SUSTech Southern University of Science and Technology

Project 1.
Speech Synthesis And Perception With Envelope Cue

Cheng PENG
Department of Biomedical Engineering
pengc@sustech.edu.cn
Fall 2020

Lab Schedule

2020年	周 次	一	二	三	四	五	六	日
11月	第10周	16 初二	17 初三	18 初四	19 初五	20 初六	21 初七	22 小雪
	第11周	23 初九	24 初十	25 十一	26 十二	27 十三	28 十四	29 十五
12月	第12周	30 十六	1 十七	2 十八	3 十九	4 二十	5 廿一	6 大雪
	第13周	7 廿三	8 廿四	9 廿五	10 廿六	11 廿七	12 廿八	13 廿九
	第14周	14 三十	15 11月小	16 初二	17 初三	18 初四	19 初五	20 初六
	第15周	21 冬至	22 初八	23 初九	24 初十	25 十一	26 十二	27 十三

← **lab 5**

← Intro to project 1

← Q&A, intro to project 2

← Presentation 1

← Project 2 Q&A

← Presentation 2

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 read/save a '*.wav' file
  - 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

Application Background: Cochlear Implants

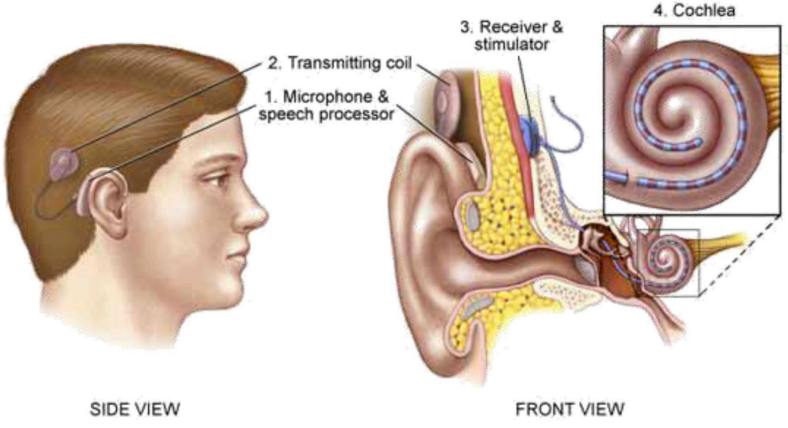


Image from: http://www.mayoclinic.org/

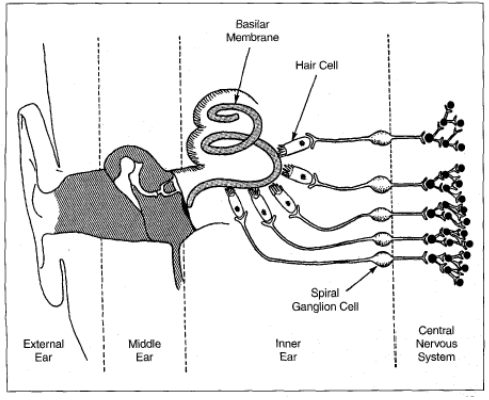
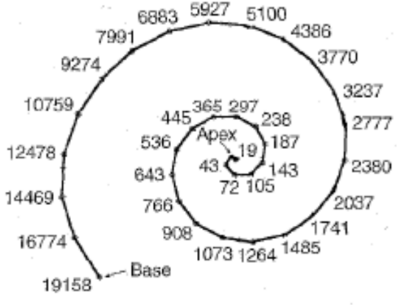
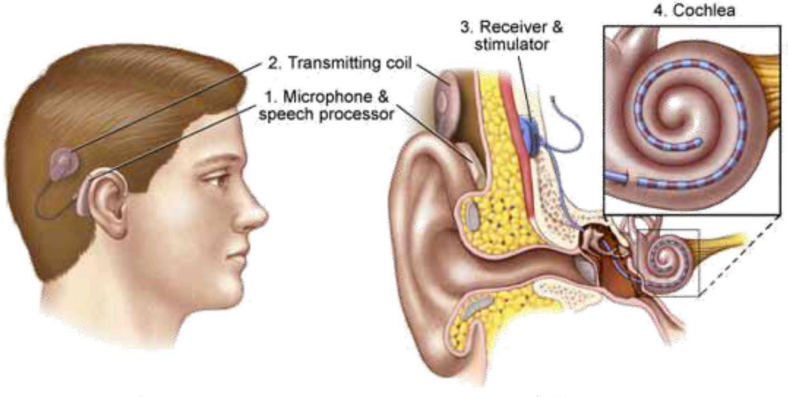
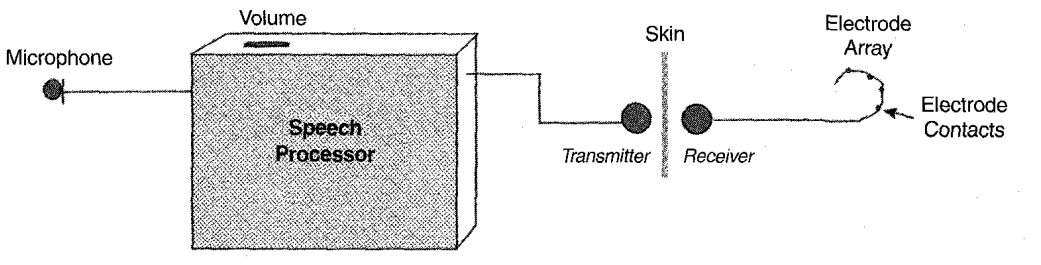


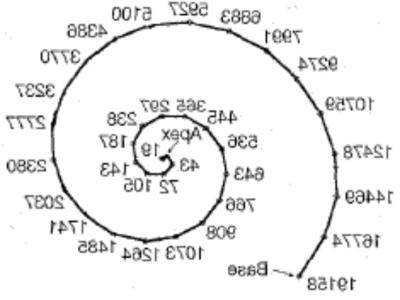
Diagram of the basilar membrane showing the base and the apex. The position of maximum vibration in response to sinusoids of different frequency (in Hz) is indicated.



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

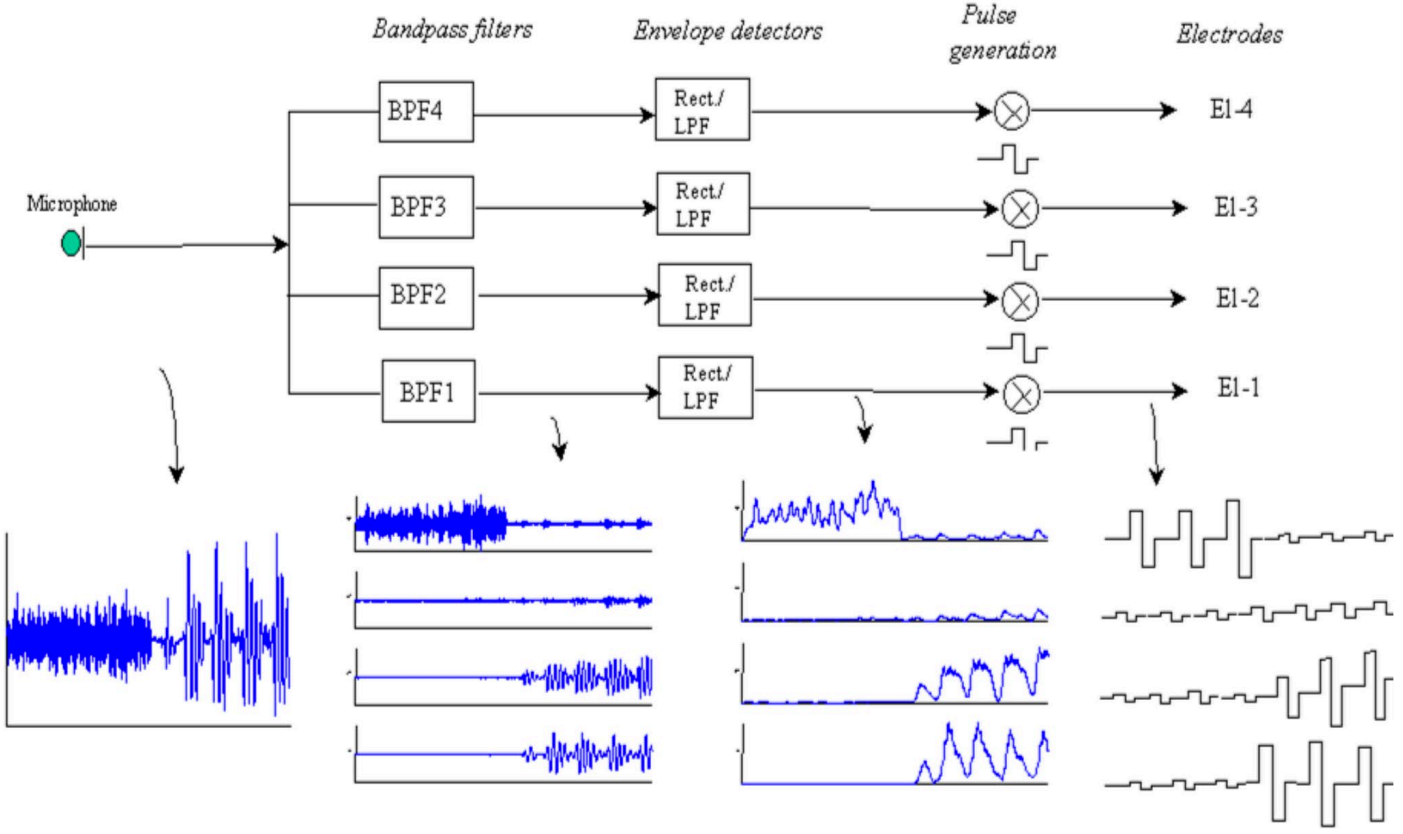






Loizou (1999) “Introduction to cochlear implants,” IEEE Eng. in Med. and Bio. Mag., 18, 32-42.

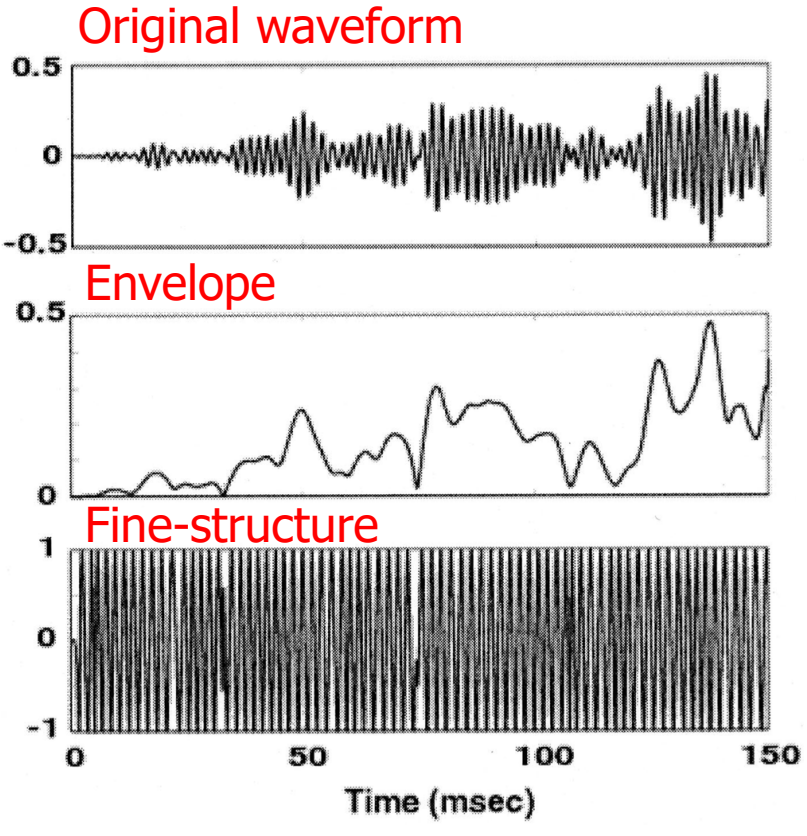
Speech processing in cochlear implants



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

Acoustic cues of speech signal

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



Greenwood, D.D., J. Acoust. Soc. Am. 1990

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

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.

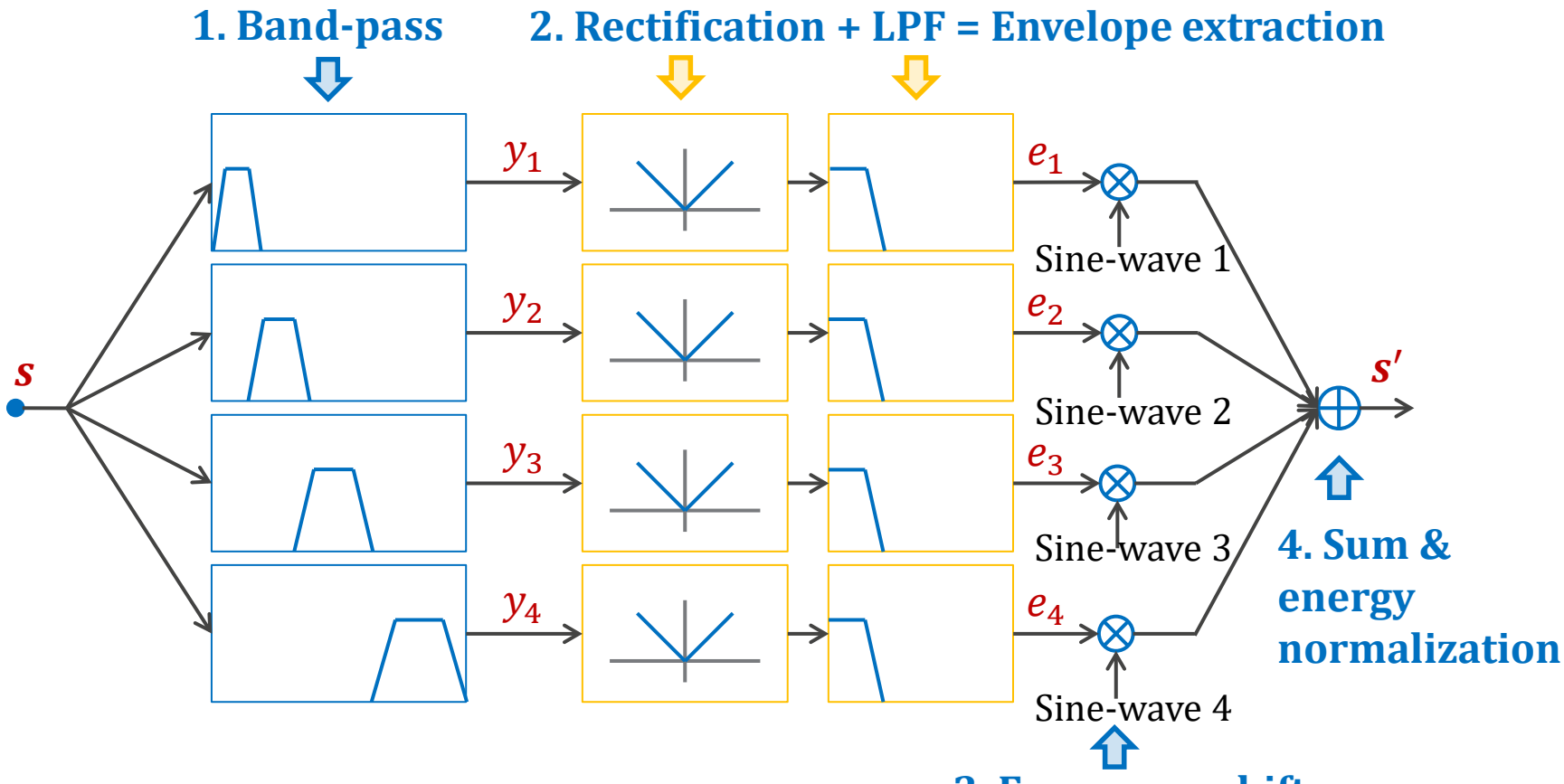
The recognition of speech has been thought to require frequency-specific (spectral) cues. Spectral energy peaks in speech (formants), for example, reflect the resonant properties of the vocal tract and thus provide acoustic information on the production of the speech sound. However, efforts to identify 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*).

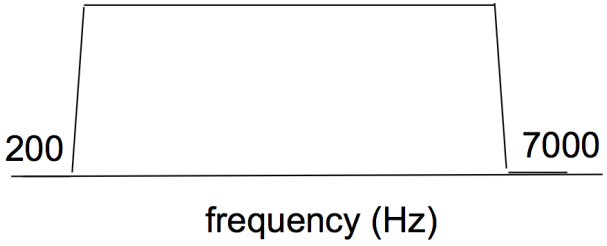
The specific reception of three speech features—voicing, manner, and place of articulation—was evaluated by information transmission analysis (*11*) on the consonant confusion matrix (Fig. 3). Information received on voicing and manner increased from one to two bands, to >90%, with no further improvement 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-

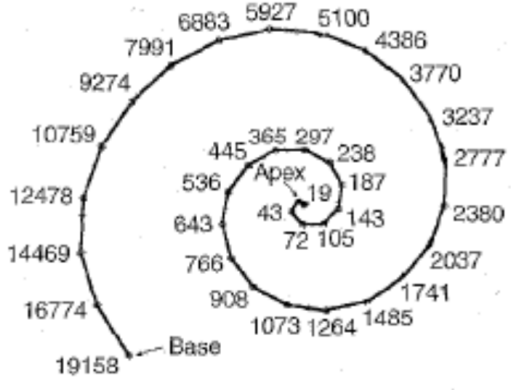
Speech synthesis with envelope cue: Tone-vocoder



1. How many channels/bands? --- N channels/bands, varies in your research
2. The pass band for each channel, how to determine?
   - Two frequencies : the lower cutoff frequency and the upper one
   - Vary among channels
3. Band pass filter design, type? order?
   - 4-order Butterworth band-pass filter
4. Low pass filter for envelope extraction
   - Same LPF for all channels/bands
   - Type? Order? --- 4-order Butterworth low pass
   - Cutoff frequency? --- varies in your research
5. Sine wave frequency?
   - Varies among channels
   - Is equal to the center frequency of band-pass filter of for each channel

Tone-vocoder (Cont. -1): determine the pass band for each channel





- The whole frequency range interested:
  - **200 Hz to 7000 Hz**
- 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 cutoff frequency, and d(mm) is the distance 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.

Multiple Choice(single) Points: 1

Setting

If N = 1, the pass band for BPF is ?

A 200Hz to 7000Hz

B 20Hz to 20kHz

C ???

D -infinity to infinity

Submit

Multiple Choice(single) Points: 1

Setting

If N = 2, the pass bands for BPFs are ?

A 200Hz to 3600Hz, 3600Hz to 7000Hz

B 200Hz to 1452.7Hz, 1452.7Hz to 7000Hz

C 200Hz to 7000Hz, 7000Hz to 14kHz

D -infinity to 0, 0 to infinity

Submit

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

- $d_{200} = {log_{10}\left(\frac{200}{165.4}+1\right)}/{0.06}$
- $d_{7000} = {log_{10}\left(\frac{7000}{165.4}+1\right)}/{0.06}$
- $d_{\text{middle}} = \frac{(d_{200}+d_{7000})}{2}$
- $f_{middle} = 165.4 \times \left(10^{0.06 \cdot d_{middle}} - 1\right)$

Tone-vocoder (Cont.-2)

- Set the number of bands to N.
- For the i^{th} band, $i = 1,2, \dots, N$
  1) Design band-pass (e.g., 400-1000 Hz) filter at i^{th}band
     >> fs=16000; %sampling frequency, depends on your signals
     >> [b, a]=butter(4, [400 1000]/(fs/2)); %band-pass filter
  2) Do band-pass filtering at i^{th} band
     >> y= filter(b, a, s); % s is speech signal, y is the band-passed signal at i^{th} band
  3) Do full-wave rectification, and low-pass filtering to get the envelope at i^{th} band
  4) Generate a sinewave, whose frequency equals to the center frequency of the i^{th} band-pass filter $f_c = (f_{lower} + f_{upper})/2$
  5) Multiply the envelope signal in 3) and sinewave in 4)

Tone-vocoder (Cont.-3)

- Repeat for all N bands
- Sum up the outputs from all bands (denoting the summed outputs as s')
- Do energy normalization, i.e., let the energy of s' equals to that of s
- Save the *.wav file for signal s'

Project tasks -1

- Sentences for Project 1: ‘C_01_01.wav’ & ‘C_01_02.wav’
- Task 1
  - Set LPF (for envelop extraction) cutoff frequency to 50 Hz.
  - Implement tone-vocoder by changing the number of bands to $N = 1, N = 2, N = 4, N = 6,$ and $N = 8$. (bandpass filters vary)
  - 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 cutoff frequency to 20 Hz, 50 Hz, 100 Hz, and 400 Hz.
  - Describe how the LPF cutoff 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.

Organization

- ≤ 4 students/group.
- Each group needs to present 1 of the 2 Lab Projects (but submit reports for both projects):
  - The presentation date is Dec. 10th for Project 1, and Dec. 24th for Project 2.
  - Please inform TA your choice before the class in next week
- Each presentation should be within 10 minutes
  - All team members need to contribute to the presentation.
  - Presenting in English (recommended) or Chinese.

The presentation and report should (but not limited to) ...

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

Grading according to ...

- Introduction 20%
- Results 30%
- Discussion 30%
- Q&A 10%
- Overall (e.g., PPT design and presentation) 10%

- Report deadline
  - for Project 1: Dec 13th, 2:00 pm
  - for Project 2: Dec 27th, 2:00 pm

- Any questions?

- Generate a sine wave
  - sine wave of frequency = f1, length = N points, and sample rate = fs

```
n = 1:N;     % sample counts sequence
dt = n * (1/fs);     % time sequence (in second)
sinesig = sin(2*pi*f1*dt)     % get the sine wave
```