EECS225d Audio Signal Processing in Humans and Machines **Spring 1995**

Lecture 15: February 22nd, 1995

Lecturer: Tom Holton *Scribe: Brett Walker*

15.1 Models of the Auditory System to Improve Speech Recognizer Front Ends

Today our guest lecturer, Tom Holton of San Francisco State University, gave us a review of some of his research. He concentrated on the use of models of the auditory system to come up with radical new ideas for speech recognition algorithms.

15.1.1 Speech Recognition

Speech Recognition can be broken up into the following sequence of functional units. Each function is performed differently in man than in machine.

- Transducer: *Man* uses the ear. *Machine* uses a microphone.
- Signal Processor: *Man* uses the cochlea. *Machine* uses some form of spectral analysis.
- Feature Extraction: *Man* uses the auditory nerve. *Machine* uses formant trackers, phoneme detectors, etc.
- Utterance Classifier: *Man* uses the Left Cerebral Cortex. *Machine* uses computers

15.1.2 Ear Models

Ear Models were once again introduced and explained. For a more detailed description of the auditory system see *Lecture 12*. Important points to note are:

- The cochlea performs a spectral analysis. Nerves along the Basilar Membrane have peak sensitivity at different frequencies. The base corresponds to high frequencies while the apex corresponds to low frequencies.
- The nerves connected to the Basilar Membrane fire in response to an audio stimulus. The nerve spikes tend to correspond to the peaks in the signal but have a certain degree of randomness which can be thought of as background noise.
- A nerve will adapt to an input stimulus, that is it will fire most often at the onset of the stimulus and will gradually fire less as the stimulus continues unchanged. This suggests that the auditory system is most sensitive to impulsive audio signals.
- Normal speech is very complicated and thus hard to analyse. The responses of the cochlea and the cochlear nucleus are also very complicated. The cochlear nucleus contains many different kinds of cells which respond differently to the hair cell inputs.

15.1.3 Models

All sections of the auditory chain have been effectively modelled. Due to the large amount of data we need to decide what is important and what is not. It has been found that the *timing* rather than simply the *rate* of nerve-fiber discharge is important. For example, when the input volume (Sound Pressure Level) is raised, the profile of the rate of nerve-fiber discharge in the cochlea as a function of frequency (a sort of cochlea energy spectrogram) loses information on the location of formants. By contrast, one can still find a representation the formant information at all sound-pressure levels in the synchrony of timing of nerve-fiber discharge. From this we can infer that timing is more important than the spectral energy distribution.

15.1.4 Holton's Approach

Holton's approach is to detect speech information by breaking the nerve responses up into an impulse response (I) and the formant response (S). Since the nerves are sensitive to impulses they all tend to fire at the beginning of an utterance. Nerves most sensitive to the frequencies contained in the input stimulus will continue to fire for the duration of the input while the other nerves activity will quickly die out. Tom plotted each fibres' activity against time to obtain a series of traces with increasing frequency represented by the y-axis and time the x-axis. These traces are then passed through a level detector, a mark is left if the activity is above a set threshold. From the resulting diagram we can see:

- Vertical lines: long vertical lines represent the impulse or onset of an utterance. An impulse shows up as a vertical line because all the nerves are sensitive to impulses and all fire at the same time.
- Patches: Patches of short vertical lines just after the impulse represent the formants of the input. The lowest patch is the first formant, the second lowest the second formant, etc.

To effectively detect an utterance (e.g the vowel /a/) we wait for the impulse then look at the formant patterns that follow the impulse. To detect the impulse each fibre can poll its neighbours when it fires to see if they also have fired. If the result is a yes for most of the fibres then it can be concluded that an utterance has just begun. Tom has built an electronic model that uses this scheme to extract the salient features of the incoming speech signal. Some of the important properties are:

- Still effective in the presence of noise.
- Not responsive to chords. This implies that the system is specific to speech and will ignore other non-speech signals.
- Can follow signals with varying pitch easily. Autocorrelators can not.
- Can detect voiced and non-voiced signals.

These properties suggest that such a system will be useful as the front end of a speech recognizer because it has the ability to extract the useful features of speech in the presence of other interfering signals.

15.1.5 Conclusions

- Goal - To detect speech features using auditory models.
- Approach - Local time domain phase correlation.
- Results - Robust algorithms for detection of speech.