# INCORPORATING INFORMATION FROM SYLLABLE-LENGTH TIME SCALES INTO AUTOMATIC SPEECH RECOGNITION

*Su-Lin Wu, Brian E. D. Kingsbury, Nelson Morgan and Steven Greenberg*

International Computer Science Institute,
1947 Center Street, Suite 600, Berkeley, CA 94704, USA
and the University of California at Berkeley, Berkeley, CA 94720, USA
{sulin,bedk,morgan,steveng}@icsi.berkeley.edu

## ABSTRACT

Including information distributed over intervals of syllabic duration (100–250 ms) may greatly improve the performance of automatic speech recognition (ASR) systems. ASR systems primarily use representations and recognition units covering phonetic durations (40–100 ms). Humans certainly use information at phonetic time scales, but results from psychoacoustics and psycholinguistics highlight the crucial role of the syllable, and syllable-length intervals, in speech perception. We compare the performance of three ASR systems: a baseline system that uses phone-scale representations and units, an experimental system that uses a syllable-oriented front-end representation and syllabic units for recognition, and a third system that combines the phone-scale and syllable-scale recognizers by merging and rescoring $N$-best lists. Using the combined recognition system, we observed an improvement in word error rate for telephone-bandwidth, continuous numbers from 6.8% to 5.5% on a clean test set, and from 27.8% to 19.6% on a reverberant test set, over the baseline phone-based system.

## 1. INTRODUCTION

Automatic speech recognition (ASR) systems typically focus on short-time information distributed over periods of 10–100 ms. A speech signal is partitioned into overlapping frames of 20–30 ms for purposes of feature extraction. Frames are classified into phone or sub-phone classes typically using features from one to five contiguous frames. These frame-level classifications are then assembled together via a decoding process incorporating constraints from a set of word models which describe words as sequences of phone or sub-phone units, and a model of the grammar of the language, to produce a hypothesized word sequence. Some strategies for improving the robustness of the front-end speech representation, e.g., cepstral normalization and speaker adaptation, employ information from longer segments of speech.

While the successful application of current speech recognition technology to a range of tasks clearly demonstrates the utility of short-time speech representations and units, there is ample evidence that information distributed over longer periods of time, especially over durations of syllabic length (100–250 ms), should also be incorporated into the recognition process. To test the usefulness of syllable-time-scale information in automatic speech recognition, we developed an ASR system that focuses on information encoded over syllabic durations and compared its performance on clean and reverberant speech with that of a baseline recognizer focusing on information at the phonetic segment scale, as well as with a recognizer that combines the syllable-based and phone-based recognizers into a single system. While the baseline recognizer has better performance than the syllable-based recognizer on both the clean and reverberant test sets, the recognizer that combines both the phone-scale and the syllable-scale systems is significantly better than the baseline on both clean and reverberant speech.

We briefly review some of the evidence for the importance of syllable-scale information in human speech perception. Next, we describe the speech materials and recognition systems used in our experiments, with an emphasis on the methods used to focus on syllable-scale information in the front-end speech representation, speech unit classification, and speech decoding stages of an ASR system. Next, we review different methods for combining speech recognition systems and explain the utterance-level combination method we used. Finally, we discuss the results of our experiments.

## 2. THE SYLLABLE IN SPEECH RECOGNITION

Although the role of the syllable in human speech processing is the subject of continuing investigation, psychoacoustic and psycholinguistic studies suggest that the syllable plays a central role in human speech perception.

Studies of human speech perception demonstrate the dependence of speech intelligibility on relatively slow changes in the spectrum of the speech signal. These changes manifest themselves as amplitude modulations at rates of 2–16 Hz in subbands following a critical-band spectral analysis. The first evidence for this relationship between slow modulations and speech intelligibility emerged from the work of Homer Dudley and his colleagues at Bell Labs on the channel vocoder [6]. Subsequently, work on the prediction of speech intelligibility in reverberant and noisy rooms [12] and over nonlinear communications channels [17] have highlighted the importance of slow modulations. Recent perceptual studies [5, 1] show that suppression of modulations in the 2–8 Hz range significantly degrades speech intelligibility. Modulations in this frequency range correspond to the typical durations of single syllables and metrical feet.

A second line of evidence for the role for syllables in speech recognition comes from the study of echoic memory. Measurements of the capacity of the human echoic memory [13, 15] demonstrate that this preperceptual buffer can store roughly 250 ms. In conversational English nearly 80% of syllables have a duration of 250 ms or less [9]; thus, the syllable may constitute a natural unit for the segmentation and recognition of the speech signal, as it is the largest unit of speech which can fit into preperceptual storage.

That the syllable plays an important role in identification can be surmised from [18], in which the author asserts that "temporal compounds" are formed by acoustic elements and that the human perceptual system can identify these compounds more readily than constituent sounds. These temporal compounds were found to be longer than the typical phoneme length through experiments with loud, clear, repeating vowel acoustic elements catenated together.

The syllable was proposed as a unit for automatic speech recognition as early as 1975 [8]. Since then, the syllable has been revisited in a number of speech recognition systems in several languages and in various capacities, most recently in [19, 7]. The syllable is an attractive unit for recognition for several reasons:

1. Syllable representations and durations may exhibit greater stability relative to phoneme-based representations and durations.
2. Syllables appear to offer a natural interface between speech acoustics and lexical access.
3. Syllables constitute a convenient framework for incorporating suprasegmental prosodic information into recognition.

The implementation of syllable-based recognizers has proven to be challenging, however.

## 3. EXPERIMENTAL MATERIALS AND RECOGNIZERS

### 3.1. Speech Material

Recognition experiments were performed on a subset of the OGI Numbers corpus [4], a set of continuous, naturally spoken utterances collected from many different speakers over the telephone. The 32-word vocabulary is restricted to numbers, including such confusable sets as "four," "fourteen," and "forty." A sample utterance from the database is "eighteen thirty one." The subset of the Numbers corpus used in these experiments has a training set of about three hours of speech (3600 utterances) and a development test set and an evaluation test set each containing about one hour of speech (1200 utterances). Recognizer performance was measured on unmodified "clean" and digitally reverberated versions of the evaluation test set, while recognizer training was performed on the clean training set. The reverberant test sets were generated by convolving the clean sets with an impulse response measured in a room having a reverberation time of 0.5 s and a direct-to-reverberant energy ratio of 0 dB.

### 3.2. Baseline Recognizer

Each of the speech recognition systems implemented for these experiments was a hybrid hidden Markov model/multilayer perceptron (HMM/MLP) speech recognizer [3] in which the phonetic classification was performed with a single-hidden-layer MLP and speech decoding was performed by NOWAY [16], a start-synchronous stack decoder. The recognizers all used the same backoff bigram grammar, derived from the Numbers training set, for language modeling.

The baseline recognizer used eighth-order log-RASTA-PLP [11] features computed over 25-ms windows with a 10-ms window step, supplemented with delta features computed over a 9-frame window. RASTA-PLP features include a filtering operation on critical-band spectral trajectories, using a filter with a 1–12 Hz passband. The MLP phonetic classifier, with 400 hidden units, took features from 105 ms (9 frames) of speech and classified them into 32 phone categories. A multiple-pronunciation lexicon with simple minimum-duration modeling was developed for the baseline recognizer based on the phonetic hand-transcriptions of the training data. Embedded Viterbi alignment was applied iteratively to optimize the lexicon pronunciations, minimum phone durations, and training labels.

### 3.3. A Syllable-based Recognizer

The syllable-based recognizer used modulation spectrogram features [10] for the front-end speech representation. To compute these features, speech sampled at 8 kHz is analyzed into 15 quarter-octave channels using an FIR filterbank. In each channel, an amplitude envelope is computed by performing half-wave rectification, low-pass filtering with a 20-Hz cutoff frequency and decimation by a factor of 80 on the filter output. These steps produce a spectral shape estimate with critical-band-like resolution. Next, each amplitude envelope signal is normalized by its average value, computed over an entire utterance, to provide a crude model of the auditory adaptation. The normalized envelope signals are filtered, to model auditory sensitivity to slow modulations, compressed with a cube-root nonlinearity, and normalized to a range of $[-1, +1]$. Two different modulation filters are used in parallel: a low-pass filter with a cutoff frequency of 8 Hz and a band-pass filter with cutoff frequencies of 2 Hz and 8 Hz. This severe modulation filtering blurs envelope fluctuations at the phonetic segment scale (12–20 Hz), while emphasizing changes at the syllabic scale. The most important difference between these features and the RASTA features is the much narrower filter applied to the envelope signals, which leads to comparatively more temporally-smeared features.

Similar to the baseline system, the MLP in the syllable-based system had a single hidden layer of 400 units, though the syllable-based system used an extended context window of 185 ms (17 frames) and classified the features into 124 "semi-syllable" categories.

The syllable-based lexicon was derived from the baseline system's lexicon via a direct mapping from phones to semi-syllable units. For each pronunciation in the lexicon, the corresponding phone sequence was partitioned into syllables via an automatic syllabification program. The syllables were split into semi-syllables at the midpoint of the nucleus. Minimum duration constraints for the decoding process were derived directly from the minimum durations for the constituent phones of each semi-syllable. Although a forced-alignment procedure was used to realign the training labels, the lexicon was not further optimized.

### 3.4. Combining Recognizers

Combinations of speech recognition systems can potentially achieve better accuracy than individual recognizers if the recognizers being combined tend to make independent errors and the combination method allows correct answers to override incorrect ones. Merging methods (classifier fusion techniques) at the frame level [14] and at the syllable level [7] that blend diverse information sources have been applied to obtain substantial improvements in accuracy. For our experiments, we have focused on merging recognition systems at the *whole-utterance* level.

To combine the results of the baseline and syllable-based recognizers at the whole-utterance level, each recognizer is used to generate a word lattice for each input utterance. For each utterance, an $N$-best list with a maximum length of 150 hypotheses is generated from each lattice, the two $N$-best lists are concatenated, and duplicate hypotheses are eliminated. For each hypothesis in this merged list two acoustic scores are calculated via forced alignment using the baseline and syllable-based recognizers and a language model score is calculated from the backoff bigram grammar. The final score for each utterance is a weighted sum of the two acoustic scores and the language model score. An empirically determined weighting factor can be applied to influence the fusion of the acoustic scores from the different recognizers. In a series of experiments with the cross-validation portion of the training set, we found that the optimal weighting factor varied with task and system parameters. In the absence of a method for determining a good weighting on-line, an equal weighting of the acoustic scores seemed to be the best operating point. From the list of rescored hypotheses, the top-scoring word sequence was selected as the recognized result for the utterance.

As a theoretical note, the adding of log likelihoods in the recognition fusion procedure assumes that the likelihoods are independent. Though this assumption is clearly incorrect, the combining method described still proves to be effective in the experiments described below.

## 4. RESULTS AND DISCUSSION

No optimization of recognizer parameters was performed on the evaluation test set. The cross-validation portion of the training set was used for optimization of parameters such as the language scaling factor, word transition penalty and the relative weighting of the two recognizers in the combined system. After this optimization process, we had four recognition systems defined that incorporated slightly different sets of syllable-based information. A single syllable-based system was selected based on the performance of these four systems on the development test, and only this system was used for the evaluation test set results.

Because the combined system has many more parameters than either of the constituent systems, we conducted several experiments with expanded hidden layers and larger feature sets on the development test set to rule out the possibility that the observed performance improvement is merely the result of the increased number of recognizer parameters. When the hidden layer of the MLP was expanded to 1000 hidden units to equalize the total number of parameters with the combined system, we observed no improvement with the clean version of the development test set, and a modest improvement, 5.7% relative, with the reverberant test set. Combining the RASTA and modulation spectrogram acoustic features at the input to a 400-hidden-unit MLP (which also equalizes the total number of parameters) resulted in a 1% relative increase in word error rate for the clean development test set and a 17% relative decrease in word error rate for the reverberant version of the development test set. None of the improvements in accuracy were as large as that achieved by combining at the frame or utterance level.

Table 1 summarizes the performance of the different recognition systems on the clean and reverberant evaluation test sets. The syllable-based system is less accurate than the baseline system for both the unmodified and digitally reverberated test sets. The reported performance of this system and also the combined system is conservative because the lexicon used for recognition was tuned for the baseline system and not optimized for the syllable system. Table 1 also shows that the combined system is significantly more accurate that either of the two constituent systems (statistically significant at the 0.05 level). The clean performance represents a 19% relative reduction in error over the baseline system performance, and the reverberant performance is a 29% relative reduction in error compared to the baseline on the evaluation test set. These findings closely resemble the performance of the system on the development test set, where the introduction of the combined system reduced the error rate on the clean development test set from 7.0% to 5.6% and on the reverberant development test set from 29.2% to 20.0%. The performance improvement achieved for reverberant speech is similar to that obtained by us using a frame-level combination method (unpublished observations), as well as to the results in [2], where a multiresolution channel normalization method that operates over time spans of 10 sec. is used to compensate for reverberation.

| System | Clean | Reverb. |
|---|---|---|
| RASTA, phone units,<br>105-ms context window<br>Baseline | 6.8% | 27.8% |
| ModSpec, syllable units,<br>185-ms context window | 9.8% | 30.9% |
| Combined | 5.5% | 19.6% |

Table 1: Word-error rates for each individual system and combined for evaluation test set.

## 5. CONCLUSIONS

A syllable-length time scale appears to be a fundamental property of the human speech recognition system. We have implemented an automatic speech recognition system with syllables and syllabic time scales integral to the processing at the feature-extraction level, at the input to the phonetic classification stage and as the unit of speech recognition. Although this system did not perform as well as our baseline system, a combination of the two systems, through a mechanism that merges and rescores $N$-best lists of whole utterances, significantly outperforms the baseline system on both clean and reverberant versions of the test data. We believe this result is due to mutual compensation of the blended recognizers that offsets the weaknesses of the individual systems.

## 6. ACKNOWLEDGMENTS

We are grateful to Jim West and Gary Elko, from Bell Labs, and Carlos Avendano, now at the University of California, Davis, for collecting a set of room impulse responses and making them available to us. Bill Fisher's TSYLB2 program, from NIST, and Dan Ellis's user interface were invaluable for automatically syllabifying phonetic transcriptions. The lexicon used for these experiments was based on the work of Dan Gildea. We very much appreciate help from Steve Renals with NOWAY and from Tony Robinson with SLIB, of which the $N$-best list generation program is a part. We also gratefully acknowledge assistance from Eric Fosler.

This work was supported in part by a Joint Services Electronics Program grant F49620-94-C-0038 and an Office of Naval Research grant N00014-92-J-1617. Additional support for this project came from a European Community Basic Research grant (Project Sprach) and from the International Computer Science Institute.

## 7. REFERENCES

[1] Takayuki Arai, Misha Pavel, Hynek Hermansky, and Carlos Avendano. Intelligibility of speech with filtered time trajectories of spectral envelopes. In *ICSLP*, volume 4, pages 2490–2493, 1996.

[2] Carlos Avendano, Sangita Tibrewala, and Hynek Hermansky. Multiresolution channel normalization for ASR in reverberant environments. In *Eurospeech*, volume 3, pages 1107–1110, Rhodes, Greece, September 1997. ESCA.

[3] Hervé Bourlard and Nelson Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[4] R. A. Cole, M. Noel, T. Lander, and T. Durham. New telephone speech corpora at CSLU. In *Eurospeech*, volume 1, pages 821–824, September 1995.

[5] Rob Drullman, Joost M. Festen, and Reinier Plomp. Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 95(2):1053–1064, February 1994.

[6] Homer Dudley. Remaking speech. *Journal of the Acoustical Society of America*, 11(2):169–177, October 1939.

[7] Stéphane Dupont, Hervé Bourlard, and Christophe Ris. Using multiple time scales in a multi-stream speech recognition system. In *Eurospeech*, pages 3–6, Rhodes, Greece, October 1997.

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

[9] Steven Greenberg. On the origins of speech intelligibility in the real world. In *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Channels*, pages 23–32. ESCA, 1997.

[10] Steven Greenberg and Brian E. D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *ICASSP*, volume 3, pages 1647–1650, Munich, Germany, April 1997. IEEE.

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

[12] T. Houtgast and H. J. M. Steeneken. The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acustica*, 28:66–73, 1973.

[13] A. W. F. Huggins. Temporally segmented speech. *Perception and Psychophysics*, 18(2):149–157, 1975.

[14] Brian E. D. Kingsbury and Nelson Morgan. Recognizing reverberant speech with RASTA-PLP. In *ICASSP*, volume 2, pages 1259–1262, Munich, Germany, April 1997. IEEE.

[15] Dominic W. Massaro. Preperceptual images, processing time and perceptual units in auditory perception. *Psychological Review*, 79(2):124–145, 1972.

[16] 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.

[17] Herman J. M. Steeneken and Tammo Houtgast. A physical method for measuring speech-transmission quality. *Journal of the Acoustical Society of America*, 67(1):318–326, January 1980.

[18] Richard M. Warren. Perceptual processing of speech and other perceptual patterns: Some similarities and differences. In Steven Greenberg and William Ainsworth, editors, *Listening to Speech: An Auditory Perspective*. Oxford University Press, 1998. To appear.

[19] Su-Lin Wu, Michael L. Shire, Steven Greenberg, and Nelson Morgan. Integrating syllable boundary information into speech recognition. In *ICASSP*, volume 1, Munich, Germany, April 1997. IEEE.