Signals and Systems (Lab)

Project 1. Speech synthesis and perception with envelope cue

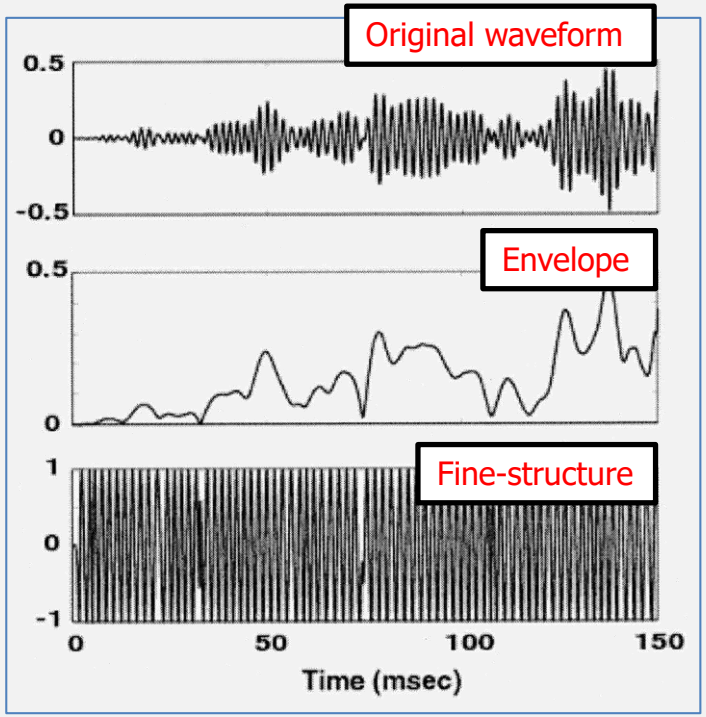
Dr. Wu Guang
Email: wug@sustc.edu.cn
Electrical & Electronic Engineering
Southern University of Science and Technology

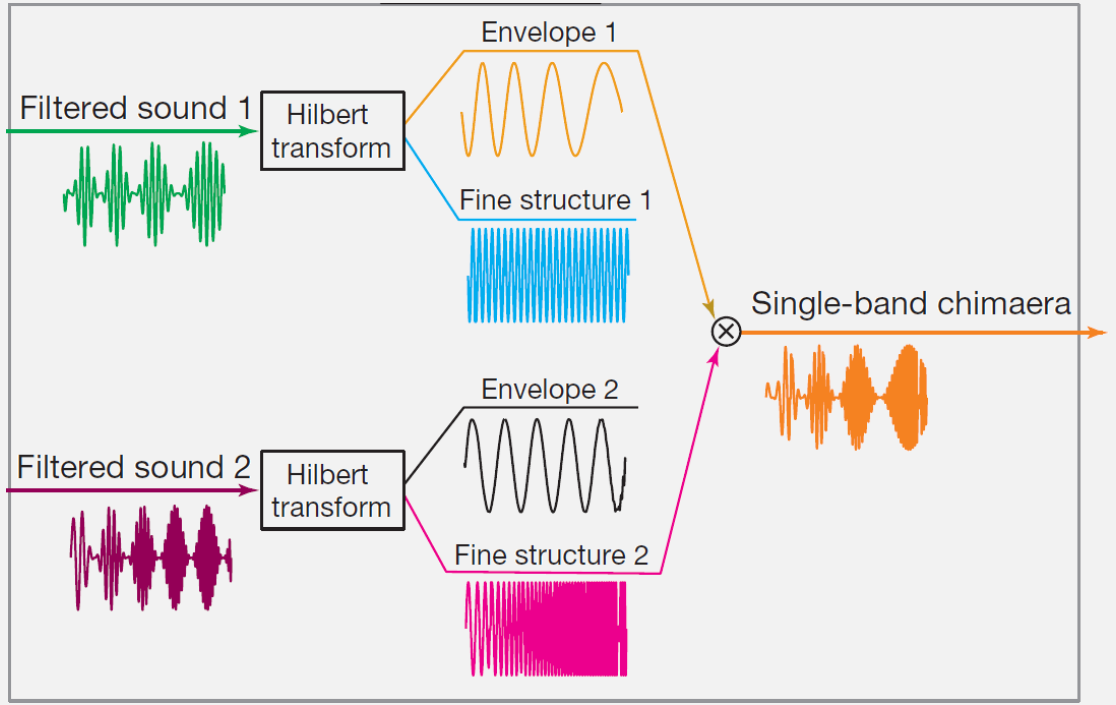
南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Overview

- In this tutorial, you will learn to synthesize a speech signal based on ***multi-band envelope cues***.
- In lab 5, you've learned:
  - how to design a low-pass/band-pass filter
  - how to extract envelope
  - how to generate a speech-spectrum shaped noise (SSN)
  - how to do energy normalization
  - how to read/save a '*.wav' file

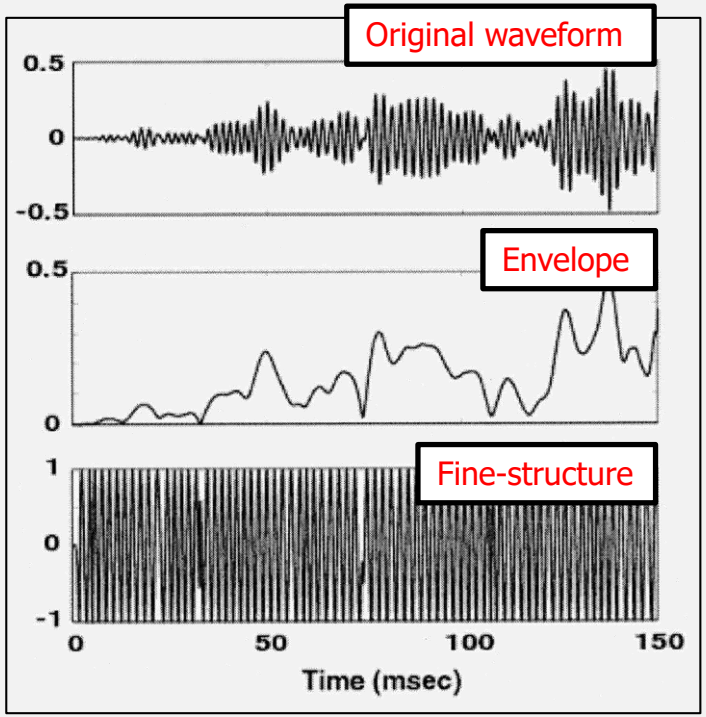
Acoustic cues of speech signal





Z. M. Smith et al. (2002), "Chimaeric sounds reveal dichotomies in auditory perception ," Science.

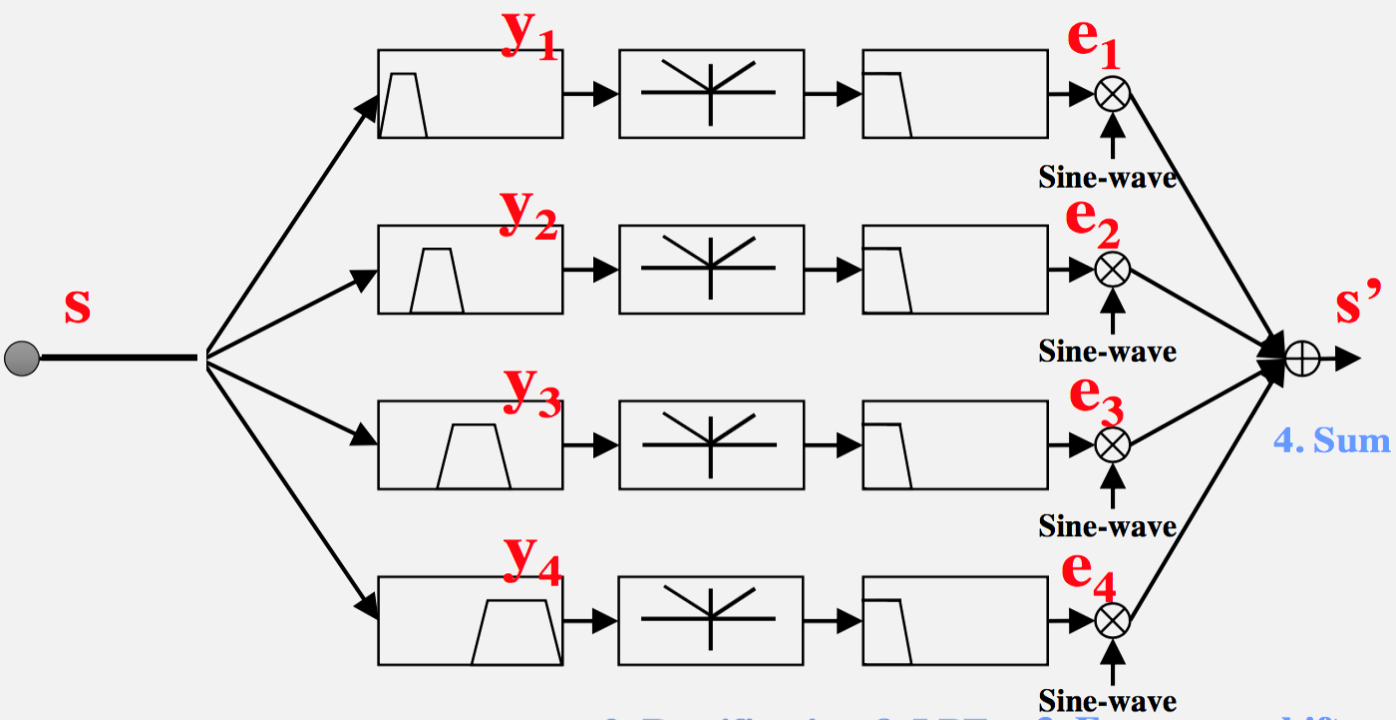
Acoustic cues of speech signal



- Envelope and fine- structure
  - **Envelope:** amplitude modulation, low- frequency
  - **Fine-structure:** frequency modulation, high- frequency

B.S. Wilson et al. (2005), "Two new directions in speech processor design for cochlear implants," Ear Hear.

Speech synthesis with envelope cue: Tone-vocoder

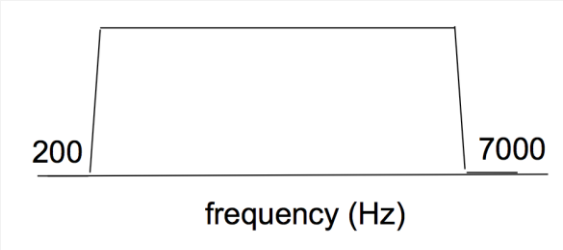


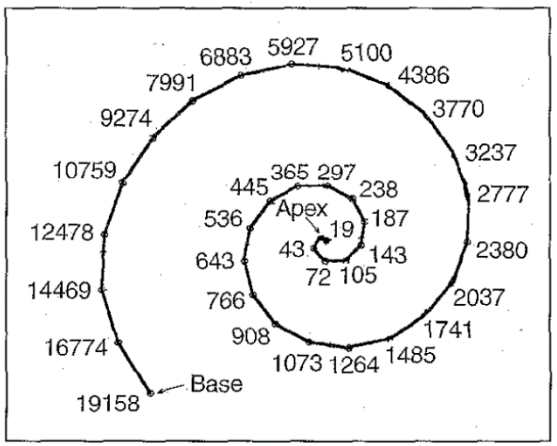
Band-pass filter

- **How to divide pass-band from 200 Hz to 7000 Hz?**
  - Equally divide the cochlea length.
  - Frequency-to-place mapping as:

$$f = 165.4 \times \left(10^{0.06 \cdot d} - 1\right)$$

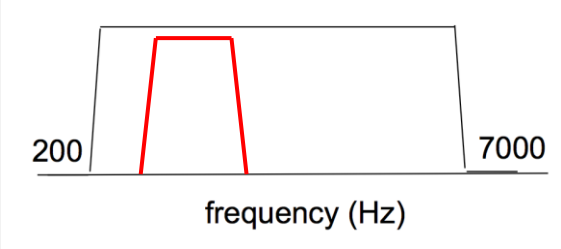
  - where f is -3 dB cut-off frequency, and d is the distance (in millimeter) along the cochlea.





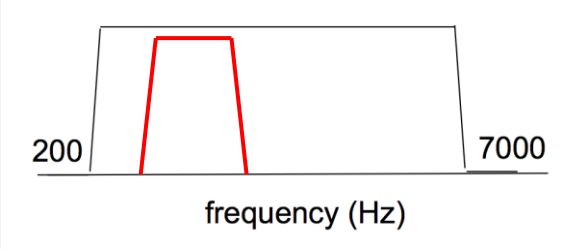
Greenwood, D.D., 1990. A cochlear frequency-position function for several species—29 years later. J. Acoust. Soc. Am. 87, 2592–2605.

Example



- Set the number of bands to N.
- For the i^{th} band
  1) Design band-pass (e.g., 400-1000 Hz) filter at band i
     >> fs=16000; %sampling frequency
     >> [b, a]=butter(4, [400 1000]/(fs/2)); %band-pass filter
  2) Do band-pass filtering at band i
     >> y= filter(b, a, s); % s is speech signal, and y is the band-passed signal at band i
  3) Do full-wave rectification, and low-pass filtering to get the envelope at band i
  4) Generate a sinewave, whose frequency equals to the center frequency of the i^{th} band-pass filter
  5) Multiply the envelope signal in 3) and sinewave in 4)
  6) Repeat for all N bands
  7) Sum up the outputs from all bands (denoting the summed outputs as s')
  8) Do energy normalization, i.e., letting the energy of s' equals to that of s
  9) Save the wavefile for signal s'

Example



- Set the number of bands to N.
- For the i^{th} band
  1) Design band-pass (e.g., 400-1000 Hz) filter at band i
     >> fs=16000; %sampling frequency
     >> [b, a]=butter(4, [400 1000]/(fs/2)); %band-pass filter
  2) Do band-pass filtering at band i
     >> y= filter(b, a, s); % s is speech signal, and y is the band-passed signal at band i
  3) Do full-wave rectification, and low-pass filtering to get the envelope at band i
  4) Generate a sinewave, whose frequency equals to the center frequency of the i^{th} band-pass filter
  5) Multiply the envelope signal in 3) and sinewave in 4)
  6) Repeat for all N bands
  7) Sum up the outputs from all bands (denoting the summed outputs as s')
  8) Do energy normalization, i.e., letting the energy of s' equals to that of s
  9) Save the wavefile for signal s'

Project tasks -1

- **Sentences for pro 1: ‘C_01_01.wav’ & ‘C_01_02.wav’**
- **Task 1**
  - Set LPF cut-off frequency to 50 Hz.
  - Implement tone-vocoder by changing the number of bands to N=1, N=2, N=4, N=6, and N=8.
  - Save the wave files for these conditions, and describe how the number of bands affects the intelligibility (i.e., how many words can be understood) of synthesized sentence.

Project tasks -2

- **Task 2**
  - Set the number of bands N=4.
  - Implement tone-vocoder by changing the LPF cut-off frequency to 20 Hz, 50 Hz, 100 Hz, and 400 Hz.
  - Describe how the LPF cut-off frequency affects the intelligibility of synthesized sentence.

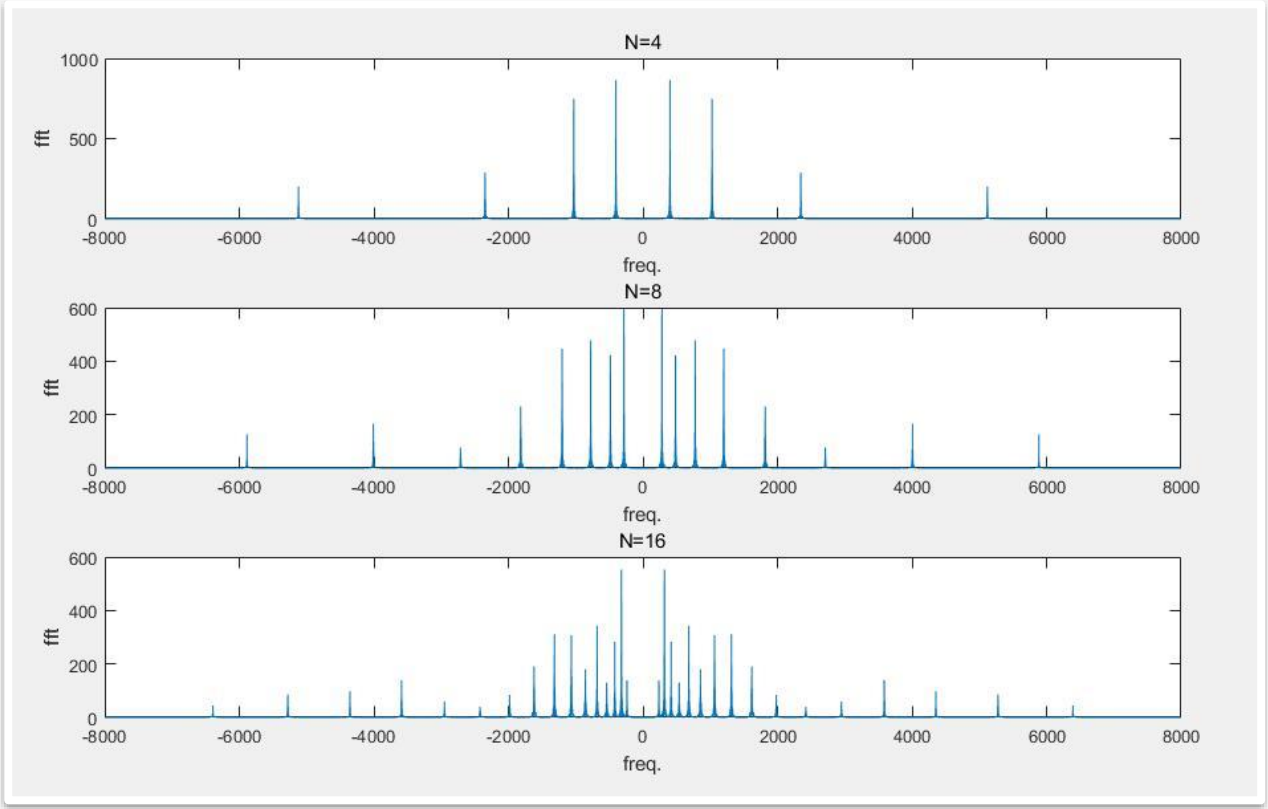
Project tasks -3

- **Task 3**
  - Generate a noisy signal (summing clean sentence and SSN) at SNR -5 dB.
  - Set LPF cut-off frequency to 50 Hz.
  - Implement tone-vocoder by changing the number of bands to N=2, N=4, N=6, N=8, and N=16.
  - Describe how the number of bands affects the intelligibility of synthesized sentence, and compare findings with those obtained in task 1.

Project tasks -4

- **Task 4**
  - Generate a noisy signal (summing clean sentence and SSN) at SNR -5 dB.
  - Set the number of bands to N=6.
  - Implement tone-vocoder by changing the LPF cut-off frequency to 20 Hz, 50 Hz, 100 Hz, and 400 Hz.
  - Describe how the LPF cut-off frequency affects the intelligibility of synthesized sentence.

Results



Application: Cochlear Implants

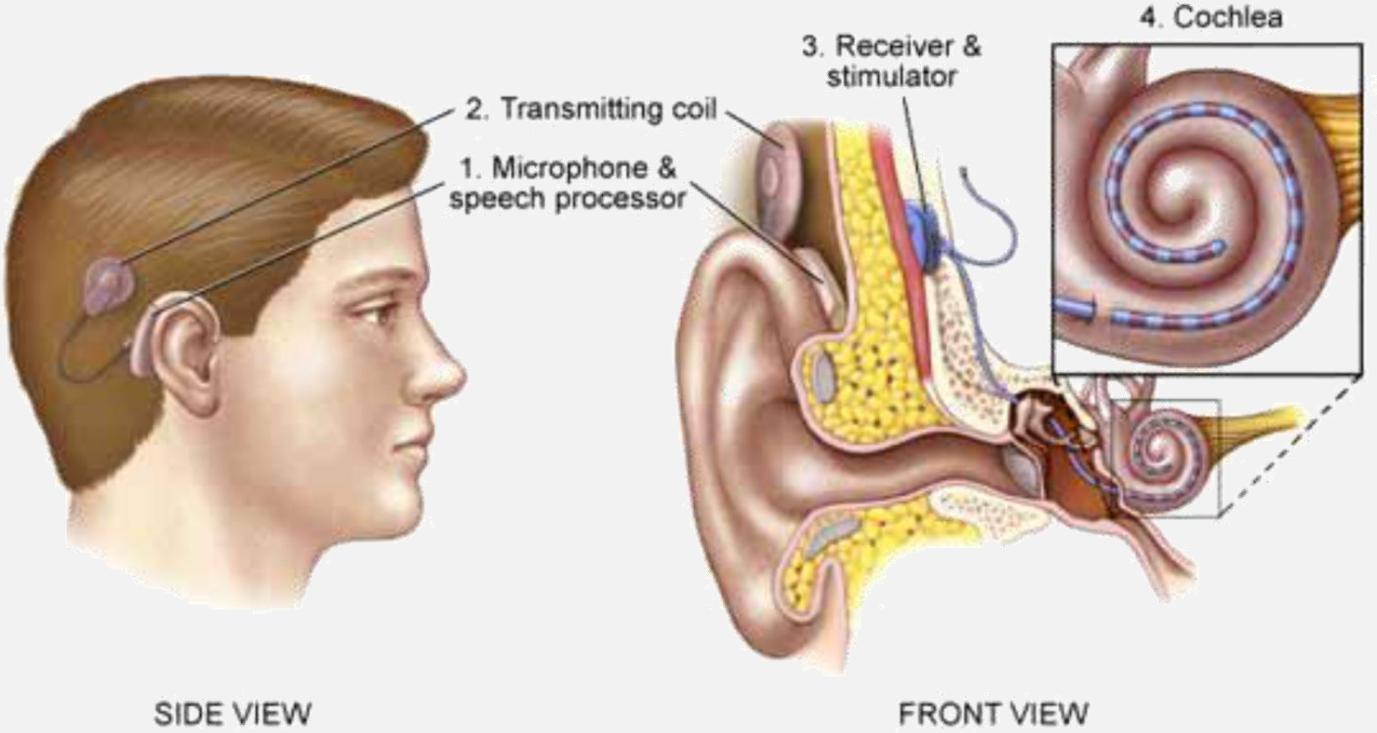
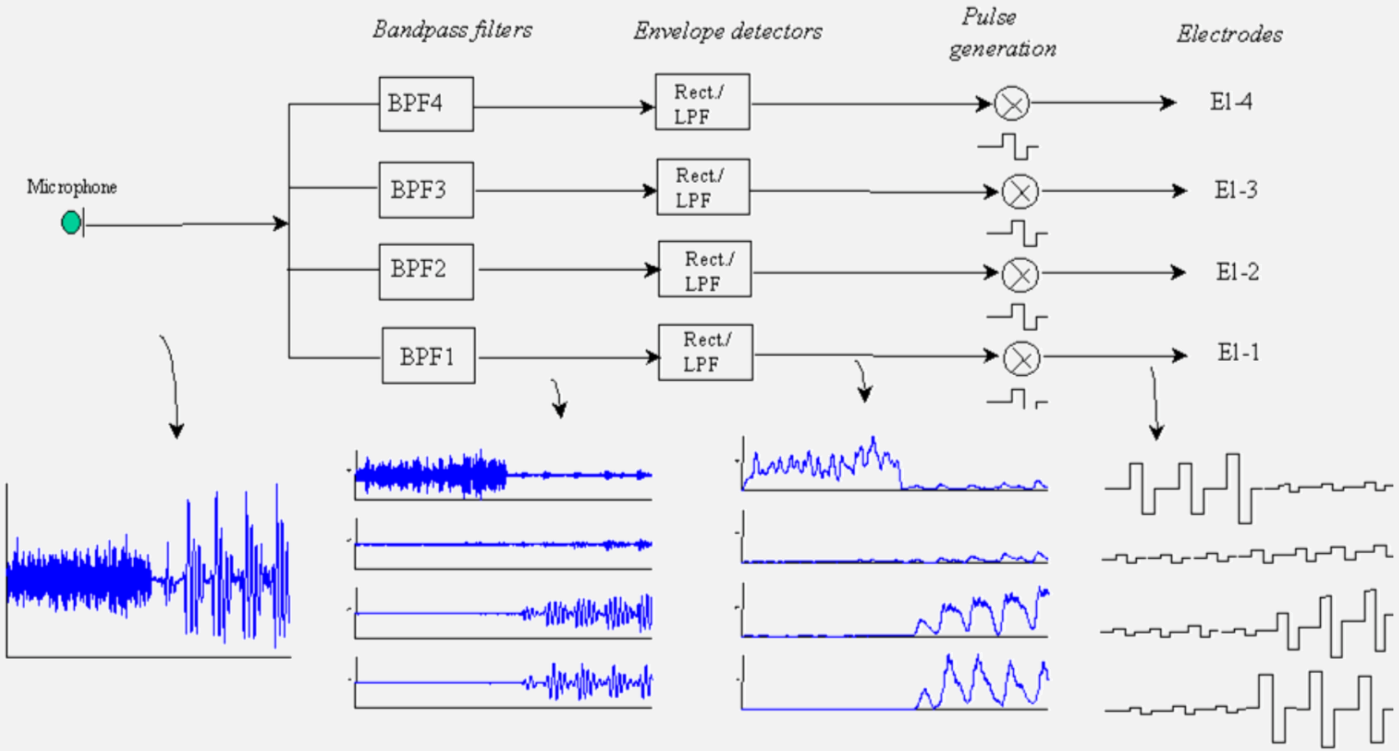


Image from: http://www.mayoclinic.org/

Speech processing in cochlear implants



Loizou (1999) “Introduction to cochlear implants,” IEEE Eng. in Med. and Bio. Mag., 18, 32-42.

Speech Recognition with Primarily Temporal Cues

Robert V. Shannon,* Fan-Gang Zeng, Vivek Kamath, John Wygonski, Michael Ekelid

Nearly perfect speech recognition was observed under conditions of greatly reduced spectral information. Temporal envelopes of speech were extracted from broad frequency bands and were used to modulate noises of the same bandwidths. This manipulation preserved temporal envelope cues in each band but restricted the listener to severely degraded information on the distribution of spectral energy. The identification of consonants, vowels, and words in simple sentences improved markedly as the number of bands increased; high speech recognition performance was obtained with only three bands of modulated noise. Thus, the presentation of a dynamic temporal pattern in only a few broad spectral regions is sufficient for the recognition of speech.

The recognition of speech has been thought to require frequency-specific (spectral) cues. Spectral energy peaks in speech (formants), for example, reflect the resonant properties of the vocal tract and thus provide acoustic information on the production of the speech sound. However, efforts to identify acoustic cues that convey phoneme identity reliably under various listening conditions and with various talkers

16, 50, 160, and 500 Hz were used for envelope extraction to evaluate the effect of reducing the bandwidth of temporal envelope information. The envelope signal was used to modulate white noise, which was then spectrally limited by the same bandpass filter used for the original analysis band (*7*). Thus, temporal and amplitude cues were preserved in each spectral band, but the spectral detail within each band was

under conditions of reduced spectral cues, slowly varying temporal information (<50 Hz) can yield relatively high speech recognition performance. This result is consistent with the observation of poor speech discrimination in children who have central processing disorders that disrupt temporal processing in the 20- to 50-ms range (*10*).

The specific reception of three speech features—voicing, manner, and place of articulation—was evaluated by information transmission analysis (*11*) on the consonant confusion matrix (Fig. 3). Information received on voicing and manner increased from one to two bands, to >90%, with no further improvement as the number of bands increased to three or four. Thus, binary information on the spectral distribution of energy, when combined with temporal cues, is sufficient to convey almost all information on voicing and manner. Voicing and manner have similar patterns of results as a function of the number of spectral bands, and both cues show maximum performance with only two spectral bands; these findings reinforce the hypothesis (*4*) that both categories of information, although labeled according to vocal produc-

Organization

- **Each group consists of four students.**
- **Each group need present the two Lab projects (submit reports for both projects):**
  - The presentation date is 2020.05.21 for project 1
- **Each presentation is 10 minutes (including Q & A)**
  - All team members need to contribute to the presentation.
  - Presentation in English (recommended) or Chinese.

The presentation should...

- Introduce
  - team
  - objective of the project
  - background review (search more additional information)
  - methodology
- Discuss
  - what you have learned from this study?
  - problems during this project and your solution
  - investigation beyond project tasks
  - critical thinking
- Present
  - relevant data, figure, etc.
  - the results for project tasks (e.g., with demo, Figure, etc.)
  - interpretation of project findings
- Appendix (if any)
- Team effort (e.g., individual contribution)
- Reference
- Q & A (answer questions raised from audience)

Questions