Hierarchical Speech Recognition

1. Research Team

Project Leader: Prof. Shrikanth Narayanan, *Electrical Engineering*

Graduate Students: Abhinav Sethy

2. Statement of Project Goals

Speech recognition is an essential component of any Human Computer Interaction (HCI) scheme, which aspires to be natural. Thus, high accuracy speech recognition is of critical importance in making natural man-machine interfaces. Most systems today are based on phonemes, which are considered to be the fundamental units of speech based communication. For recognition purposes the phoneme provide a convenient unit in terms of training data requirements and availability. However the short duration of the phoneme limits us to correlations and information present in time scales of around 30-40ms. This puts fundamental limits on the recognition accuracy that can be achieved. The goal of this project is to design training and recognition algorithms for building systems, which will use units such as syllable or word to provide a much larger acoustic context for recognition. In addition larger units are more robust in handling pronunciation variations, which are common in a diverse cultural society such as the USA. Based on our current experimental results we can confidently say that such hierarchical systems will clearly outperform existing phoneme based systems.

3. Project Role in Support of IMSC Strategic Plan

The proposed work contributes to enabling natural and customizable interactions, a key element of IMSC's strategic plan. We aim to use multiple time scale units to improve speech recognition accuracy. This work will be of especial significance for multicultural scenarios where there are significant pronunciation and accent variations.

4. Discussion of Methodology Used

The major challenge in using syllables and word level units for recognition is the training data sparsity problem. The number of syllables and words in a language like English is more than 100 times the number of phonemes, so a large number of these units will have little or no acoustic training data and this could lead to poor performance. We have addressed this problem in two steps: First we use context dependent phonemes to initialize the longer duration units in a manner, which minimizes the impact of training data sparsity. Subsequently we split the lexicon into units of different acoustic length based on an analysis of the training data. The second step also ensures that the larger units are used only if they help in improving recognition accuracy.

5. Short Description of Achievements in Previous Years

We have designed and evaluated different algorithms for training and unit selection for hierarchical speech recognition.

5a. Detail of Accomplishments During the Past Year

For a medium vocabulary speech recognition task, our initialization from CD phone scheme allowed us to improve recognition accuracy substantially. As can be seen in the table below, the syllable and word level units when initialized from CD phonemes are able to give equal performance without any further training. Thus even with limited retraining on limited acoustic material we can get substantially better performance.

Recognizer Type (ms)	First Reestimation	Third Reestimation
Context Free Syllable	72	85
Context Free Word	74	87
Context Dependent Phoneme	74	74

As can be seen in the table below, proper unit selection helps in further improving accuracy with a decrease in system complexity.

Recognizer Type (ms)	Accuracy	Number of model states in recognizers
Context Free Word	87	43380
Context Free Syllable	85	24460
Mixed Unit Recognizer	90	13450

6. Other Relevant Work Being Conducted and How this Project is Different

Previous efforts at using cross phoneme correlations have focused on using techniques like parameter HMMs and multi path HMMs [4,5] in a phoneme recognition framework. However these techniques have led to marginal improvements, which highlights the fact that long-term correlations cannot be captured in a phoneme-based system.

Work on syllable-based recognition [3,6,7] has not addressed the training and lexical selection problems, which are our primary goals. These schemes essentially use only syllable units, whereas we use units of different lengths together to achieve better performance without substantial increase in system perplexity.

7. Plan for the Next Year

For the next year, we plan to extend our unit selection strategies by using MAP/MLE criteria and design algorithms to optimally training units of different acoustic length in a single framework.

8. Expected Milestones and Deliverables

Development of algorithms for robust training for multi unit recognitions

9. Member Company Benefits

N/A

10. References

[1] Abhinav Sethy, Shrikanth Narayanan and S. Parthasarthy, "A syllable based approach for improved recognition of spoken names", *Proceedings of the ISCA Pronunciation Modeling Workshop*, Estes Park, Colorado, September 2002

[2] Abhinav Sethy, Shrikanth Narayanan, "Split-Lexicon based hierarchical recognition of speech using syllable and word level acoustic units", To appear *ICASSP 2003*, Hong Kong, April 2003.

[3] Ganapathiraju, J. Hamaker, M. Ordowski, G. Doddington and J. Picone, Syllable-Based Large Vocabulary Continuous Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 358-366, May 2001.

[4] H. Gish and K.Ng, "Parameter trajectory models for speech recognition", *Proceedings of ICSLP*, Philadelphia. PA, 1996, pp 466-469.

[5] F.Kormazskiy, "Generalized mixture of HMM's for continuous speech recognition", *Proceedings of ICASSP*, Munich, Germany, 1997, pp 1443-1446.

[6] Kirchhoff, K., "Syllable-level desynchronisation of phonetic features for speech recognition", *International Conference of Spoken Language Processing* 1996, pp 2274-2276.

[7] Su-Lin Wu, Brian Kingsbury, Nelson Morgan, and Steven Greenberg, "Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition", *ICASSP-98*, Seattle, pp. 721-724.

[8] S. Greenberg, "Speaking in Shorthand - A Syllable-Centric Perspective for Understanding Pronunciation Variation", *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, The Netherlands, May 3-6, 1998.

[9] Odell J, Ollason D, Woodland P, Young S, Jansen J, *The HTK Book for HTK V2.0*, Cambridge University Press, Cambridge, UK, 1995.

[10] D. Kahn, "Syllable-Based Generalizations in English Phonology", Indiana University Linguistics Club, Bloomington, Indiana, USA, 1976.

[11] W.M. Fisher, "Syllabification Software", http://www.itl.nist.gov/div894/894.01/slp.htm, The Spoken Natural Language Processing Group, National Institute of Standards and Technology, Gaithersburg, Maryland, U.S.A., June 1997.