PERFORMANCE IMPROVEMENTS THROUGH COMBINING PHONE- AND SYLLABLE-SCALE INFORMATION IN AUTOMATIC SPEECH RECOGNITION

Su-Lin Wu,[1] Brian E. D. Kingsbury, Nelson Morgan, Steven Greenberg

International Computer Science Institute
University of California at Berkeley
{sulin, bedk, morgan, steveng}@icsi.berkeley.edu

ABSTRACT

Combining knowledge derived from both syllable- (100-250 ms) and phone-length (40-100 ms) intervals in the automatic speech recognition process can yield performance superior to that obtained using information derived from a single time scale alone. The results are particularly pronounced for reverberant test conditions that have not been incorporated into the training set. In the present study, phone- and syllable-based systems are combined at three distinct levels of the recognition process — the frame, the syllable and the entire utterance. Each strategy successfully integrates the complementary strengths of the individual systems, yielding a significant improvement in accuracy on a small-vocabulary, naturally spoken, telephone speech corpus. The syllable-level combination outperformed the other two methods under both relatively pristine and moderately reverberant acoustic conditions, yielding a 20-40% relative improvement over the baseline.

1. SYLLABLES IN ASR

Most current automatic speech recognition (ASR) systems for English rely predominantly on phone-scale information. Phonological and psychoacoustic evidence suggests, however, that syllable-length intervals also play an important role in spoken language processing by human listeners [6, 12], particularly under adverse acoustic conditions, and thus may prove of utility in speech recognition by machines [4].

Our analysis of machine recognition errors indicates that, to a certain degree, syllable-based information complements phone-based information and that these two distinct knowledge representations can be used in combination to provide better performance than is possible using either representation alone. This approach may be particularly advantageous for recognition of speech spoken under adverse acoustic conditions (such as background noise or reverberation) that typically results in a pronounced deterioration of ASR performance [11]. In this paper we demonstrate that using appropriate combination methods to integrate multiple, complementary representations of the speech input can reduce the deleterious impact of such acoustic interference (for an example of a more syllable-focused design see [5]).

[1] Su-Lin Wu is now with Nuance Communications.

2. RECOGNITION SYSTEMS

2.1. Speech Material

The speech materials used in the current set of experiments were recorded over the telephone from speakers of both genders and include native speakers from all major dialect regions of the United States as well as non-native speakers. The utterances consist of continuous, naturally spoken numbers from a vocabulary of 32 separate words (e.g., "two hundred eleven") and are derived from the Oregon Graduate Institute's Numbers corpus [1].

Approximately 1.6 hours (about 3,500 utterances containing 14,000 word tokens) of material were used for training and two separate 40-minute portions (about 1,200 utterances containing 4,700 word tokens) were used for development and evaluation testing. The evaluation set was held back from use until all recognition parameters had been determined with the training and development material.

Artificially reverberated versions of the development and evaluation test material were used as exemplars of adverse conditions not represented in the systems' training data. These materials were created by convolving the original ("clean") speech signals with a room impulse response whose reverberation time (RT60) was 0.5 s and whose direct-to-reverberant energy ratio was 0 dB.

2.2. Phone-based System (Baseline)

The baseline system was a hybrid hidden Markov model/multilayer perceptron (HMM/MLP) system in which HMM emission probabilities were estimated by a multilayer perceptron, as described in [13]. The input representation was eighth-order log-RASTA-PLP [8] computed over 25-ms frames every 10 ms, supplemented with delta features calculated over a 9-frame window. RASTA-PLP features incorporate filtering of spectral trajectories in critical-band-like channels using a filter with a 1–12 Hz passband. For purposes of probability estimation the baseline system used an MLP having a single hidden layer with 400 hidden units, a 9-frame input context window and 32 context-independent phone output categories.

The system used a multiple-pronunciation lexicon derived from phonetic transcription (by trained human listeners) of the Numbers corpus [1]. An embedded Viterbi alignment process opti-

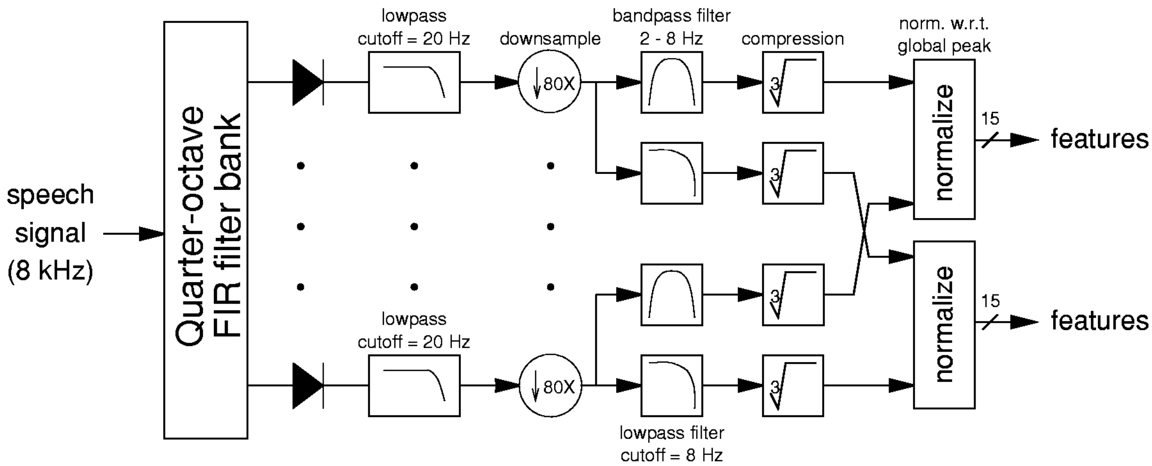

Figure 1: Modulation spectrogram feature extraction method. See text (Subsection 2.3) for a description of the algorithm.

mized the pronunciations, the models for minimum phone duration and the training labels. The language model was a backoff-bigram grammar learned from word transcriptions of the same data set used for acoustic training.

2.3. Syllable-based Systems

Two hybrid HMM/MLP syllable-based systems were also developed. Syllable-oriented design elements were incorporated into each system by using an alternative feature analysis method, the modulation spectrogram [7], and lengthening the MLP input window to 17 frames. One of the syllable-based systems used half-syllable recognition units, while the other used conventional context-independent phone units. The language model was the same as that used in the baseline system.

Modulation Spectrogram. Both experimental systems used modulation spectrogram features to incorporate syllable timing at the signal processing stage. In this process [7, 10], illustrated in Figure 1, the speech signal is decomposed into 15 quarter-octave channels using an FIR filterbank, and an amplitude envelope is calculated for each channel via rectification and lowpass filtering. The envelope signals are themselves filtered to simulate temporal response properties of the auditory cortex, then compressed and normalized. The normalization is performed by finding the point with the largest magnitude across all channels over the entire utterance and dividing the envelope signals by that magnitude. Two different modulation filters process the envelope signals in parallel: a lowpass filter with a cutoff frequency of 8 Hz and a bandpass filter with cutoff frequencies of 2 Hz and 8 Hz. This rather severe modulation filter blurs envelope fluctuations at the phonetic segment scale (12–20 Hz), while emphasizing spectral changes at the syllabic time scale. The normalization is applied separately to the outputs of the lowpass and bandpass envelope filters. These features have many characteristics in common with RASTA-PLP but differ in a number of respects. The key difference is the much narrower (and steeper) filter applied to the envelope signals in the modulation spectrogram processing which produces features that are more temporally-smeared than those produced with RASTA-PLP. A more recent, enhanced version of the modulation spectrogram processing is described in detail in [9].

Lengthening the MLP Input Window. The baseline phone-oriented system uses an MLP input window of 9 frames (roughly equivalent to 105 ms of the speech signal). By lengthening the context window to 17 frames (equivalent to 185 ms), syllable-scale timing is incorporated at the input of the probability estimation stage. In principle, the MLP can use this longer window to capture dependencies useful for the estimation of the conditional probabilities of longer time-scale states.

Syllable-based Recognition Units. To further emphasize syllable time scales, one of the syllable-based systems uses half-syllable recognition units instead of phones. 124 half-syllables were derived from the OGI Numbers lexicon described above by dividing each syllable at the midpoint of the nucleus. The resulting recognition unit typically covers a longer stretch of the speech signal than the phone. The lexicon and training labels for the syllable-based system were adapted from automatic syllabification of the phone version, but were not further optimized.

3. RECOGNIZER COMBINATION METHODS

We experimented with combining the baseline phone-based system with the syllable-based systems at three levels of the decoding process: at the frame, syllable, and whole-utterance stages. The simplest to implement, frame-level combination, multiplies corresponding phone probabilities at the output of the MLPs from each recognition system. Functionally, probabilities from corresponding phone outputs are used at the same time step and hence the two recognition streams are closely coordinated. The one-to-one correspondence needed for this combination method precluded using syllable-based recognition units. For this reason the syllable system with context-independent phone recognition units was used instead.

Combining two recognition streams at the syllable level entails multiplying corresponding syllable-string likelihoods at the end of syllable hypotheses during the decoding process. We used

Individual Systems	Clean	Reverb
Phone-based System (baseline)	6.7%	28.0%
Syllable-based System with Phone Units	8.6%	25.8%
Syllable-based System with Syl. Units	10.0%	30.1%

Table 1: Word error rates associated with each system for the clean and reverberant evaluation test sets.

Combining Method	Syllable-based System Combined with Baseline	Clean	Reverb
Frame	System with Phone Units	5.8%	17.7%
Syllable	System with Syllable Units	5.1%	16.7%
Utterance	System with Syllable Units	5.5%	19.6%

Table 2: Word error rates produced by combining the baseline phone-based system and the experimental syllable-based systems for the clean and reverberant evaluation test sets.

HMM-recombination [2] (similar to HMM-decomposition) to implement this step. Because the two recognition streams interact only at the end points of syllables, they may be desynchronized elsewhere. This potential asynchrony permits the use of half-syllable recognition units since the outputs of the neural networks need not be as closely coordinated as with frame-level combination.

The utterance combination method, discussed in more detail in [15], multiplies corresponding word-string likelihoods at the end of the entire utterance. This combination method is implemented by merging and rescoring N-best lists. Each recognizer generates a maximum of 150 hypotheses, which are then rescored using both recognition systems. The scores from the two systems are simply added together to determine the overall score for a hypothesis and the best scoring hypothesis is designated as the recognized word string. Because the two recognition streams interact only at the end of the utterance, they can be decoded asynchronously, thus permitting the use of half-syllable units.

4. RESULTS

Although each recognition system performs moderately well on its own, combining the syllable-based systems with the baseline results in significantly lower error rates. As illustrated in Table 1, neither of the syllable-based systems is as accurate as the baseline system for the “clean” test condition. This is not surprising since the syllable-based elements in the recognizers smear information germane to phonetic identity across time. In the case of the reverberant test condition the syllable-based system with phone-based recognition units achieves a slighter higher performance level than the baseline as a consequence of the modulation spectrogram features and the wider neural network context window (both of which are designed to accommodate temporal smearing of the speech information).

Combining the experimental syllable-based systems with the baseline system results in performance improvements using any of the combination methods outlined above, as shown in Table 2. The gain in performance for each case is larger than that achieved by merely increasing the number of MLP parameters. Of the three methods, the syllable-level combination displays the largest improvement over the baseline, with 20% relative improvement for clean speech and 40% relative improvement for reverberant speech. The frame-level combination is almost as effective, while having a lower implementation cost. Combination at the utterance-level, while still exhibiting considerable improvement in performance compared to the baseline, manifests the smallest performance gain.

5. ERROR ANALYSIS

Inspection of the individual recognizers’ errors helps explain the improvement obtained via these methods of combination. The recognizers generally make different types of errors and the combination methods allow correct answers to override those that are incorrect. A simple error analysis method (based on [3]) quantifies the differences between any two systems. The recognition outputs are compared with the actual word strings to measure the degree of accuracy and concordance between systems. The outputs from the two recognizers (corresponding to a given correct word) can be categorized into one of five categories:

Both Correct Both systems recognized the word correctly.

Phone System Only Correct Only the baseline system recognized the word correctly.

Syllable System Only Correct Only the experimental syllable-based system recognized the word correctly.

Both Incorrect and Different Both systems recognized a given word incorrectly, but produced different hypotheses.

Both Incorrect and Identical Both systems recognized a given word incorrectly, but produced the same hypothesis.

Tables 3 and 4 illustrate this analysis for a syllable-based system and the baseline system. Analogous procedures for frame-level, syllable-level and utterance-level analyses are described in [14].

Identical errors are generally the most difficult type to compensate for using simple combination methods, while other types of error are more readily corrected. For this reason, the proportion of identical errors can serve as an indirect measure of the potential for improvement via combining. For each case, the percentage of identical errors is relatively small (and hence the potential for decreasing the error rate through combination is high). Comprehensive comparisons between the baseline system and other experimental systems [14] show that introducing syllable-based design elements reduces the number of identical errors between systems and thus increases the potential for gains in performance.

6. CONCLUSIONS AND FUTURE WORK

The combination of a relatively conventional phone-based ASR system with an experimental system incorporating some measure of syllable-time-scale information produced significant performance improvements for both clean and reverberant test material. The superiority of the syllable-level combining strategy may reflect the temporal organization of speech. For practical purposes

	Both Correct	Phone System Only Correct	Syllable System Only Correct	Both Incorrect and Different	Both Incorrect and Identical
word count	4142	331	153	76	92
percentage	86.4%	6.9%	3.2%	1.6%	1.9%

Table 3: Comparing recognition behavior of the phone-based system (baseline) and the syllable-based system with syllable output units for clean speech (development test set).

	Both Correct	Phone System Only Correct	Syllable System Only Correct	Both Incorrect and Different	Both Incorrect and Identical
word count	2863	754	714	423	234
percentage	57.4%	15.1%	14.3%	8.5%	4.7%

Table 4: Comparison of the phone-based system and the syllable-based system with syllable output units for reverberant speech (development test set). The total word count does not equal that of the clean set due to the variability in the number of insertions.

of implementation, however, the simpler frame-level combination appears more attractive. In collaboration with Cambridge University's Connectionist ASR group, similar combination strategies are now being explored for the 1998 Broadcast News evaluation.

7. ACKNOWLEDGMENTS

We are grateful to Jim West and Gary Elko, from Bell Labs, and Carlos Avendaño, 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, was invaluable for automatically syllabifying phonetic transcriptions. We very much appreciate help from Steve Renals with NOWAY and from Tony Robinson with SLIB.

This work was supported, in part, by a Joint Services Electronics Program grant F49620-94-C-0038, an Office of Naval Research grant N00014-92-J-1617 and NSF grant IRI-9712579. Additional support for this project came from a European Community Basic Research grant (Project Sprach) and the International Computer Science Institute.

8. REFERENCES

1. R. A. Cole, M. Noel, T. Lander, and T. Durham. New telephone speech corpora at CSLU. In *Eurospeech*, pages 821–824, Sept. 1995.
2. S. Dupont, H. Bourlard, and C. Ris. Using multiple time scales in a multi-stream speech recognition system. In *Eurospeech*, pages 3–6, October 1997.
3. K. R. Farrell, R. P. Ramachandran, and R. J. Mammone. An analysis of data fusion methods for speaker verification. In *ICASSP*, pages 1129–1132. IEEE, Apr. 1998.
4. O. Fujimura. Syllable as a unit of speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1):82–87, Feb. 1975.
5. A. Ganapathiraju, V. Goel, J. Picone, A. Corrada, G. Doddington, K. Kirchhoff, M. Ordowski, and B. Wheatley. Syllable — a promising recognition unit for LVCSR. In *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 207–214, Santa Barbara, California, Dec. 1997. IEEE.
6. S. Greenberg. On the origins of speech intelligibility in the real world. In *Proc. of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 23–32. ESCA, Apr. 1997.
7. S. Greenberg and B. E. D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *ICASSP*, pages 1647–1650. IEEE, Apr. 1997.
8. H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, Oct. 1994.
9. B. E. D. Kingsbury. *Perceptually-inspired signal processing strategies for robust speech recognition in reverberant environments*. PhD thesis, University of California, Berkeley, California, 1998. To appear.
10. B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 1998. In press.
11. R. Lippmann. Speech perception by humans and machines. In *Workshop on the Auditory Basis of Speech Perception*, pages 309–316. ESCA, July 1996.
12. D. W. Massaro. Preperceptual images, processing time and perceptual units in auditory perception. *Psychological Review*, 79(2):124–145, 1972.
13. N. Morgan and H. Bourlard. Continuous speech recognition. *IEEE Signal Processing Magazine*, 12(3):25–42, May 1995.
14. S.-L. Wu. *Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition.* PhD thesis, University of California, Berkeley, California, May 1998.
15. S.-L. Wu, B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Incorporating information from syllable-length time scales into automatic speech recognition. In *ICASSP*, pages 721–724. IEEE, Apr. 1998.