had 8 Gaussian mixture components per state.

The models in this system, however, were poorly trained due to the nature of SWB. Of the 9,000 syllables appearing in the training database, over 8000 of these have fewer than 100 training tokens. From Figure 2 it is clear that a small portion of the syllabary is sufficient to cover nearly all of the database — 275 syllables cover 80% of the database. Hence, we chose to evaluate our approach using a system consisting of 800 syllables and replacing the remaining poorly trained syllables in the lexicon by their phonemic representation. An example is the word “access,” which is represented in this hybrid system as “_ae_k s eh s.” The symbol “_ae_k” represents a syllable model, while “s” and “eh” represent its phone constituents. Approximately 10% of the entries in the lexicon have syllable-only representations.

Note that the phones used for recognition were not trained in the context of syllables, but were trained separately as CI phones using 32 Gaussian mixture components per state. All of these phone models, which we refer to as “glue” phones, consisted of three state left-to-right models. This methodology of combining syllables and phones is not entirely appropriate and is the subject of on-going research since it is a combination of disparate model sets estimated in isolation. However, this approach represents a pragmatic solution to the problem of poorly trained acoustic models.

3. ENHANCED SYSTEMS

Though the use of syllables in conjunction with phones in lexical representations circumvented the problem of undertrained syllable models, model mismatch at phone-syllable junctions was still a significant problem. Below, we describe several modifications that were made to address this problem.

3.1. HYBRID SYLLABLE SYSTEM

We first approached the mismatch issue by building a system consisting of the 800 most frequent syllables and CI phones from scratch (rather than bootstrapping the models). 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 example, “shading” became “sh ey d d_ih_ng.”

The number of states for each syllable model was proportional to its median duration, while the phone models used standard three state left-to-right topologies with no skip states. The final models had 8 Gaussian mixture components per state. An evaluation of the system on the 2427 utterance test set resulted in a 57.8% word error rate (WER). A analysis of the errors occurring in the above experiment revealed that a very high percentage of words with an all-phone lexical representation or a mixed syllable/phone lexical representation were in error. Table 1 provides an analysis of the errors by word category.

This analysis motivated the development of a hybrid system using syllables and word-internal triphones. Following the approach above, CI phones were replaced with the corresponding CD phones (for example, “_ah_n k” became “_ah_n n-k”, and “p _t_ih_ng” became “p+t _t_ih_ng”). Syllable models from the baseline syllable system and triphone models from the baseline word-internal triphone system were combined and reestimated using 4 passes of Baum-Welch over the entire training database. Table 2 gives the performance of the hybrid system with CD phones.

3.2. MONOSYLLABIC WORD MODELING

One reason SWB is a difficult corpus for LVCSR is the variability in word pronunciations. Since the syllable is a longer acoustic unit compared to the phone, the need to explicitly provide pronunciations for all variants can be alleviated. It is possible for the syllable model to automatically consume the acoustical variation of the pronunciation of a word/syllable in the model parameters. A closer look at the training data in terms of its word content revealed some interesting facts. Table 3 shows the distribution of words in the 60+ hour training set. There were a total of 529 monosyllabic words in the training data. However, these 529 monosyllabic words covered 75% of the total number of word tokens in the training set. The top 200

monosyllabic words covered 71% of the total number of word tokens in the training set. Additionally, 82% of the recognition errors when using the hybrid syllable system were monosyllabic words. This suggested the need to explicitly model monosyllabic words.

In the monosyllabic word system, monosyllabic words with multiple pronunciations were represented by one model that represented all pronunciations. Table 4 provides some examples of this modification. Another example not mentioned in the table is the word “and.” In the lexicon its only pronunciation is “_ae_n_d.” However, in conversational speech the possible common alternative pronunciations could have deletions of “_ae,” “_d,” or both. Using the larger acoustic unit, it can be seen that the word model is not dependent on the lexical realization, and variation in pronunciation can be modeled by the HMM structure directly.

However, we decided to use separate models for words with different baseforms. Table 5 provides some examples of this modification. In spontaneous conversational speech, the monosyllabic word “no” has a significant durational variation depending on its position in a sentence. It is unlikely that the monosyllabic word “know” has this same characteristic. The difference in the duration of these models is, on an average, 80ms. This difference, therefore, necessitates two separate models for these homonyms with a different number of states in each model. The word model “know” was constructed with 9 states and the word model “no” had 13 states. Another example of this type of monosyllabic word is “to,” which is more likely to be pronounced as “_t_ax” rather than “_t_uw.” In this case the number of states in the word model “to” is 4 compared to 10 states for the word “two,” and 11 states for the word “too.”

Yet another ramification of introducing monosyllabic word models is the relationship between word models that were previously represented as syllables. The 800 syllables in the baseline system were replaced by 200 word models + 632 syllables. Some of the syllables were only trained from monosyllabic word tokens while others have training tokens from both monosyllabic and polysyllabic words. However, when using word models, some of the original syllables would have insufficient training material to reliably train both a word model and a syllable model. In other words, most of the occurrences of a given syllable were as a monosyllabic word. Separating the two occurrences resulted in a poorly trained syllable

version of the model.

The 200 word models were seeded with the most frequent syllable representation for that word. The number of states in the syllable and word models were reestimated by relabeling the forced alignments with the 632 syllables and 200 word models. As before, the number of states for each model was one half the median duration. The seed models of this monosyllabic word system (200 word models + 632 syllables + word internal triphones) that were obtained from the hybrid syllable system discussed in the previous section were reestimated using 4 iterations of Baum-Welch reestimation on the 60+ hour training set.

3.3. FINITE DURATION MODELING

As previously explained, a syllable is expected to be durationally more stable than a phone. However, when we examined the forced alignments using our hybrid syllable system, we noticed very long tails in the duration histograms for many syllables. The duration histogram for the syllable "_ae_n_d" is shown in Figure 3. The peak at the 8th frame in Figure 3 is an artifact of the requirement that all instances of the unit during forced alignment need to be at least 8 frames long (the number of states in the model) since the models do not have skip states. We also observed a very high word deletion rate. The deletions are somewhat attributable to the long tails in the duration histograms of syllable models. These facts suggested a need for some additional durational constraints on our models.

To explore the importance of durational models, we decided to evaluate a finite duration [21] topology. The finite duration topology we chose is shown in Figure 4. A model was created by using the corresponding infinite duration model as a seed, and replicating each state in the finite duration model P times, where P is obtained from

$$P = E[S] + 2 \cdot stddev(S) , \quad (1)$$

and S is the number of frames that have been mapped to that state for a given syllable token. The number of times the state is replicated is roughly proportional to the self-loop probability for the given state. Assuming a Gaussian distribution for the duration of a syllable, the above equation guarantees that at least 90% of the training tokens for the syllable can be explained by the estimated probability. The observations

of each replicated state are tied to the observations of the entry state so we maintain a manageable number of free variables for a model and so there are at least 100 instances of training data per parameter. To achieve a quick turnaround time we decided not to do a complete training of the models. Rather, we did a four pass reestimation of the finite duration monosyllabic words and syllables from the monosyllabic word system described in the previous section.

4. EVALUATIONS

In this section, we demonstrate improved performance for our syllable system on two drastically different tasks involving speech collected from the public telephone network: the SWITCHBOARD (SWB) Corpus [11] which consists of conversational speech and the OGI Alphadigit (AD) Corpus [12] which consists of spoken letters and numbers.

4.1. SWITCHBOARD

The syllable system trained on SWB data was evaluated on the test set created for the 1997 LVCSR Summer Workshop at Johns Hopkins University [17]. In the first full evaluation, we attempted to test fairly generic phone and syllable systems. The syllable system consisted of the 800 most frequently occurring syllables and 42 monophones. The results of these experiments are shown in Table 2. The syllable system performs exactly as we would expect, better than the context-independent monophone and worse than the context-dependent triphone system, at 56.4% WER.

An error analysis of the baseline syllable system reveals some interesting facts. Table 1 shows the word errors segregated by their representation in the test lexicon: words with an all-syllable representation, words with a mixed representation (syllable and phone) and words with an all-phonemic representation. The error rate in words with a mixed representation and an all-phone representation is rather high. This suggests a mismatch in using phones and syllables together, where each is trained separately. The hybrid syllable system tries to address this problem by using 800 syllables and word-internal triphones together. These models were reestimated together to better model contextual effects. The hybrid system improved the performance significantly to 51.7% WER. Tables 1 and 3 also indicate the need for better modeling of monosyllabic words, not just because of a high incidence of errors for these words but also because they

cover over 80% of the database.

The monosyllabic word system consisted of 200 monosyllabic words, 632 syllables and word-internal triphones. This system, with an error rate of 49.3%, gave a reduction in error rate of 2.4% absolute, compared to the hybrid syllable system. Using finite duration topology for this system reduced the WER marginally by 0.2% to 49.1%. This result needs further investigation since durational analysis of forced alignments suggest that finite duration topology would have a considerable effect in reducing the deletion rate. The deletions however reduced only marginally and the disparity between deletion and insertion rate was not effected (13.3% deletions vs. 3.6% insertions).

4.2. ALPHADIGITS

To obtain a measure of the extensibility of our continuous speech recognition systems developed on the SWB data, we also investigated performance on a smaller scale task: alphadigits. The reason this was appealing was that the alphadigit task was far removed from the task of large vocabulary continuous speech recognition. The corpus we chose comprised of short controlled segments of prompted speech — much different than the spontaneous, unpredictable SWB utterances. On this type of speech data, read speech, one would expect minimal differences in the performance of the syllable and triphone systems due to less significant coarticulatory effects and pronunciation variations. Another reason we were drawn to this task was its real-world applicability. Telephone alphadigit recognition has been of interest to Bell Labs and others since the 1970's [22]. Many applications (security, automated telephone services, etc.) would be enhanced if a user's spelled or spoken response could reliably take the place of the keypads which are pervasive today. Evaluation on alphadigit data would also give us a measure of the portability of our syllable system where we could ascertain that syllables would give us a performance improvement irrespective of the domain of application.

A robust and reliable alphadigit system has long been a goal for speech recognition scientists. Most recent work on both alphabet and alphadigit systems focuses on resolving the high rates of recognizer confusion for certain word sets, particularly the E-set (B, C, D, E, G, P, T, V, Z, THREE) and A-set (A, J, K, H, EIGHT). The problems with these tasks occur mainly because the acoustic differences between the letters

of the sets are minimal. For instance, the letters B and D differ primarily in the first 10-20 ms during the consonant portion of the letter [23]. Many techniques have been used to deal with these minimal-pair distinctions, such as an inclusion of weighting functions in dynamic time warping [24], and knowledge-based approaches [25]. In the 1980's, HMM-based systems [26] were first applied to this task. Enhancements to these HMM-based systems have yielded the best performance to date, 8.3% WER, on alphabets [23]. Table 6 summarizes many important systems for alphadigit recognition developed over the years [23]. These results, however, are not a fair comparison because the evaluations were done on different databases and different feature extraction techniques.

The corpus we chose to use for alphadigit recognition was the OGI Alphadigit corpus [12]. This corpus bears a large resemblance to the SWB Corpus in that the data collection conditions were similar: both were collected digitally over T1 telephone lines. The corpus consists of approximately 3000 subjects, each of whom spoke some subset of 1102 phonetically-balanced prompting strings of letters and digits. All experiments were performed on a training set of 51545 utterances from 2237 speakers and evaluated on 3329 utterances from 739 speakers [29]. Three systems were developed: cross-word triphones (CW-TRI), word-internal triphones (WI-TRI) and context-independent syllables (CI-SYL).

Table 7 summarizes the performance of the AD experiments described in this paper. The CI-SYL system not only performs better than its triphone counterpart, WI-TRI, (by approximately 3% absolute), it also performs better than the CW-TRI system by a 2% absolute difference. It is also interesting to note the word error rates for both the alphabets and digits separately. The syllable system makes its greatest gains in recognition of the alphabets whereas it lags in performance on the digit recognition.

Table 8 gives an analysis of the primary contributors to the error rate. It is somewhat curious to note that the syllable models perform better than the triphone models in E-set, A-set and S-F pair recognition. One would expect the phones to do better on this portion of the task given their fine-grain phonetic contexts. On the other hand, the phone system performs better on nasals. The word SIX accounts for a large portion of the errors on digits for the syllable system. If we analyze the errors in the syllable and triphone systems without the word SIX, the syllable system's performance exceeds that of the triphone system,

5.3% vs. 5.7% WER. The syllable _s_ih_k_s that is used to model SIX, is a CVCC syllable. Owing to the low bandwidth telephone quality data being used, the consonants in this case are easily confusable. Cross-word triphone models do a better job of modeling this context.

Not only do the syllable models achieve a lower word error rate, but they do so in a more efficient manner. Table 9 shows some complexity statistics for both triphone systems and our best syllable system. In this table, the term “logical model” refers to a triphone model that is generated as a concatenation of existing models. In contrast, the term “physical model” refers to the a HMM that is estimated after a decision-tree based context tying [18]. Note that there has been a considerable reduction in the number of models from context-dependent cross-word triphones to the context-independent syllables. The number of total states has been reduced considerably in the phone-based systems by way of state-tying. Though the syllable system contains more states than the word-internal triphones, the search space for the syllable system is significantly smaller. Both these factors result in a recognition speedup by a factor of **seven** over the triphone system.

5. CONCLUSIONS

We have presented a series of recognition experiments on data comprising two vastly different speaking styles (spontaneous telephone speech and read telephone alphadigits) using a syllable-based acoustic model. Our results are summarized in Table 10. Results indicate that syllables are a promising unit for recognition in LVCSR. The major innovation of our syllable system is the smooth integration of a large inventory of syllable models and a mixture of acoustic models ranging from monosyllabic words to context-dependent phones. The syllable-based system with monosyllabic word models gives an absolute 1% reduction in WER on a standard SWB evaluation set as compared to a similar word-internal triphone system. The best WER achieved on SWB was 49.1%. Additionally, the complexity of the syllable-based system is lower than the comparable triphone system when state-tying and clustering are not employed. Evaluations on the OGI Alphadigit data confirmed the gain we achieve by using syllables as a replacement for triphone units. The syllable-based alphadigit system achieved a performance of 10.4% WER, an absolute 2.9% decrease in WER compared to a cross-word triphone system.

Our work also shows the need for better modeling of monosyllabic words, which form a large portion of the SWB database. A performance improvement (a 2.4% absolute reduction in WER) with monosyllabic word models and syllables was achieved. This gain 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”). If this effect scales well with other improvements, we believe that the syllable alleviates the need for explicit pronunciation modeling in SWB [14] (current approaches to pronunciation modeling are very compute intensive).

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 recently performed experiment, word models were used in conjunction with triphones and resulted in only a marginal improvement in performance. This seems to indicate that mixing models of significantly different contexts and lengths may not be very useful. This could also explain the minimal gain we see in SWB experiments as compared to Alphadigit data where the system was purely syllable-centric. Another important area of interest is the introduction of context-dependent syllables in a constrained way to keep the number of free variables in the system manageable. We believe that additional syllable models can be introduced without a significant increase in the overall system complexity by using state-tying and clustering.

6. REFERENCES

[1] J. Godfrey, E. Holliman, and J. McDaniel, “SWITCHBOARD: Telephone Speech Corpus for Research and Development,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 517-520, San Francisco, California, USA, March 1992.

[2] H. Bourlard, H. Hermansky and N. Morgan, “Copernicus and the ASR Challenge -- Waiting for

Kepler," *Proceedings of the DARPA Speech Recognition Workshop,* pp. 157-162, Harriman, New York, USA, February 1996.

[3] K. F. Lee, "Context-Independent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 38, pp. 599-609, April 1990.

[4] M. Riley, et. al., "Stochastic pronunciation modeling from hand-labeled phonetic corpora," Proceedings of the Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, The Netherlands, pp. 109-116, May 1998.

[5] M. Saraclar, et. al., "Pronunciation Modeling Using a Hand-Labeled Corpus for Conversational Speech Recognition," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 313-316, Seattle, Washington, USA, May 1998.

[6] 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.

[7] S. Greenberg, "The Switchboard Transcription Project", presented at the *1996 LVCSR Summer Research Workshop,* Johns Hopkins University, Baltimore, Maryland, USA, August 1996.

[8] H. Gish, and K. Ng, "Parameter Trajectory Models for Speech Recognition," *Proceedings of the IEEE International Conference on Speech and Language Processing,* pp. 466-469, Philadelphia, Pennsylvania, USA, October 1996.

[9] F. Korkmazskiy, "Generalized Mixture of HMMs for Continuous Speech Recognition," *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing,* pp. 1443-1446, Munich, Germany, April 1997.

[10] M. Ostendorf and S. Roukos, “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.

[11] J. Picone, “WS’97 Databases and Tools”, *http://www.clsp.jhu.edu/~support/www97*, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland, U.S.A., August 1997.

[12] M. Noel, “Alphadigits,” *http://cslu.cse.ogi.edu/corpora/alphadigit*, Center for Spoken Language Understanding, Oregon Gradaute Institute of Science and Technology, Portland, Oregon, U.S.A., May 1997.

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

[14] M. Ostendorf, et. al., “Modeling Systematic Variations in Pronunciations via a Language-Dependent Hidden Speaking Mode,” presented at the *1996 LVCSR Summer Research Workshop,* Johns Hopkins University, Baltimore, Maryland, USA, August 1996.

[15] C. McLemore, “Pronlex Transcription,” *http://www.cis.upenn.edu/~ldc/readme_files/com-lex.readme.html*, The Linguistic Data Consortium, The University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A., September 1997.

[16] 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.

[17] G. Doddington, et. al., “Syllable-Based Speech Recognition.” WS’97 Final Technical Report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland, U.S.A., December 1997.

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

[19] L. Xiaoqiang and F. Jelinek, "Probabilistic Classification of HMM States for Large Vocabulary Continuous Speech Recognition," presented at the 9th Hub-5 Conversational Speech Recognition Workshop, Linthicum Heights, Maryland, USA, September 1998.

[20] A. Ganapathiraju, V. Goel, J. Picone, A. Corrada, G. Doddington, K. Kirchoff, M. Ordowski and B. Wheatley, "Syllable - A Promising Recognition Unit for LVCSR," *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 207-214, Santa Barbara, California, USA, December 1997.

[21] J. Picone, "Continuous Speech Recognition Using Hidden Markov Models," *IEEE Signal Processing Magazine*, vol. 7, no. 3, pp. 26-41, July 1990

[22] L. Rabiner, and J. Wilpon, "Some performance benchmarks for isolated word speech recognition systems," *Computer Speech and Language*, vol. 2, pp. 343-357, 1987.

[23] L. Philipos and A. Spanias, "High-Performance Alphabet Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 6, pp. 430-445, November 1996.

[24] L. Rabiner, L. Levinson, A. Rosenberg, and J. Wilpon, "Speaker independent recognition of isolated words using clustering techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-27, pp. 336-349, August 1979.

[25] R. Cole, R. Stern, and M. Lasry, "Performing fine phonetic distinctions: Templates vs. features," in Inference and Variability of Speech Processes, J. Perkell and D. Klatt, Eds. New York: pp. 325-341 Lawrence Erlbaum, 1986.

[26] L. Rabiner and J. Wilpon, "Isolated word recognition using a two-pass pattern recognition

approach," *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 2, pp. 724-727, Atlanta, Georgia, USA, April 1981.

[27] S. Euler, B. Juang, S. Lee and F. Soong, "Statistical segmentation and word modeling techniques in isolated word recognition," *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 2, pp. 745-748, Albuquerque, New Mexico, USA, April 1990.

[28] E. Huang and F. Soong, "A probabilistic acoustic MAP based discriminative HMM training," *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 693-696, Albuquerque, New Mexico, USA, April 1990.

[29] J. Hamaker, A. Ganapathiraju, and J. Picone, "A Proposal for a Standard Partitioning of the OGI AlphaDigit Corpus," *http://www.isip.msstate.edu/projects/speech/software/asr/research/syllable/alphadigits/data/ogi_alphadigits/eval_trans.text*, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, U.S.A., August 1997.

List of Figures

List of Tables

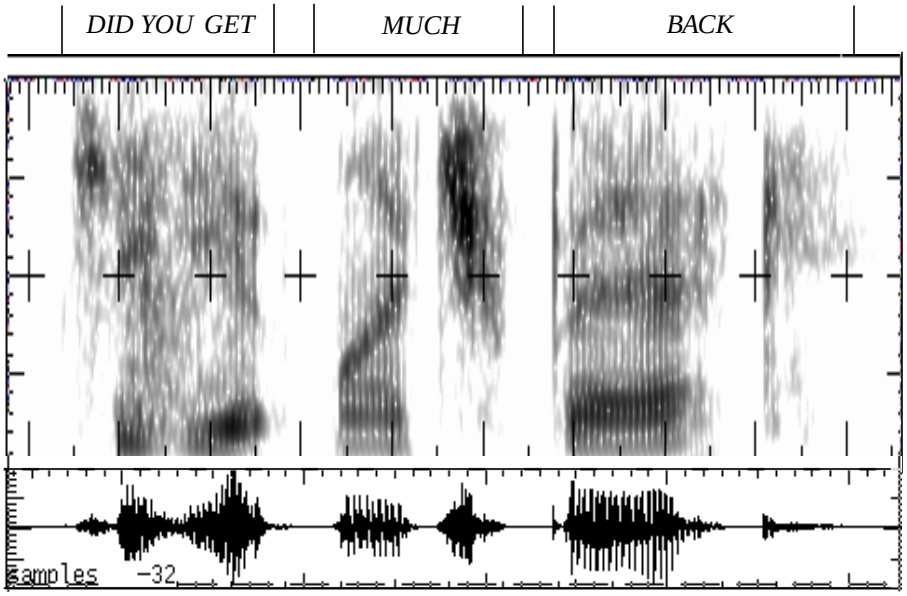

Figure 1. An example of phone deletion in the phrase “Did you get much back?” excised from a Switchboard conversation. “Did” is reduced and merged into “you get” such that the resulting word is pronounced “jyuge.”

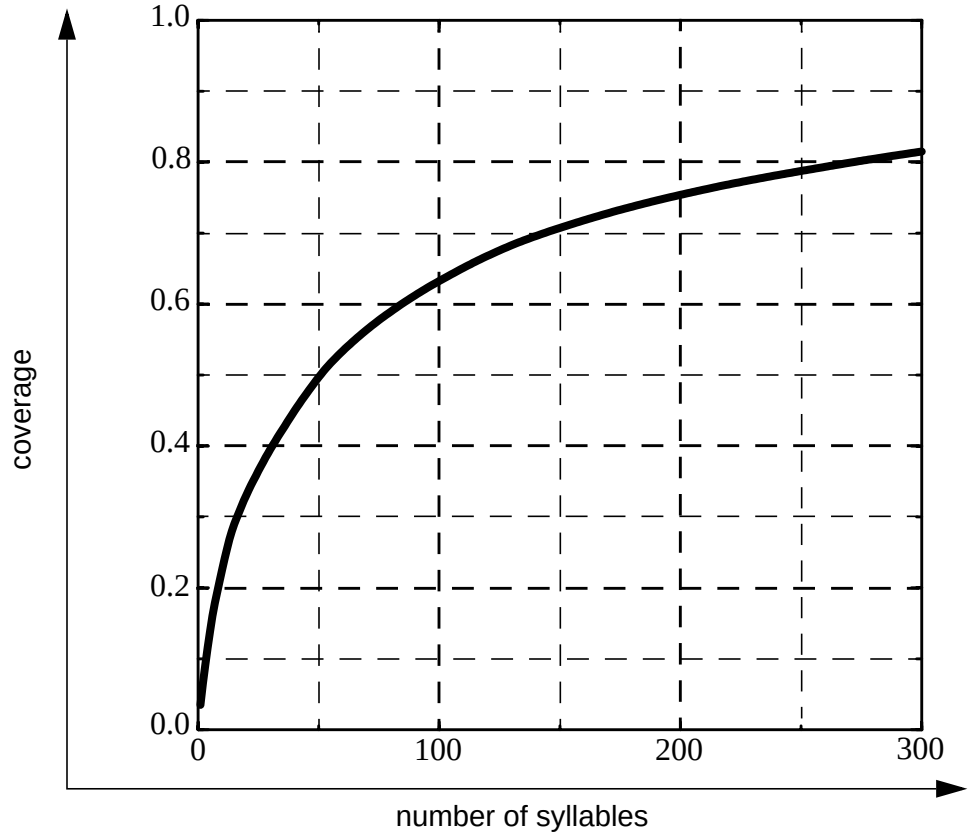

Figure 2. A cumulative histogram of the syllable tokens appearing in SWB transcriptions.

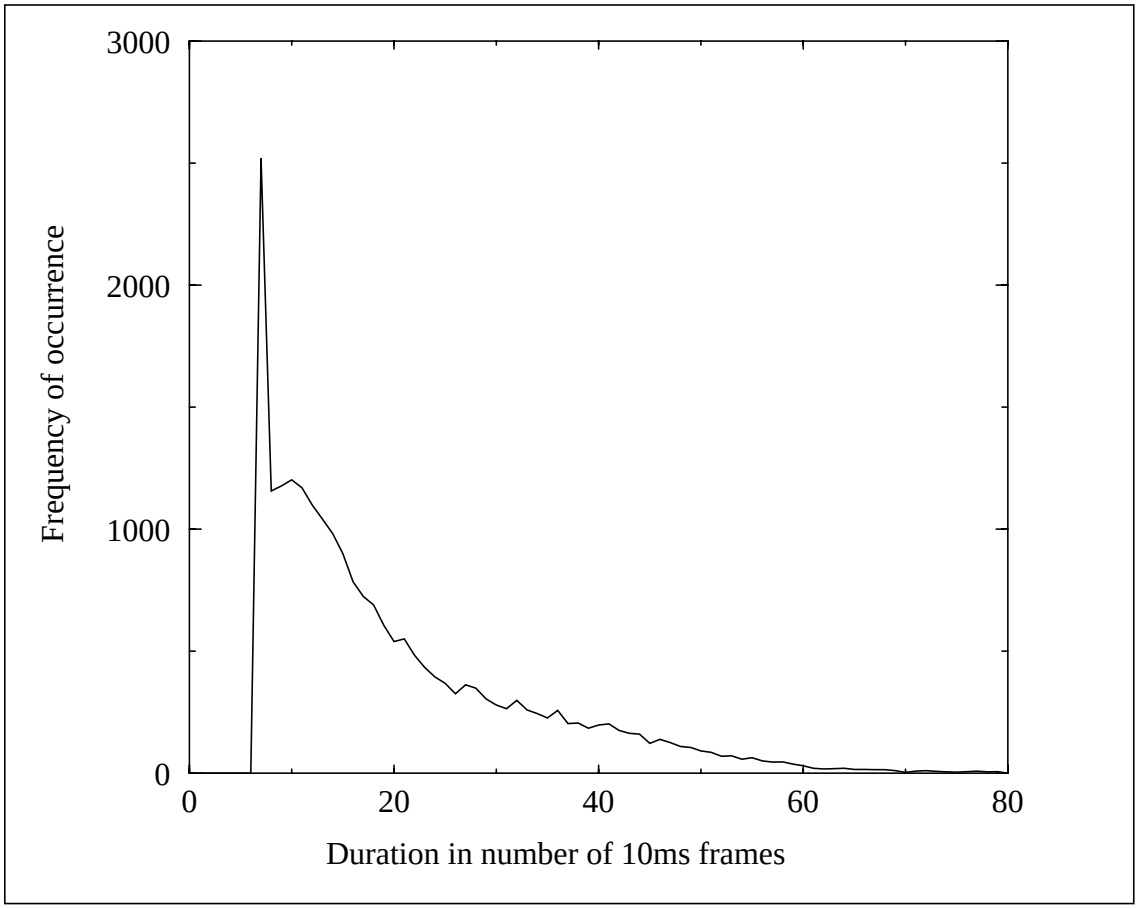

Figure 3. Duration histogram for the syllable "_ae_n_d."

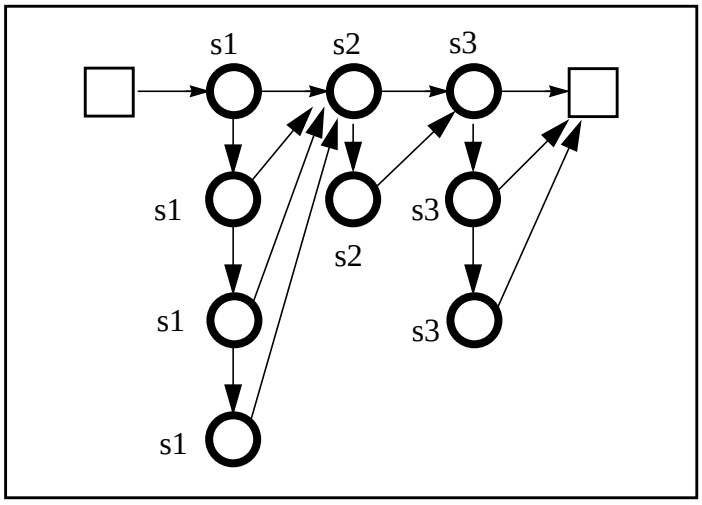

Figure 4. A finite duration HMM topology.

Lexical Representation	Total words	Words in error	Word Error Rate
Words with syllables only	15732	7321	47%
Monosyllabic words	13585	6504	48%
Words with syllables + phones	1186	663	56%
Words with phones only	1226	830	68%

Table 1: Error analysis of the hybrid syllable + monophone system.

System	Word Error Rate
Context-Independent monophones	62.3%
Word-internal triphones	49.8%
800 Syllables and 42 monophones	56.4%
800 Syllables and word-internal triphones	51.7%
632 Syllables, 200 monosyllabic words and word-internal triphones	49.3%
Finite duration monosyllabic words and syllables and word-internal triphones	49.1%

Table 2: Summary of system performance with CD phones.

Category	Count/Percentage
Unique Words	15,127
Number of Word Tokens	659,713
Number of Monosyllabic Words (dependent on lexicon/alignments)	529
Monosyllabic word tokens covered by the top 200 Monosyllabic words	95%
Word tokens covered by the 529 Monosyllabic words	75%
Word tokens covered by the top 200 Monosyllabic words	71%

Table 3: Analysis of the frequency of words appearing in the training data.

Word Model	Pronunciation Variants
THE	_dh_ah _dh_iy _dh_ax
FOR	_f_er _f_ow_r
TO	_t_ax _t_uw

Table 4: Examples of a monosyllabic word model representing all pronunciation variants.

Words	Word Models
KNOW, NO	__know, __no
THERE, THEIR	__there, __their
TO, TOO	__to, __too

Table 5: Monosyllabic words for which separate word models were used for each baseform.

Researchers	Year	Bandwidth	Speaker Dependent	Speaker Independent
Rabiner, et al. [24]	1979	3.2 kHz	---	21.0
Rabiner and Wilpon [26]	1981	3.2 kHz	11.5	15.4
Rabiner and Wilpon [22]	1987	3.2 kHz	10.5	---
Euler et al. [27]	1990	3.2 kHz	7.0	---
Huang and Soong [28]	1990	3.2 kHz	10.0	---

Table 6: A summary of prior alphadigit related research in terms of WER.

System	Total WER	Alphabet WER	Digit WER
Cross-word Triphones	13.3%	16.5%	5.4%
Word-internal Triphones	15.2%	19.4%	5.1%
Syllables	10.4%	12.1%	6.3%

Table 7: Alphadigit WER analysis by word category: alphabet or digit.

Confusion set	Triphone System WER	Monosyllabic Word System WER
E-Set	22.2%	18.5%
S-F pair	19.4%	17.2%
A-Set	16.7%	10.0%
Nasals	11.1%	14.3%

Table 8: Error analysis by word category: E-Set, S-F pair, A-Set, or Nasal.

System	Logical Models	Real Models	Number of States
Cross-word Triphones	25202	3225	2045
Word-internal Triphones	25202	1011	249
Syllables	42	42	900

Table 9: Relative complexity of systems for alphadigit recognition.

System	Model Set		Features
	Training	Testing	
Baseline Triphone	11344 word-internal triphones	11344 word-internal triphones	State tying used 8 Gaussians/state 3 states per model
Preliminary Syllable	9023 syllables	800 syllables and context-independent phones	8 Gaussians/state Number of states proportional to avg. duration
Hybrid Syllable	800 syllables and context-independent phones	800 syllables and context-dependent phones	8 Gaussians/state Number of states proportional to avg. duration
Monosyllabic Word	200 monosyllabic words, 632 syllables and context-dependent phones	200 monosyllabic words, 632 syllables and context-dependent phones	8 Gaussians/state Number of states proportional to avg. duration
Finite Duration	200 monosyllabic words, 632 syllables and context-dependent phones	200 monosyllabic words, 632 syllables and context-dependent phones	8 Gaussians/state Max. stay duration in syllable and word models

Table 10: Summary of LVCSR systems discussed in the paper.

Reply to comments from reviewer A:

Section 3.2 illustrates how using a monosyllabic word model for AND obviates the need for explicit pronunciation modeling.

Table/Figure references have been corrected.

Terms like XWRD, WINT have been removed and substituted by the fully expanded terms, especially in the tables. The definition of the terms ‘logical’ and ‘real’ models has been clarified.

The digit SIX accounts for a large chunk of the errors on digits while using the syllable system. If we analyze the errors in the syllable and triphone systems without the digit SIX, the syllable system outperforms the triphone system, 5.3% vs. 5.7% WER. So the main reason the syllable system does worse than the triphone system on digits is its performance on the digit SIX. The syllable _s_ih_k_s that is used to model SIX, is a CVCC syllable. Owing to the low bandwidth telephone quality data being used, the consonants in this case are easily confusable. Cross-word triphone models do a better job at modeling this situation.

The theoretical part advocating the advantages of using syllables and disadvantages of using phone based systems has been expanded in Section 1.

Reply to comments from reviewer B:

The syllable definition section and ambisyllabic syllable treatment has been expanded.

The motivation for the specific choice of finite duration topology and the associated mathematical formulation has been explained.

The advantages/disadvantages of having a stable acoustic unit have been described.

Though the finite duration topology experiment did not yield significant improvements, it did try to address the high deletion rate in evaluations, which occurs due to long model durations. This technique has been proved to be effective on word models used for digit recognition at TI.

Concerning the issue of not comparing the syllable system to the best triphone system developed at WS’97, the intent of our work is to explore the feasibility of using syllables as recognition units. Since this work is exploratory in nature, it only makes sense to start with a simple system devoid of more complicated features like MLLR, VTN etc. As explained in Section 3, MLLR and VTN were not used in the syllable system and hence it is only fair to compare with a triphone system that did not use these

techniques. In a paper presented at the recent LVCSR Workshop [19], the improvements achieved by systematically adding VTN and MLLR to a baseline system in HTK are clearly shown. MLLR gives about 3% improvement in WER and VTN gives a 1% improvement. This incremental behavior has been shown for several configurations of phone based systems. Since we did not explore cross-word context for syllables, we used a word-internal triphone system as the baseline. The difference between going from cross-word to word-internal is about 3% in terms of WER. Hence, the baseline system we chose to compare to is both a fair and valid data point for the syllable system.

Reply to comments from reviewer C:

The baseline training dictionary was derived from a more complete lexicon created during the LVCSR Summer Workshop ‘96 at JHU, which also had stress markings apart from syllable boundaries and phonemic definitions marked. It is similar to the Pronlex dictionary but with a smaller phone set of 42 phones as compared to 48.

The HTK system was not modified at all to train the syllable system. In fact the training procedure of the syllable and triphone systems has the exact same stages except for stage where context dependency is incorporated.

All grammatical errors have been corrected.

The differences between the various syllable and triphone systems discussed in the paper have been tabulated in Table 10 as suggested. All acronyms that were previously undefined have been either defined or have been replaced by their expanded forms.

All systems are now compared in terms of WER instead of a mixture of accuracy and WER.

It has been made clear in section 7, that the finite-duration topology has been applied to monosyllabic word and syllable models only.