# Implicit Pronunciation Modeling for Speech Recognition Using Syllable-Centric Models

## 1. Research Team

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

Post Doc(s): Panayiotis Georgiou

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. 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. Larger units can implicitly handle minor base form variations without the need for dictionary augmentation. This is of importance in a diverse cultural group such as the USA. Based on our current experimental results we can confidently say that mixed syllable and phoneme based systems can help improve ASR performance significantly.

## 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

For a medium vocabulary speech recognition task based on TIMIT, 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 |

## 5a. Detail of Accomplishments During the Past Year

Syllable modeling based ASR was tried in the context of the NSF funded MALACH , (Multilingual Access to Large Spoken Archives) project. The MALACH project is an ongoing effort that aims to achieve a quantum leap in our ability to access the contents of large, multilingual, spoken archives by advancing the state of the art in automated speech recognition (ASR), information retrieval (IR) and other component technologies, by utilizing the world's largest digital archive of video oral histories collected by VHF1. VHF was created to record the firsthand accounts of Holocaust survivors, liberators, rescuers and witnesses and disseminate that information to future generations [2]. The MALACH corpus consists of unconstrained, natural speech filled with disfluencies, heavy accents, age-related coarticulations, uncued speaker and language switching and emotional speech collected in the form of interviews from over 52000 speakers in 32 languages. A good number of words uttered in this corpus are foreign words or sequences of words spoken in a foreign language, unfamiliar names and places. The corpus consists of elderly speech, where the age of the interviewees range from 56 years to 90 years. The age-related coarticulation effects (natural deletion of phones) contribute significantly to the high word error rates seen in this corpus.

Various variations on syllable modeling were tried on the MALACH corpus and we were able to achieve a 0.7% gain in WER and around 17% gain in Named entiry recognition [1] using a dual pronunciation phonetic-syllabic system in a rescoring framework.

## 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 [5.6] 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 [4.7.8] 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 use information theoretic criteria to select the optimal units to represent lexicon words and experiment with syllable length segmental trajectory models.

## 8. Expected Milestones and Deliverables

**2002-2004 -** Development of algorithms for robust training for multi unit recognitions
**2004-2006 -** Automatic clustering of lexical items to generate acoustic units of varying length
**2006-2008 -** Implementation and demonstration in human-machine interaction systems

## 9. Member Company Benefits

N/A

## 10. References

[1] Abhinav Sethy, Bhuvana Ramabhadran and Shrikanth Narayanan, "Improvements in English ASR for the MALACH project using syllable centric models",*IEEE ASRU 2003* , US Virgin Islands, December 2003.
[2] 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
[3] 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.
[4] 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.
[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.