Presented at ICASSP-97, Munich, vol. 2 pp. 987-990.

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 recognition has been successfully used in several languages (including Spanish [1] and Chinese [2]), the syllable has been 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 psycholinguistical research [3, 4, 5], as well as a model by 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 reexamined (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 spontaneous 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.

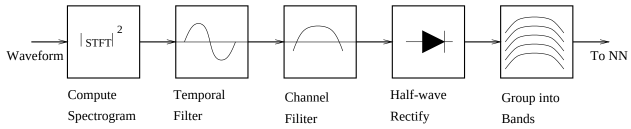

Figure 1. Major processing steps for the syllable onset features.

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-

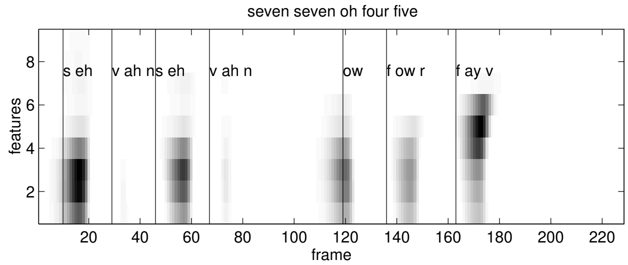

Figure 2. Example of onset features derived for the utterance 'seven seven oh four five'. The vertical lines denote syllable onsets as derived from hand-transcribed phone labels.

going energy regions where hypothesized syllable onset characteristics occur. The channel outputs are subsequently averaged over a region spanning nine critical bands [17], the result of which is illustrated in Figure 2.

Features derived from this procedure are updated every 10 ms. The resulting vectors are concatenated with log-RASTA [18] features computed over a 25-ms frame every 10 ms, and this combination is used as the input to a neural network classifier for estimating the location of syllabic onsets. A single-hidden layer, fully connected, feed-forward multilayer perceptron with 400 hidden units was trained to estimate the probability that a given frame is a syllable onset, given the acoustic patterns described above. For the purposes of training, the syllable onset (as derived from phonetically transcribed segmentation) was represented as a series of five frames, in which the initial frame corresponded to the actual onset.

A simple numeric threshold applied to the probability estimates generated by the neural net determined the identification of any given frame as a syllabic onset. This procedure correctly detected 94% of the onsets computed from phonetically transcribed data (within the five-frame tolerance window defined for training). The procedure also mistakenly inserted syllabic onsets where there were none (false positives) in 15% of the frames outside the tolerance window of any onset. These onset decisions were used by a syllable-based decoder as frames corresponding to syllable onsets.

3. SYLLABLE-BASED DECODING

A speech decoder was designed to incorporate an intermediate syllabic level of abstraction between the phone and word/sentence tiers. The decoder processes phonetic probabilities from a neural network using a conventional Viterbi algorithm using a bigram syllable grammar and creates a syllable graph (a derivative of the word graph as defined in [19]). The syllable graph serves as input to the program's stack decoder [20, 21], along with a bigram word grammar, to determine the most likely sequence of words. This procedure is a variation on the multiple-pass decoding method (related to the approach used in [22] and [23]) and enables the use of a complex language model at a higher stage of linguistic representation. The additional complexity of the decoder design permits the explicit representation of the relationship of phones to syllables and syllables to words. Syllabic onset information is introduced as an additional probability input into the decoder at the level of the syllable graph.

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

System	Word Error Rate sub./ins./del.
no onset information	10.8% 5.8%/3.1%/1.8%
with known syllable onset times Total frs./onset = 1	6.7% 4.4%/0.7%1.6%

Table 1. Word-error rates for decoding using a single-pronunciation lexicon, with and without artificial syllabic onsets derived from forced alignment.

Number of frames about each onset	Error Rate sub./ins./del.
Total frs./onset = 5 centered on onset	7.3% 4.9%/0.9%/1.5%
Total frs./onset = 9 centered on onset	7.8% 5.1%/1.3%/1.4%
Total frs./onset = 13 centered on onset	8.5% 5.2%/1.9%/1.4%

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.

increase in word error rate as the onset window is increased from one to thirteen frames, consistent with the hypothesis that syllabic onset information of intermediate accuracy is of potential utility in speech recognition systems.

4.2. Experiments with Acoustically Determined Syllabic Onsets

Speech recognition systems do not typically possess detailed *a priori* information concerning the temporal loci of syllabic boundaries. Rather, they must infer the syllable boundaries from other information sources. We are in the initial stages of integrating the acoustically-derived onset information described above into the decoding process.

In order to provide a closer match between the phonetically transcribed material and the syllabic onsets derived from the neural network training procedure, a new set of lexicons and grammars were developed, specifically based on the transcription data from the training set. These materials included 32 words (and their range of 178 possible pronunciations), comprising 118 separate syllabic forms. The spectrum of pronunciations included cover approximately 90% of the pronunciation variations in the corpus, as reflected in the phonetic transcription material. The durations of phonetic segments were also computed from the transcription of the training materials. The word grammar (derived from the word transcriptions of the training set) was identical to the one described for the initial series of recognition experiments in the last section. However, the syllable-level grammar was, by necessity, specifically adapted to this language model set.

The decoder used a simple threshold applied to the output of the neural network in order to ascertain the occurrence of a syllabic onset. The algorithm set temporal restrictions on the syllabic models such that they were required to begin no sooner than five frames preceding the time of the estimated syllabic onset. By this metric it was possible to reduce the number of potential starting frames for syllabic models by 58%.

System	Error Rate sub./ins./del.
with data-derived lexicon no onset information	9.1% 5.3%1.3%2.4%
with data-derived lexicon onset used with threshold only	8.2% 4.8%1.3%2.1%

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

When such acoustically-derived syllabic onset information is incorporated into the decoding process the recognition performance improves slightly. The word error rate decreases by 10% which, while not quite reaching statistical significance (for $p < 0.05$), is still indicative of the potential performance benefit to be derived from including temporal information pertaining to syllabic boundaries.

The incorporation of multiple pronunciations in the recognition lexicon improved the performance of the baseline system and served to provide further details concerning the specific relationship between syllabic boundary information and word models.

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 syllabification 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/ /eyt/, while the phonetic realization is more typically [fay] [veyt]. 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

[1] Antonio Bonafonte, Rafael Estany, and Eugenio Vives. Study of subword units for spanish speech recognition. In *Eurospeech*, volume 3, pages 1607–1610, Madrid, Spain, September 1995. ESCA.

[2] Sung-Chien Lin, Lee-Feng Chien, Keh-Jiann Chen, and Lin-Shan Lee. A syllable-based very-large-vocabulary voice retrieval system for Chinese databases with textual attributes. In *Eurospeech*, volume 1, pages 203–206, Madrid, Spain, September 1995. ESCA.

[3] Dominic W. Massaro. Perceptual units in speech recognition. *Journal of Experimental Psychology*, 102(2):199–208, 1974.

[4] Douglas O'Shaughnessy. *Speech Communication*, chapter 5, pages 164–203. Addison-Wesley Publishing Company, Reading, Massachusetts, 1987.

[5] Juan Segui, Emmanuel Dupoux, and Jacques Mehler. The role of the syllable in speech segmentation, phoneme identification and lexical access. In Gerry Altmann, editor, *Cognitive Models of Speech Processing*, chapter 12, pages 263–280. MIT Press, 1990.

[6] Steven Greenberg. Understanding speech understanding: Towards a unified theory of speech perception. In *Proceedings of the ESCA Workshop (ETRW) on The Auditory Basis of Speech Perception*, pages 1–8, Keele, United Kingdom, July 1996. ESCA.

[7] Osamu Fujimura. Syllable as a unit of speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1):82–87, February 1975.

[8] Paul Mermelstein. Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. Am*, 58(4):880–883, October 1975.

[9] M.J. Hunt, M. Lennig, and P. Mermelstein. Experiments in syllable-based recognition of continuous speech. In *ICASSP*, volume 3, pages 880–883, Denver, Colorado, April 1980. IEEE.

[10] P.D. Green, N. R. Kew, and D. A. Miller. Speech representations in the sylk recognition project. In M. P. Cooke, S. W. Beet, and M. D. Crawford, editors, *Visual Representation of Speech Signals*, chapter 26, pages 265–272. John Wiley, 1993.

[11] Kenneth W. Church. Phonological parsing and lexical retrieval. In Uli H. Frauenfelder and Lorraine Komisarjevsky Tyler, editors, *Spoken Word Recognition*, Cognition Special Issues, chapter 3, pages 53–69. MIT Press, 1987.

[12] W. Reichl and G. Ruske. Syllable segmentation of continuous speech with artificial neural networks. In *Eurospeech*, pages 1771–1774, Berlin, Germany, September 1993.

[13] Katrin Kirchhoff. Syllable-level desynchronisation of phonetic features for speech recognition. In *ICSLP*, volume 4, pages 2274–2276, Philadephia, Pennsylvania, October 1996.

[14] Norman R. French, Charles W. Carter, Jr., and Walter Koenig, Jr. The words and sounds of telephone conversations. *The Bell System Technical Journal*, IX:290–325, April 1930.

[15] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP*, volume 1, pages 517–520, San Francisco, California, March 1992. IEEE.

[16] Steven Greenberg, Joy Hollenback, and Dan Ellis. The Switchboard transcription project. Technical report, International Computer Science Institute, 1997.

[17] Donald D. Greenwood. Critical bandwidth and the frequency coordinates of the basilar membrane. *JASA*, 33:1344–1356, 1961.

[18] Hynek Hermansky and Nelson Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[19] Martin Oerder and Hermann Ney. Word graphs: An efficient interface between continuous-speech recognition and language understanding. In *ICASSP*, volume 2, pages 119–122, Minneapolis, Minnesota, April 1993. IEEE.

[20] Frederick Jelinek. Fast sequential decoding algorithm using a stack. *IBM J. Res. Develop.*, 13:675–685, November 1969.

[21] Steve Renals and Mike Hochberg. Efficient evaluation of the LVCSR search space using the noway decoder. In *ICASSP*, volume 1, pages 149–152, Atlanta, Georgia, May 1996. IEEE.

[22] Frank K. Soong and Eng-Fong Huang. A tree-trellis based fast search for finding the N best sentence hypotheses in continuous speech recognition. In *ICASSP*, volume 1, pages 705–708, Toronto, Canada, May 1991. IEEE.

[23] P. Kenny, R. Hollan, V. Gupta, M Lennig, P Mermelstein, and D. O'Shaughnessy. A^*-admissible heuristics for rapid lexical access. In *ICASSP*, volume 1, pages 689–692, Toronto, Canada, May 1991. IEEE.

[24] Center for Spoken Language Understanding, Department of Computer Science and Engineering, Oregon Graduate Institute. Numbers corpus, release 1.0, 1995.

[25] Godfrey Dewey. *Relative Frequency of English Speech Sounds*, volume 4 of *Harvard Studies in Education*. Harvard University Press, Cambridge, 1923.

[26] Zhihong Hu, Johan Schalkwyk, Etienne Barnard, and Ronald Cole. Speech recognition using syllable-like units. In *ICSLP*, volume 2, pages 1117–1120, Philadephia, Pennsylvania, October 1996.