Lecturer: Oded Ghiza *Scribe: Joseph Saracino*

Auditory Models and Human Performance, Part I

39.1 Overview

- Motivation
- Discussion of Database
- Model Evaluation Methodology
- Diagnostic Rhyme Test
- Extracting Diphones from Speech

39.2 Motivation

Auditory Models that are capable of achieving human performance would provide a basis for realizing effective speech processing systems. We look for an auditory model that can predict or mimic human performance. This brings up the question: How do we know the model mimics human performance in the right way? The model is tested by means of a quantative methodology; it is run on a comprehensive test database that scans the complete space of possible speech inputs. Unless this type of testing is performed, the model remains only a potential model.

Speech Coding: Saving bits in speech coders relies on a perceptual tolerance to acoustic deviations from the original speech. Purpose: We want to take speech, compress it to a minimum, transmit it, then reproduce the original signal with good results. In order to do this we need a good auditory model.

Front-End Block: Perceptual invariance to adverse signal conditions (noise, microphone and channel distortions, room reverberations) and to phonemic variability (due to the non-uniqueness of articulatory gestures) may provide a basis for robust speech recognition. Currently there does not exist a reliable acoustic to phonemic transduction method that works robust settings. We can apply an understanding of Human performance to speech processing front-ends.

39.3 Discussion of Database

- Speech Space defined over phonemic Dimensions. We try to find an orthogonal basis of compact phonetic features.
- Uniform coverage of the speech space. The model should show how it deals with all signals in the subspace.

- Non-cognitive psychophysical paradigm. We don't want to involve higher cognitive influences outside the scope of the model.
- Choice: Voiers' Diagnostic Rhyme Test (DRT)

39.4 Diagnostic Rhyme Test

The Diagnostic Rhyme Test (DRT) is a technique for evaluating the capabilities of our isolated word recognition model. Developed by Voiers, it consists of a database of 96 pairs of confusable words spoken in isolation. It covers initial diphones (see below for explanation of diphones) of the consonant-vowel-consonant (CVC) type. Each word pair differs only in its initial consonant. These pairs are further divided up into six distinctive phonemic dimensions. Each dimension is binary – corresponding to each half of its word-pair grouping. The test is to distinguish between these binary groupings. Key points summarized below.

- Two Alternative Forced Choice Test
- Voiers Database
  - 96 Pairs of confusable words (CVC type – consonant-vowal-consonant)
  - Words in a pair differ in their initial C.
  - 6 distinctive features (16 word-pairs per feature)

39.4.1 Procedure using the DRT

There were eight subjects for the test. Each was given all 192 words from the database in random order. When the acoustic stimulus was presented, the subject was shown the binary grouping of the stimulus and asked to choose which one was presented. We assume that all cognitive information is available to the listener and that error is due to inaccuracy in the auditory representation.

39.5 Relation of Human Auditory System to Model

The Human Auditory System is made of complex, highly-interlinked organs. At every step of processing there exists feedback from higher organs allowing the entire auditory system to adjust its function to the optimum operating point. This can be seen in the attached document. However, as mentioned in discussion, our model does not account for feedback.

Our model deals with the lower portions of the auditory system, in particular, the cochlea and the auditory nerve, and how they extract features for processing

39.6 Extracting Diphones from Speech

A *Diphone* is the smallest acoustic unit from a perception point of view. Multiple diphones are linked together to form words in speech. For our model it is important that we segment the diphones within the acoustic speech signal.

We start by noticing that there almost never exists a steady state vowel in speech. Rather, vowels are a target that we try to reach from consonants. We use a mechanism sensitive to contrasts between consonants and vowels and, use this “edge” as a pointer. Then the diphones are segmented surrounding this pointer.

Pre-existing research has provided a vocaulary of “diphone states”, or perceptual magnets, that indicate likely diphone combinations in real speech. A match is then attempted between the diphone segments and the existing database of diphone states.

The model and a sample of diphone segmentation from a speech signal are illustrated in the attached document.