Combining Phones, Syllable and Phonological Features in Speech Recognition Using Neural Networks

Arturo Geigel – M.S. Student
Electrical & Computer Engineering
and Computer Science Department
Email: ageigel@telinga.com

Prof. Kay Berling – Advisor
Polytechnic University of Puerto Rico

1 Abstract

The work proposes a new system that uses phones, phonological features and the syllable in the analysis of continuous speech recognition. The work also puts forth the autonomy of specific property of phonological features for neural networks. The proposed architecture takes advantage of the layering of the stages of the processing of the acoustic-representation, the phonological features, and the phone to achieve better performance. The implementation of the system also uses multiple neural networks to carry out a new analysis, in which the part of syllable is attached to each individual phonological feature with the goal of providing greater discriminatory power to the neural net.

2 Phonological Features, Phonemes and Syllables

- **Phoneme**: A phoneme is an abstraction devised to identify distinct speech sounds. Examples of phonemes classification are the International Phonetic Alphabet (IPA) and the ARPABET.
- **Phonological Feature**: The phonemes themselves are not individual units and can be decomposed based on common characteristics such as being covered by the sound tract (consonants), raising of the tongue body (high), etc. Such characteristics are called phonological features.

3 Problem Formulation

- **Recognizing Speech**: While the spectrogram representation of the speech signal shown in Figure 9 is important for the manual analysis of the speech signal, a more efficient method is required to automatically analyze and recognize the speech signal.
- **Current systems use Hidden Markov Models (HMMs) to recognize speech.** An simplified system using HMMs is shown in Figure 4.

4 Proposed Solution

- **Proposed Solution**: The proposed solution consists of an architecture of two layers of neural networks that feed from the output signal representation.
  - **Signal Representation**: Consists of a pre-emphasis of 0.97 using 30 ms frames and 10 ms frame shift. The Signal representation is selected with:
    - 12 Mel Frequency Cepstral Coefficients (MFCC)
    - 14 Linear Predictive Coding Coefficients (LPCC)
  - **The final signal representation for each feature in the first layer will be selected on the basis of performance comparison between MFCC and LPC on the specific feature.**
  - **First Layer**: 15 neural networks each consisting of 250 hidden units and 2 output nodes for the baseline system, and 6 output nodes for the PCE system.
  - **Second Layer**: 1 neural network that takes as input the result from the 15 neural networks, and has 78 input nodes, 250 hidden nodes and 190 output nodes.

5 Results

- **Results**: The overall difference between monolithic and distributed neural networks for phonological feature recognition is less than 1% and in some instances the distributed performed better than the monolithic baseline.
- **The neural net did not perform above chance in the LPC case as opposed to the MFCC, which performed substantially above average.**

6 Conclusions

- **Results provide evidence to support the thesis there is an autonomy of specific property of phonological features for monolithic neural networks, when used to make a distributed neural network architecture for phonological feature recognition.**
- **The results provide evidence to support the alternative hypothesis that the crisp resolution of theers displayed by LPC does not help the detection of phonological features.**
- **The addition of syllable parts to the phonological feature shows that addition to specific features can increase the performance of the features substantially.**
- **However, on average it increased 14% above the baseline.**
- **Results provide evidence to support the thesis that using a distributed architecture shows a clear advantage at the phonological feature level were different methods can be used on individual features to achieve higher accuracy of feature detection.**

7 References