USE OF SYLLABLE NUCLEI LOCATIONS TO IMPROVE ASR

Chris D. Bartels and Jeff A. Bilmes

University of Washington
Department of Electrical Engineering
Seattle, WA 98195
{bartels,bilmes}@ee.washington.edu

ABSTRACT

This work presents the use of dynamic Bayesian networks (DBNs) to jointly estimate word position and word identity in an automatic speech recognition system. In particular, we have augmented a standard Hidden Markov Model (HMM) with counts and locations of syllable nuclei. Three experiments are presented here. The first uses oracle syllable counts, the second uses oracle syllable nuclei locations, and the third uses estimated (non-oracle) syllable nuclei locations. All results are presented on the 10 and 500 word tasks of the SVitchboard corpus. The oracle experiments give relative improvements ranging from 7.0% to 37.2%. When using estimated syllable nuclei a relative improvement of 3.1% is obtained on the 10 word task.

Index Terms— Automatic speech recognition, dynamic Bayesian networks, syllables, speaking rate

1. INTRODUCTION

Conventional automatic speech recognition systems based on a Hidden Markov Model (HMM) use a tweak factor that penalizes the insertion of words. Without this factor, known as the word insertion penalty (WIP), most recognizers will incorrectly insert a large number of words, many of which have unrealistically short durations. The WIP clearly has an effect on decoded word durations, but it is a single parameter that stays the same regardless of any variation in the rate of speech, the length of words, or any changes in the acoustics. There are a few reasons why such a penalty is necessary. First, the duration model in a typical recognizer is quite weak. It consists of a transition probability for each state in the pronunciation, making the duration distribution a sum of geometric models with a (short) minimum duration of one frame per state. The state transition probability has a small dynamic range and no memory of how long the model has been in the current state. Although the duration model allows for longer words, the acoustic model, which is applied every 10 milliseconds, has a relatively large dynamic range and an acoustic match can overwhelm the scores given by the transition probabilities. The WIP is a balancing value, independent of both the word and the acoustics, that lowers the probability of sentence hypotheses that have too many short words over the duration of the utterance. Second, the acoustic observation variables are independent of past and future observation variables given their corresponding state, so acoustic cues can only affect duration and segmentation via the scoring of individual sub-phone states. Standard recognition features use a time window that is only 25 milliseconds, and when longer time scale features (such as [1]) are used they are often appended to the standard observation vector and, again, can only change the segmentation via the acoustic match to the a sub-phone state. In a typical system, the transition probabilities themselves have no direct relation to the acoustics of an individual utterance.

The first goal of this work is to enhance the standard model in a novel way with additional state to better model word duration. The second goal is to use long time scale features to influence duration and segmentation directly, without having to “pass through” a sub-phone state variable. The particular acoustic queues used are estimates of syllable nuclei locations derived from a spectral correlation envelope [2, 3, 4]. A dynamic Bayesian network (DBN) is used to integrate a state variable that counts syllable nuclei with a traditional recognizer (that uses a WIP).

The use of syllable information in automatic speech recognizers has been a topic of research in the past. The syllable was proposed as a basic unit of recognition as early as 1975 [5]. In [6], the utterances were segmented via syllable onset estimations as a precursor to template matching, and in [7] syllables were employed as the basic recognition unit in an HMM . The most closely related method to this paper was presented by Wu in 1997 [8, 9]. In that work, syllable onsets are detected by a neural network classifier, and this information is then used to prune away hypotheses in a lattice. In [10], a standard phone based recognizer is fused with a syllable based recognizer using asynchronous HMMs that are fused together at dynamically located ”recombination states”, and

978-1-4244-1746-9/07/$25.00 ©2007 IEEE

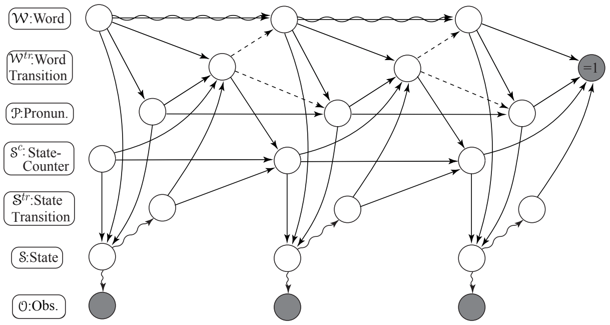

Fig. 1. Baseline Model [13, 14]. This is a standard speech HMM represented as a DBN. Hidden variables are white while observed variables are shaded. Straight arrows represent deterministic relationships, curvy arrows represent probabilistic relationships, and dashed arrows are switching relationships.

in [11, 9] phone and syllable based recognizers are combined using N-Best lists. Syllable nuclei estimates via a spectral correlation measure were first used to estimate speaking rate (one of the 3 measures in *mrate*) [2]. This idea was expanded on by Wang to include temporal correlation and a number of other improvements [3, 4], and this is the method employed in this work. Wang used this detection method in [12] to create speaking rate and syllable length features for automatic speech prominence detection.

This work does not attempt to use syllables as a recognition unit. All models in this paper use a phone based recognizer with a 10 millisecond time frame. This basic recognizer is then supplemented with information about syllable nuclei (rather than onsets), and this information uses a DBN to influence the probabilities in first pass decoding (rather than pruning segmentations in a lattice). Three experiments are presented in this paper. The first is an oracle experiment that requires the total number of syllables in the decoded hypothesis be equal to the total number of syllables in the reference hypothesis. The second experiment also uses oracle information. It generates simulated syllable nuclei locations using the reference hypotheses, and each individual decoded word must contain the correct number of syllables within its time boundary. Finally, the last experiment is performed using syllable nuclei estimated from the acoustics.

2. MODELS AND EXPERIMENTS

All experiments were performed on the 10 and 500 word tasks of the SVitchboard corpus [15]. SVitchboard is a subset of Switchboard I [16] chosen to give a small, closed vocabulary. This allows one to experiment on spontaneous continuous speech, but with less computational complexity and experiment turn-around time than true large vocabulary recognition. The A, B, and C folds were used for training, in the 10 word experiments the D fold was used as the development-test set, in the 500 word experiments the D_short fold was the development-test set, and for both tasks E was used as the evaluation set. All models were trained and decoded using The Graphical Models Toolkit (GMTK) [17].

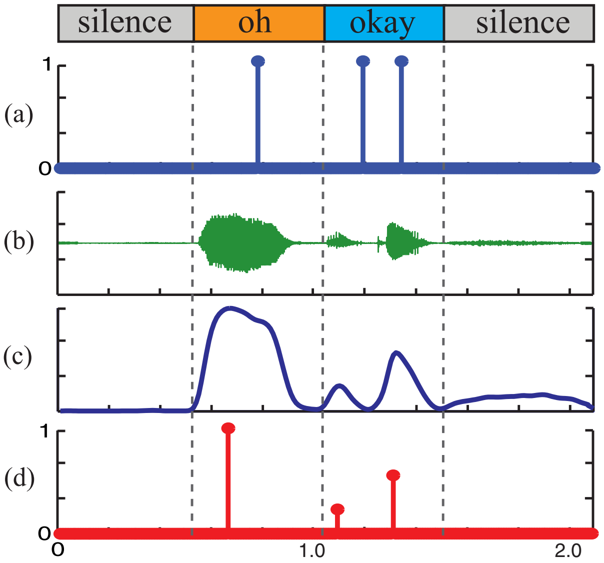

Fig. 2. Illustration of syllable nuclei features. (a) word level oracle features, these are binary features evenly spaced within the word boundary, (b) acoustic waveform, (c) correlation envelope, (d) local maxima of the correlation envelope are potential syllable nuclei (maxima in silence and unvoiced regions are removed)

The baseline systems are HMMs implemented using the DBN shown in Figure 1. This DBN and the baseline systems were developed in [13]. For more on DBNs in automatic speech recognition see [18, 14]. The 10 word experiments used three state left-to-right monophone models, and the 500 word experiments used state clustered within-word triphones with the same topology. The features are 13 dimensional PLPs normalized on a per conversation side basis along with their deltas and double-deltas. The language model scale and penalty were determined using a grid search over the development test set. Grid searches were performed separately for the 10 and 500 word experiments.

2.1. Oracle Experiments

An important part of all of the experiments is the mapping from a word and pronunciation to the number of syllables. This is determined by counting the number of vowels in the pronunciation (we call this the canonical number of syllables). Although this definition matches human intuition for most words, the precise definition of a syllable is not universally agreed upon. For some words the number of syllables is not clear, especially when several vowels appear consecutively and when vowels are followed by glides. For example, one could reasonably argue for either two or three syllables in the word "really" when pronounced "r iy ax l iy". Fortunately we do not need to know the "true" definition of syllable, we only need a mapping that is consistent with the output of our signal processing.

The first oracle experiment, called *Utterance Level*, uses the DBN in Figure 3. This DBN will only decode hypotheses that have the same total number of syllables as the reference hypothesis. The portion of the graph below the "Word" variable remains the same as the baseline, and all the trained parameters from the baseline model are used unchanged. The variable "Word Syllables", $\mathcal{S}^w$, gives the number of canonical syllables in the given word/pronunciation combination. At each word transition the value of $\mathcal{S}^w$ is added to the variable "Syllable Count", $\mathcal{S}^c$. Hence, in the last frame $\mathcal{S}^c$ contains the total number of canonical syllables in the hypothesis. The variable "Count Consistency", $\mathcal{C}^c$, only occurs in the last frame and is always observed to be equal to the oracle syllable count and simultaneously is defined to be equal to $\mathcal{S}^c$. This forces all hypotheses that have a different total number of syllables than the oracle syllable count to have probability zero. Another way of viewing this is that it creates a constraint on the allowed hypotheses, and this constraint is that all decoded sentences must have the same total number of syllables as the oracle syllable count. Because some words have more than one pronunciation, and each pronunciation might have a differing number of syllables, a forced alignment is used to obtain the oracle syllable count for each acoustic utterance. The lower part of the model still requires a language model scale and penalty, and these are again determined with a grid search on the development set. The scale and penalty are optimized for this DBN separately from the baseline experiments, and different values were learned for the 10 and 500 word experiments.

The second oracle experiment is known as *Word Level* and uses the DBN given in Figure 4. In this DBN, each individual word is forced to have the correct number of syllable nuclei somewhere within its time boundary. Note that since this is based only on a count, there is flexibility in the exact placement in time of the syllable centers. Thus, the location information is used, but it does not need to be as precise as methods that segment the utterance based on syllable onsets [6, 8]. The motivation for this is that the exact placement of the onset may not be well defined due to coarticulation [19]. The first step in this experiment was to create an oracle binary observation stream, where at each frame a 1 indicates a syllable nuclei and a 0 otherwise. This observation stream is created by taking each time-aligned reference word and evenly spacing the correct number of ones within its time boundary. An example oracle observation stream is given in Figure 2(a). The word "oh" has one syllable, so there is a single 1 placed in the center of the word. The word "okay" has two syllables, so there are two 1 features evenly spaced across this word. In the DBN, this observation stream is used to set the value of the "Syllable Nuclei", $\mathcal{S}^n$, variable in each frame. Again, "Word Syllables" ($\mathcal{S}^w$) refers to the number of canonical syllables for the given word/pronunciation. The variable "Syllable Count", $\mathcal{S}^c$, keeps track of the number of syllable centers seen since the last word transition. Finally, whenever a word transition occurs "Count Consistency", $\mathcal{C}^c$, gives zero probability to any word hypothesis that does not contain the canonical number of syllable centers. Again, a forced alignment was done to determine the number of canonical syllables in each word and pronunciation , and a grid search determines the language model scale and penalty.

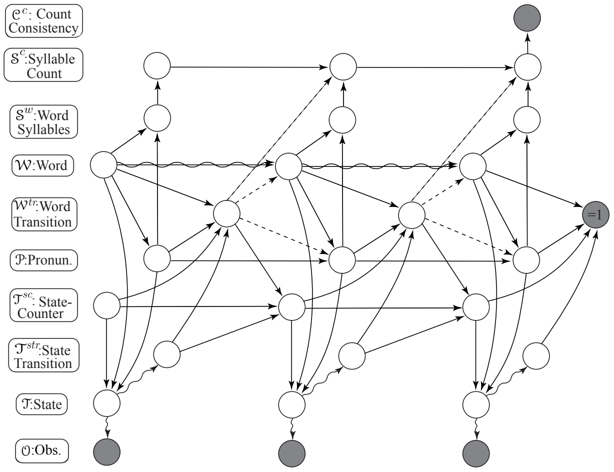
$\mathcal{C}^c$: Count Consistency
$\mathcal{S}^c$:Syllable Count
$\mathcal{S}^w$:Word Syllables
$\mathcal{W}$:Word
$\mathcal{W}^{tr}$:Word Transition
$\mathcal{P}$:Pronun.
$\mathcal{T}^{sc}$: State-Counter
$\mathcal{T}^{str}$:State Transition
$\mathcal{T}$:State
$\mathcal{O}$:Obs.
=1

Fig. 3. Utterance Level decoder with oracle syllable count (see Figure 1 for key). This DBN only allows hypotheses with the same total number of syllables as the reference transcription.

2.2. Use of Estimated Syllable Nuclei

In the third and final experiment, known as *Estimated Word Level*, the oracle syllable nuclei locations used in *Word Level* are replaced with soft estimations of nuclei locations. As will be discussed in Section 3, the oracle *Word Level* graph outperforms the oracle *Utterance Level* graph so an analogous estimated utterance level experiment was not performed. Before this DBN is presented, the feature extraction process is described. This process was given by Wang in [3, 4]. First, a 19 band filter is applied to the waveform, and the 5 bands with the most energy are selected. This filterbank uses two second-order section Butterworth band-pass filters centered at the following frequencies in Hertz: 240, 360, 480, 600, 720, 840, 1000, 1150, 1300, 1450, 1600, 1800, 2000, 2200, 2400, 2700, 3000, 3300, and 3750. Temporal correlation is performed on the selected five bands followed by spectral correlation. The resulting signal is then smoothed using a Gaussian window. An example correlation envelope can be seen

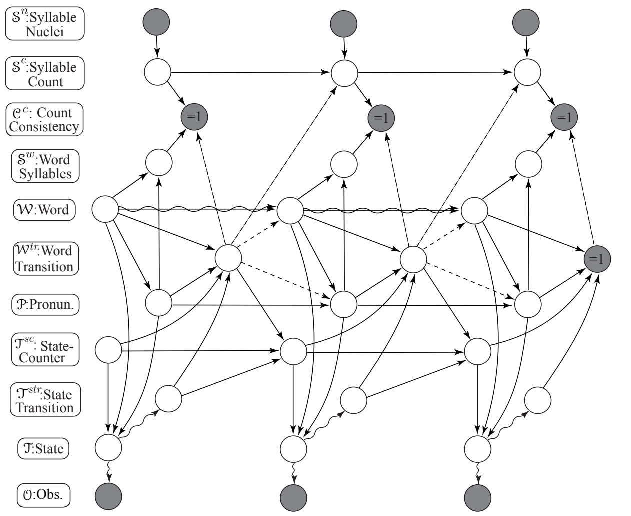
$\mathcal{S}^n$:Syllable Nuclei
$\mathcal{S}^c$:Syllable Count
$\mathcal{C}^c$: Count Consistency
$\mathcal{S}^w$:Word Syllables
$\mathcal{W}$:Word
$\mathcal{W}^{tr}$:Word Transition
$\mathcal{P}$:Pronun.
$\mathcal{T}^{sc}$: State-Counter
$\mathcal{T}^{str}$:State Transition
$\mathcal{T}$:State
$\mathcal{O}$:Obs.
=1

Fig. 4. Word Level decoder with oracle syllable nuclei (see Figure 1 for key). This graph only allows word hypotheses that are consistent with the oracle syllable nuclei locations.

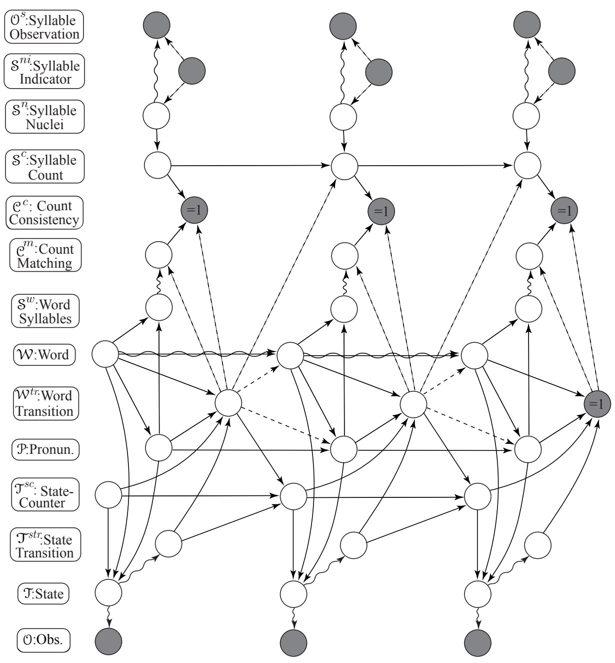
$\mathcal{O}^s$:Syllable Observation
$\mathcal{S}^{ni}$:Syllable Indicator
$\mathcal{S}^n$:Syllable Nuclei
$\mathcal{S}^c$:Syllable Count
$\mathcal{C}^c$: Count Consistency
$\mathcal{C}^m$:Count Matching
$\mathcal{S}^w$:Word Syllables
$\mathcal{W}$:Word
$\mathcal{W}^{tr}$:Word Transition
$\mathcal{P}$:Pronun.
$\mathcal{T}^{sc}$: State-Counter
$\mathcal{T}^{str}$:State Transition
$\mathcal{T}$:State
$\mathcal{O}$:Obs.
=1

Fig. 5. Word Level decoder with estimated syllable nuclei (see Figure 1 for key). This DBN estimates the number of syllable nuclei in each word and models the probability that this estimate matches the word hypothesis.

in Figure 2(c). The next step is to find the local minima and maxima of the correlation envelope. The height of each minimum is subtracted from the maximum that follows it, and the resulting maxima heights are normalized by height of the largest peak in the utterance. This method can produce spurious peaks in non-speech and unvoiced regions, so a pitch detector is applied to the waveform and all peaks corresponding to unvoiced and non-speech segments are removed. In [4], it was reported that this method correctly placed nuclei in 80.6% of the syllables in a hand transcribed test set. In [3, 4], peaks that fall below a minimum threshold are rejected and the result is a binary feature. For our experiments we do not make a hard decision, instead we retain all the maxima points and use the actual height value as a feature. This allows us to make a soft decision on if a particular local maximum is a syllable center, with a lager value indicating a higher probability. An example of the resulting features can be seen in Figure 2(d).

We now have features for estimating syllable nuclei and can move to the discussion of the *Estimated Word Level* DBN, as seen in Figure 5. The variable “Syllable Indicator”, $\mathcal{S}^{ni}$, is a binary feature indicating if the current frame is a local maximum in the correlation envelope, “Syllable Observation”, $\mathcal{O}^s$, is the magnitude of the local maximum, and “Syllable Nuclei”, $\mathcal{S}^n$, is a hidden variable that decides if the current frame is or is not a syllable nuclei. When $\mathcal{S}^{ni}$ is “false” it indicates that we are not at a local maximum in the correlation curve, and $\mathcal{S}^n$ is forced to be false and $\mathcal{O}^s$ has no bearing on the probability. When $\mathcal{S}^{ni}$ is “true”, we are at a maximum and there is a potential syllable nuclei in the frame. In this case, $\mathcal{S}^n$ is true with probability $p(\mathcal{S}^n = true)p(\mathcal{O}^s|\mathcal{S}^n = true)$ and false with probability $p(\mathcal{S}^n = false)p(\mathcal{O}^s|\mathcal{S}^n = false)$, where $p(\mathcal{S}^n = true)$ and $p(\mathcal{S}^n = false)$ are discrete probabilities and $p(\mathcal{O}^s|\mathcal{S}^n = true)$ and $p(\mathcal{O}^s|\mathcal{S}^n = false)$ are single dimensional Gaussians. This is implemented by making $\mathcal{S}^{ni}$ a “switching parent” [17] of $\mathcal{O}^s$ and $\mathcal{S}^n$, meaning $\mathcal{S}^{ni}$ controls the choice of its children’s’ distributions but does not appear as a conditioning variable in their conditional probability tables (CPTs).

As in the oracle *Word Level* model, the variable “Syllable Count” ($\mathcal{S}^c$) counts the number of syllable centers since the last word transition and “Word Syllables” ($\mathcal{S}^w$) is the number of canonical syllables. The variable “Count Consistency”, $\mathcal{C}^c$, forces “Count Matching”, $\mathcal{C}^m$, to be equal to $\mathcal{S}^c$ at word transitions. $\mathcal{C}^m$ and its CPT, $p(\mathcal{C}^m|\mathcal{S}^w)$, are the probabilistic glue between the phone recognizer and syllable counting stream. In the oracle experiment, the value of $\mathcal{S}^c$ equals the value of $\mathcal{S}^w$ with probability 1. In the estimated DBN, $p(\mathcal{C}^m|\mathcal{S}^w)$ gives a distribution where (ideally) these two values have a high probability of matching. This CPT along with $p(\mathcal{S}^n)$ and the two Gaussians ($p(\mathcal{O}^s|\mathcal{S}^n)$) are trained using EM, while the parameters of the phone recognizer are held fixed with their values from the baseline. In the 10 word case, a four dimensional grid search was performed over the language model scale, the language model penalty, a scaling factor for $p(\mathcal{C}^m|\mathcal{S}^w)$, and a scaling factor for the Gaussians $p(\mathcal{O}^s|\mathcal{S}^n)$. The scaling factor for the Gaussians did not improve results, so only a three dimensional grid search was done in the 500 word case.

		10 Word Vocabulary								500 Word Vocabulary							
		Dev				Eval				Dev				Eval			
		S	D	I	WER	S	D	I	WER	S	D	I	WER	S	D	I	WER
Baseline		183	86	25	18.1%	187	130	16	19.6%	585	234	157	53.2%	6815	3122	1803	58.6%
Oracle	Utterance Level	175	24	5	12.5%	189	39	3	13.6%	655	140	87	48.1%	6927	2935	1046	54.5%
	Word Level	178	8	2	11.5%	188	18	4	12.3%	628	50	41	39.2%	7418	913	603	44.6%
Estimated	Word Level	174	77	31	17.3%	180	125	18	19.0%	583	233	153	52.8%	6824	3114	1798	58.6%

Table 1. Table of Results. S, D, and I are counts of substitutions, deletions, and insertions. WER is percent word error rate.

3. RESULTS

The results for all experiments are given in Table 1. The 500 word baseline system has a small improvement over the results presented in [13]. Note that systems that train with additional data outside the designated SVitchboard training sets have reported lower word error rates [13, 20].

The *Utterance Level* oracle DBN gives a substantial improvement over the baseline. The improvement is much larger in the 10 word case than in the 500. The first reason for this is that the utterances in the 10 word data set are shorter than in the 500 word set, and when the syllable count is larger more valid hypotheses are possible. Second, the "Syllable Count" state variable needs to be quite large in the 500 word set and this makes decoding more difficult and more susceptible to search errors. The word error rate improvement comes in the form of a reduction in deletions and insertions, but with a rise in substitutions. The primary cause of increased substitutions is the case when the baseline hypothesis has a deletion and the oracle constraint forces the addition of a word which is incorrect.

The *Word Level* oracle DBN performs better than the *Utterance Level* DBN in both the 10 and 500 word vocabulary systems. This gives us two pieces of information. First, the location of the syllable nuclei is of more use than having only the syllable count. Second, it tells us that if we had perfect syllable detection and a perfect match from detection to the words, we could see a substantial word error rate improvement. On caveat with this experiment is that the oracle syllable centers are evenly spaced which may not always be indicative of the true locations. One can conceive of a case where a simulated center of a two syllable word is so far off that two one syllable words would not align correctly. Having the centers in locations more consistent with the acoustics could increase the confusability in such a case.

The *Estimated* syllable nuclei DBN gave a substantial result on the 10 word system, but its performance was similar to the baseline on the 500 word task. This experiment is successful at lowering deletions and substitutions, but has less impact on insertions. The problem in the oracle graphs where deletions are changed to substitutions does not occur often because the matching between the syllable count and word hypothesis is soft, and the removal of the deletion will not happen unless the acoustics in the word recognizer supports this. The reason that there is no improvement on the 500 word task is likely because the syllable nuclei detection is working much better on the short and isolated words that predominate the 10 word system. In the 500 word system the entropy of $p(\mathcal{C}^m|\mathcal{S}^w = x)$ for $x = 0...4$ is $0.04, 1.00, 1.27, 1.54$, and 1.50. This is evidence that the more syllables there are in a word, the more difficulty our system has detecting the proper number.

	10 Words		500 Words	
	Full	Reduced	Full	Reduced
Baseline	19.6%	19.9%	58.6%	59.7%
Estimated Word Level	19.0%	19.2%	58.6%	59.7%

Table 2. Results are % WER. Full is full eval set (as in Table 1), reduced is the eval set with the STP data removed

There is one possible caveat about the above experimentation that still needs to be addressed here, namely that in the development of the syllable nuclei features in [3, 4] the parameters were tuned using data from the Switchboard Transcription Project (STP) [21], and some STP data is included in our test set. In this last set of results we run an experiment that controls for this and shows that our results still hold. Table 2 gives baseline and estimated results for the "Reduced" test set, which contains the SVitchboard E fold minus any speech from any speaker included in the STP data. This set is approximately 80% of the full test set. Note that the relative differences between the baseline and *Estimated Word Level* results are approximately the same.

4. CONCLUSION

The oracle experiments present empirical evidence that syllable nuclei locations have the potential to give large word error rate improvements in automatic speech recognition. In our experiments with estimated syllable centers, an improvement was seen in the 10 word task but no performance gain was seen on the longer words and utterances found in the 500 word task. There are many possible directions for improving the results without oracle information. First, additional features for detecting syllables derived from differing signal processing methods could be employed. The simple counts could be replaced by a more sophisticated recognition stream where syllable onsets are also considered. Another direction is that instead of using the canonical number of syllables, the mapping of words to the number of detected syllables could be learned. This mapping could make use of individual syllable identities as well as their contexts. Finally, additional ways of modeling the mismatch between the detection and prediction scheme could be employed. In particular, the detection could be matched after each individual syllable instead of after each word. Given the potential gain seen in the oracle experiments and the encouraging results with estimated nuclei, all of these directions will be pursued.

Acknowledgments

This work was supported by ONR MURI grant N000140510388 and by NSF grant IIS-0093430. Thanks to Dagen Wang and Shrikanth Narayanan for providing us with the code for syllable nuclei estimation.

5. REFERENCES

[1] H. Hermansky and S. Sharma, "TRAPs - Classifiers of temporal patterns," in *Proc. of the Int. Conf. on Spoken Language (ICSLP)*, 1998.

[2] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.

[3] D. Wang and S. Narayanan, "Speech rate estimation via temporal correlation and selected sub-band correlation," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[4] D. Wang and S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Speech, Audio and Language Processing*, 2007.

[5] O. Fujimura, "Syllable as a unit of speech recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 82–87, Feburary 1975.

[6] M.J. Hunt, M. Lennig, and P. Mermelstein, "Experiments in syllable-based recognition of continuous speech," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1980.

[7] P.D. Green, N. R. Kew, and D. A. Miller, "Speech representations in the SYLK recognition project," in *Visual Representation of Speech Signals*, Martin Cooke, Steve Beet, and Malcolm Crawford, Eds., chapter 26, pp. 265–272. John Wiley & Sons, 1993.

[8] Su-Lin Wu, M. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1997.

[9] Su-Lin Wu, *Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition*, Ph.D. thesis, University of California, Berkeley, Spring 1998.

[10] S. Dupont, H. Bourlard, and C. Ris, "Using multiple time scales in a multi-stream speech recognition system," in *Proc. of the European Conf. on Speech Communication and Technology*, 1997.

[11] Su-Lin Wu, E.D. E.D. Kingsbury, N. Morgan, and S. Greenberg, "Incorporating information from syllable-length time scales into automatic speech recognition," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.

[12] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Transactions on Speech, Audio and Language Processing*, vol. 15, no. 2, pp. 690–701, Feb 2007.

[13] K. Livescu et al., "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[14] J. Bilmes and C. Bartels, "A review of graphical model architectures for speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 89–100, September 2005.

[15] S. King, C. Bartels, and J. Bilmes, "SVitchboard: Small-vocabulary tasks from switchboard," in *Proc. of the European Conf. on Speech Communication and Technology*, 2005.

[16] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992.

[17] J. Bilmes, *GMTK: The Graphical Models Toolkit*, 2002.

[18] G. Zweig, *Speech Recognition with Dynamic Bayesian Networks*, Ph.D. thesis, University of California, Berkeley, Spring 1998.

[19] A. Subramanya, C. Bartels, J. Bilmes, and P. Nguyen, "Uncertainty in training large vocabulary speech recognizers," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2007.

[20] Ö. Çetin et. al., "An articulatory feature-based tandem approach and factored observation modeling," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[21] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the switchboard corpus," in *Proc. of the Int. Conf. on Spoken Language (ICSLP)*, 1996.