INTEGRATING SYLLABLE BOUNDARY INFORMATION INTO SPEECH RECOGNITION

Su-Lin Wu, Michael L. Shire, Steven Greenberg, Nelson Morgan
International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704-1198, USA
University of California at Berkeley, Berkeley, CA 94720, USA
{sulin, shire, steveng, morgan}@icsi.berkeley.edu

ABSTRACT
In this paper we examine the proposition that knowledge of the timing of syllabic onsets may be useful in improving the performance of speech recognition systems. A method of estimating the location of syllable onsets derived from the analysis of energy trajectories in critical band channels has been developed, and a syllable-based decoder has been designed and implemented that incorporates this onset information into the speech recognition process. For a small, continuous speech recognition task the addition of artificial syllabic onset information (derived from advance knowledge of the word transcriptions) lowers the word error rate by 38%. Incorporating acoustically-derived syllabic onset information reduces the word error rate by 10% on the same task. The latter experiment has highlighted representational issues on coordinating acoustic and lexical syllabifications, a topic we are beginning to explore.

1. INTRODUCTION
Automatic speech recognition (ASR) systems typically rely upon phoneme- or sub-phoneme-based Hidden Markov models (HMMs) that are concatenated into word and sentence elements. Although syllable-based speech recognition has been successfully used in several languages (including Spanish [1] and Chinese [2]), the syllable has not been fully exploited for the automatic recognition of English. In this paper we investigate the possibility that syllabic onsets can be derived from the acoustic speech signal, and that this onset information can be incorporated into the decoding process in a manner sufficient to improve recognition performance.

Evidence from both psychoacoustic and psycholinguistic research [3, 4, 5], as well as by model as one of the authors [6], suggests that the syllable is a basic perceptual unit for speech processing in humans. The syllable was proposed as a basic unit of automatic (computer) speech recognition as early as 1975 [7, 8], and this idea has been periodically re-examined (e.g. in [9, 10, 11, 12, 13]). The syllabic level confers several potential benefits; for one, syllabic boundaries are more precisely defined than phonetic segment boundaries in both the speech waveform and in spectrographic displays. Additionally, the syllable may serve as a natural organizational unit useful for reducing redundant computation and storage in decoding. The syllabic abstraction may also be appropriate for the incorporation of suprasegmental prosodic information.

English is considered to possess a highly complex syllabic structure not readily amenable to automatic segmentation or identification. Detailed statistical analyses of sponta-neous informal discourse indicate that the syllabic structure of conversational English is not as complicated as has been generally supposed. For example, data gathered from telephone conversations in [14] and the Switchboard corpus [15, 16] indicate that over 80% of the word tokens in these corpora are monosyllabic, and more than 85% of the syllables are of the canonical consonant-vowel (CV), vowel-consonant (VC), V, or CVC varieties. These structural regularities can, in principle, be exploited to reliably estimate syllabic boundaries.

Previous research on detecting syllable boundaries and using this information to improve recognition accuracy is reported for English [8, 9, 10] and for German [12, 13]. In this communication we describe a perceptually-oriented method for the automatic delineation of syllabic onsets. Artificial neural networks (NNs) are used to classify both phonetic segments and potential syllabic onsets. In a departure from previous research, we focus on continuous, naturally-spoken English.

2. DETECTING SYLLABLE ONSETS
Syllable onsets are typically characterized by a pattern of synchronized rises in subband energy spanning adjacent subbands. The time course of these coordinated rises and falls in energy correspond to syllable-length intervals, on the order of 100-250 ms.

Figure 1 illustrates the signal processing procedures designed to enhance and extract these acoustic properties. The speech waveform is decomposed via short-time Fourier analysis into a narrow-band spectrogram, which is convolved with both a temporal and a channel filter, effectively creating a two-dimensional filter. The temporal filter (a high-pass filter analogous to a Gaussian derivative) smoothes and differentiates along the temporal axis, and is tuned for enhancing changes in energy on the order of 150 ms. The (Gaussian) channel filter performs a smoothing function across the channels, providing weight to regions of the spectrogram where adjacent channels are changing in coordinated fashion. Half-wave rectification is used to preserve the positive changes in energy, thus emphasizing the syllable onsets.

Large values in this representation correspond to positive-
4. RECOGNITION EXPERIMENTS

Recognition experiments were performed on a subset of the OGI Numbers corpus [24]. This corpus contains continuous, naturally spoken utterances of many different speakers saying numbers from a vocabulary of thirty words. A sample utterance from the database is “eighteen thirty one.” The example in Figure 2 is also derived from the Numbers corpus. Collected over telephone lines, the utterances exhibit large variations in speaking rate and acoustic environmental conditions. The subset includes approximately three hours (3500 sentences) of training data, and one hour (1200 sentences) each of development-test set and final-test set data. The training data, with its cross-validation subset, was used for tuning the parameters. The development test set (referred to as the “test set” in the sections below) was used for the results reported below.

4.1. Experiments with Syllabic Onsets Determined from Forced-Viterbi Alignment

In order to ascertain the potential value of syllabic onset timing, this information (derived from advance knowledge of the word-transcriptions of the test utterances) was incorporated into the decoding process.

A forced-Viterbi technique was used to generate phone alignment labels based on word transcriptions of the corpus provided for all the utterances in the test set. Artificial syllabic onsets were derived from these forced-Viterbi labels. The resulting syllabic onset information was only approximate. Many of the onsets were as much as 50 ms distant from the labelled segment boundary.

The experimental lexicon included 32 single-pronunciation words, comprising 30 different syllables. The pronunciations were derived from those developed at Carnegie Mellon University for large vocabulary recognition. The context-dependent phonetic durations used were derived from the training data using an embedded training process.

The recognition procedure used a highly restrictive criterion for syllabic decoding. A syllable was presumed to occur only when the beginning frame for the syllabic model coincided precisely with a predetermined onset. No restriction was placed on a syllable’s termination; it was theoretically possible for the end point of a postulated syllable to occur after the next Viterbi-derived onset of the following syllable. Only syllabic onset information from the test set was included in our recognition experiments. No prior knowledge of phonetic information from the test set was used.

If the dynamic programming (Viterbi) procedure and the speech decoding input elements were of the ideal form, the addition of artificially-derived syllabic boundary information would, in theory, provide little or no improvement in recognition performance. In principle the decoding process assumes that models can begin at any frame, including the ones we specified as incorporating syllabic onsets. In our experiment, incorporation of artificially-derived syllable segmentation information reduces the word error rate by 38%, from 10.8% to 6.7%, as shown in Table 1. This large reduction in word error suggests that syllabic boundary information can significantly improve speech recognition performance when directly incorporated into the decoding process.

A second series of experiments was conducted with the aim of delineating the precision required for syllabic onset information to be of significant utility in the decoding process. The temporal precision of the syllabic onset was systematically varied over several frames, as shown for selected values in Table 2. There is a small, but significant
Table 1. Word-error rates for decoding using a single-pronunciation lexicon, with and without artificial syllabic onsets derived from forced alignment.

<table>
<thead>
<tr>
<th>System</th>
<th>Word Error Rate sub./ins./del.</th>
</tr>
</thead>
<tbody>
<tr>
<td>no onset information</td>
<td>10.8%/5.8%/3.1%/1.8%</td>
</tr>
<tr>
<td>with known syllable onset times</td>
<td>6.7%</td>
</tr>
<tr>
<td>Total frs./onset = 1</td>
<td>4.4%/0.7%/1.6%</td>
</tr>
</tbody>
</table>

Table 2. Word-error rates for single-pronunciation decoding, using syllable hypotheses that are allowed to begin within several frames of artificial onsets derived from forced alignment.

<table>
<thead>
<tr>
<th>Number of frames about each onset</th>
<th>Error Rate sub./ins./del.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total frs./onset = 5 centered on onset</td>
<td>7.3%</td>
</tr>
<tr>
<td></td>
<td>4.9%/0.9%/1.3%</td>
</tr>
<tr>
<td>Total frs./onset = 9 centered on onset</td>
<td>7.8%</td>
</tr>
<tr>
<td></td>
<td>5.1%/1.3%/1.4%</td>
</tr>
<tr>
<td>Total frs./onset = 13 centered on onset</td>
<td>8.5%</td>
</tr>
<tr>
<td></td>
<td>5.2%/1.9%/1.4%</td>
</tr>
</tbody>
</table>

Table 3. Word-error rates for multiple-pronunciation (data-derived) decoding, with and without acoustically-derived onsets.

<table>
<thead>
<tr>
<th>System</th>
<th>Error Rate sub./ins./del.</th>
</tr>
</thead>
<tbody>
<tr>
<td>with data-derived lexicon</td>
<td>9.1%</td>
</tr>
<tr>
<td>no onset information</td>
<td>5.3%/1.3%/2.4%</td>
</tr>
<tr>
<td>with data-derived lexicon</td>
<td>8.2%</td>
</tr>
<tr>
<td>onset used with threshold only</td>
<td>4.8%/1.3%/2.1%</td>
</tr>
</tbody>
</table>

5. DISCUSSION

The experiments described in the section above illuminated certain limitations in the present recognition system that necessarily impact its performance. One such limitation of the current experimental paradigm pertains to the mismatch between the acoustic-phonetic and phonological representations of the syllable forms used for word recognition. The syllabic segmentation method was based largely on acoustic-phonetic criteria, while the syllabic combination of lexical items was derived from a more abstract phonological representation. An instance where this distinction is of particular significance for word sequences is one in which the syllable coda of the first word is consonantal and the onset of the following word is vocalic, as in "five eight." The phonological representation of such a sequence would be /fayv/ /eyv/, while the phonetic realization is more typically [fay] [vey]. Such "re-syllabification" phenomena are not easily accommodated within the present syllabic representational framework. Future efforts will be devoted to resolving such issues within a single, coherent representational framework.

6. SUMMARY AND FUTURE WORK

Incorporation of information pertaining to syllabic onsets has the potential to significantly increase the accuracy of word-level recognition. This syllabic information was obtained in our study from two different sources — artificial boundaries derived from prior phonetic transcriptions of the corpus materials, and acoustic segmentation derived from a signal processing method based on general principles of auditory analysis. The word-error rate was reduced by 38% for the artificially-derived boundaries and by 10% for the boundary information derived from the acoustic segmentation method. These results indicate the potential utility of incorporating syllable boundary information in future speech recognition systems. We are now working towards improving the accuracy of the acoustically-based segmentation algorithm via the incorporation of the computed probability estimates from the neural net and through optimization of the decision criterion derived from such signal.
detection theoretic parameters as the false alarm rate and response bias.

7. ACKNOWLEDGMENTS

We thank Dan Gildea for developing the data-derived pronunciations and gratefully acknowledge valuable assistance from Eric Fosler and Dan Ellis. The automatic syllabification program we used, *tsylb2*, was written by Bill Fisher of NIST.

This material is based upon research supported by the following funding sources: a National Science Foundation Graduate Research Fellowship (SW), Joint Services Electronics Program grant (SW, MS), Contract Number F49620-94-C-0038 and a DOD subcontract from the Oregon Graduate Institute. Additional support was received from the Faculté Polytechnique de Mons as part of a European Community Basic Research grant (Project Sprach). Finally, we gratefully acknowledge continued support from the International Computer Science Institute.

REFERENCES


