Teknologi Pengolahan Suara –

Automatic Speech Recognition (ASR)

Dr. ir. Kuntjoro Pinardi

International University Liaison Indonesia

IULI

bahasakita

Apa yang kita harapkan dari ASR

- Speech-to-text transcription : Pengolahan Suara Audio Menjadi Text / Deret Kata-Kata
- Kemiripan bunyi kata – kedekatan bunyi ucapan Bang - Bank
- Speaker diarization / Speaker ID: Kemampuan identifikasi pembicara
- Speech recognition: Kepastian apa yang diucapkan sama dengan apa yang akan dituliskan.
- Paralinguistic aspects: how did they say it? (timing, intonation, voice quality)

Referensi: http://www.inf.ed.ac.uk/teaching/courses/asr/lectures-2021.html

Hambatan dan Tantangan dalam ASR

- Device atau Aplikasi yang kita kenal



- Keterbatasan Resources untuk Hasil yang Maksimal – Memory dan Storage
- Variasi Sumber Suara – Lingkungan, Aksen, Perangkat Perekam / Audio
- Kondisi Bicara – Spontan atau Baca Text
- Ragam Bahasa – Campuran Bahasa dalam Percakapan

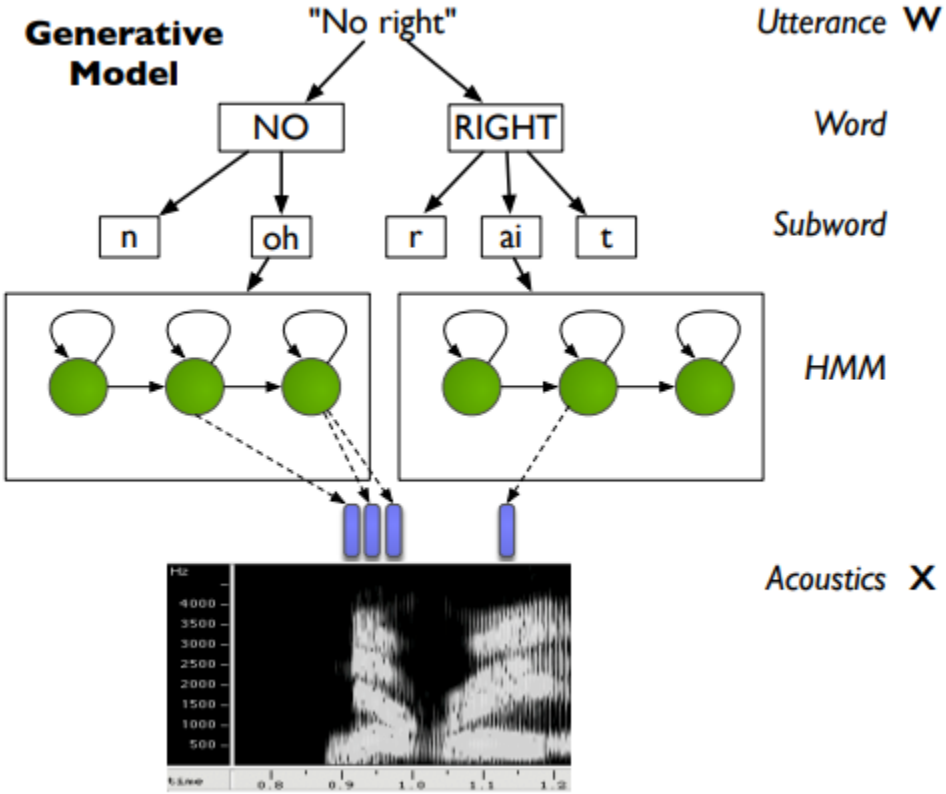
Proses Pengenalan Suara

- Audio suara merupakan sinyal elektronik dengan karakter vector akustik dengan keluaran teks yang sesuai.
  - acoustic feature vectors X and the output word sequence as W

- Pengolahan sinyal memastikan probability ada teks W dengan input vector akustik X

- Pengolahan Sinyal secara metoda statistic harus dilakukan dengan cara melakukan proses Training untuk mendapatkan model Corpus untuk setiap utterances / bunyian huruf, suku kata dan kata.

Proses Pengenalan Suara



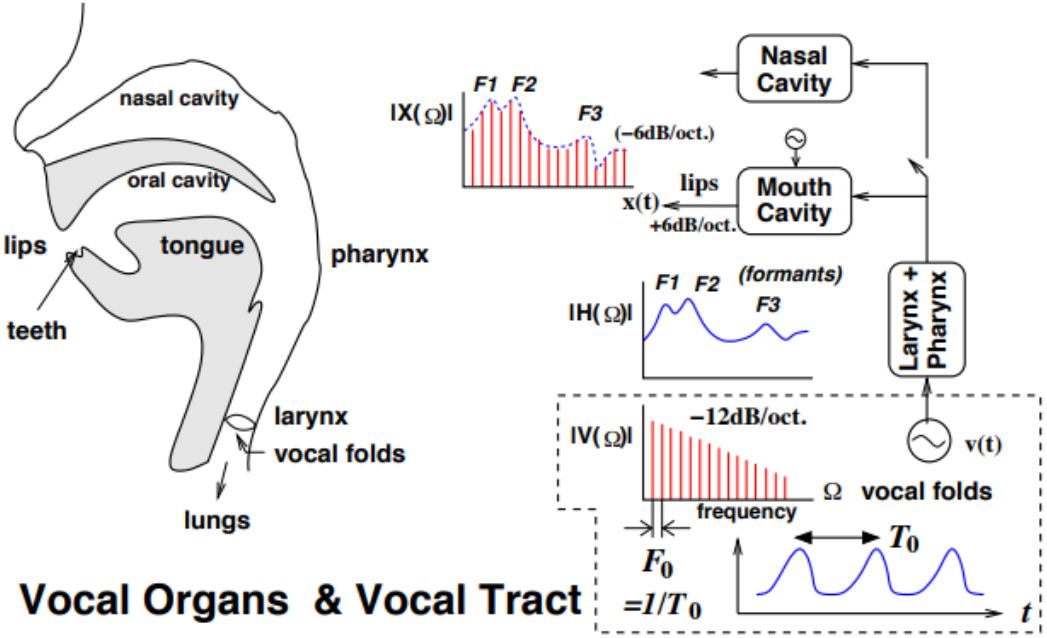
Pemodelan Statistik

- Pengolahan sinyal memastikan probability ada teks W dengan input vector akustik X atau secara Statistik :

$$W^* = \arg\max_W P(W \mid X)$$

- Apakah kita harus menghitung secara continue ?

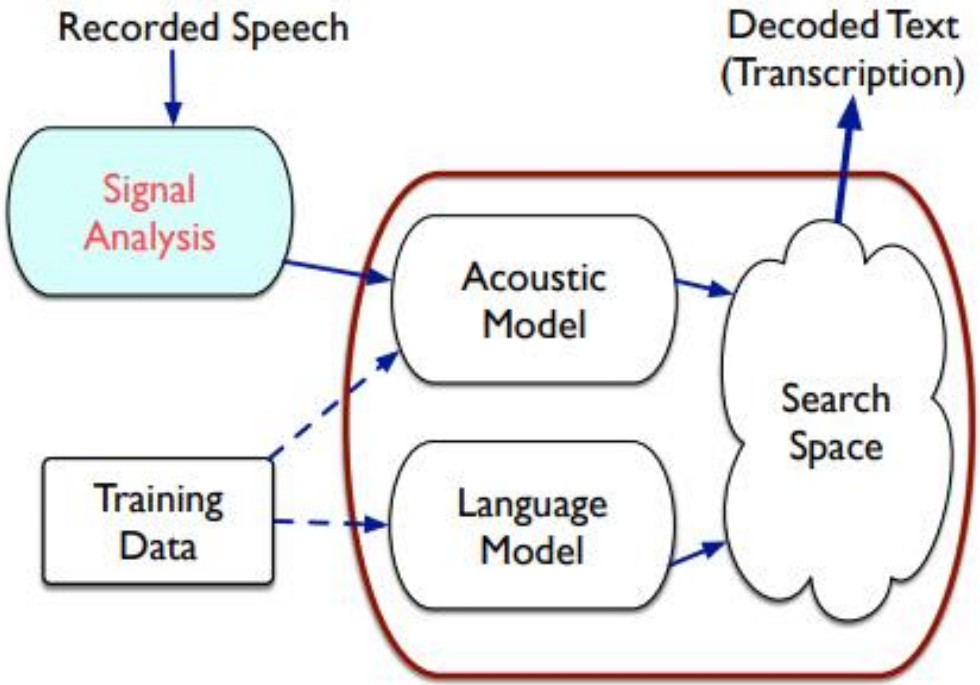
Karakter Suara Manusia



Vocal Organs & Vocal Tract

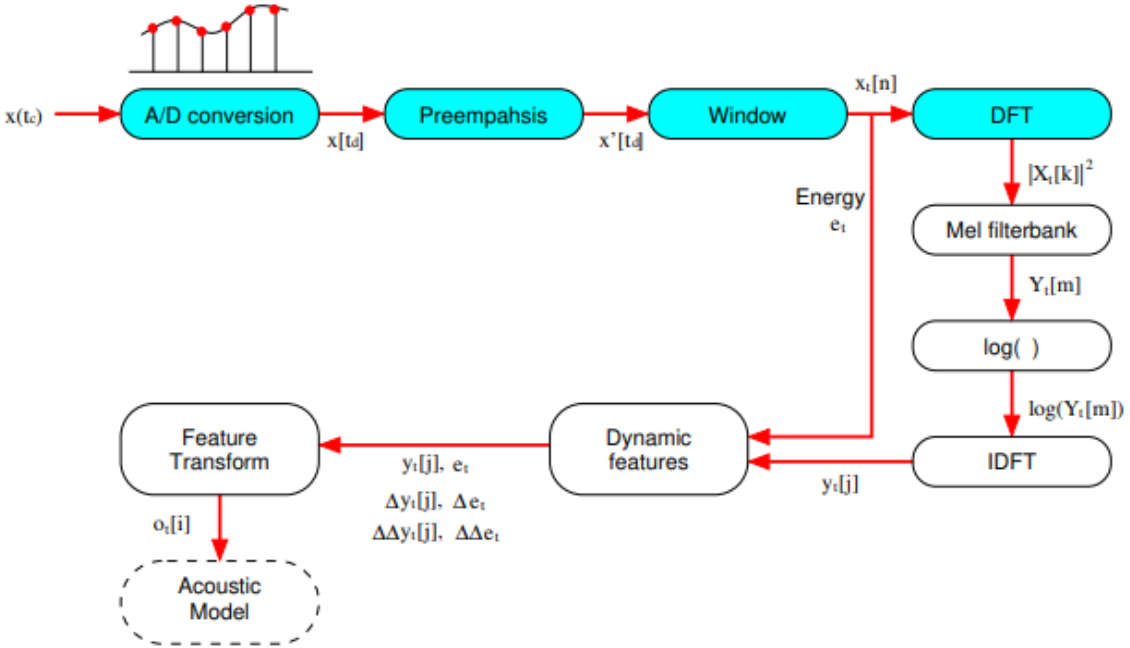
(F_0 : fundamental frequency)

Diagram Alur Proses ASR

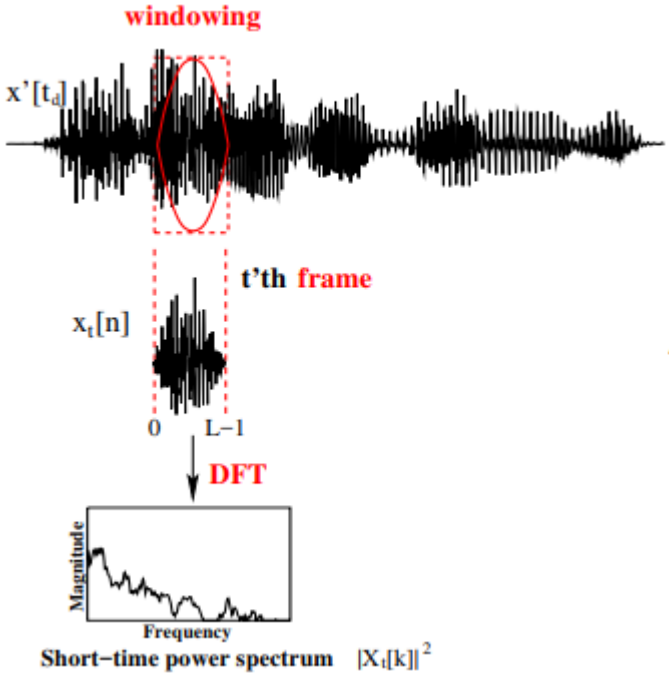


Pembangunan Model Akustik

- Penentuan Karakteristik Sinyal Suara Manusia
  - Mel Frequency Cepstral Coefficients (MFCC)
    - A short-period power spectrum of sound wave representation. **Mel frequencies** are based on the human ear's respond to bandwidth variation below 1 kHz **frequencies** .



Pembangunan Model Akustik



- For ASR:
  - frame width $\sim 25ms$
  - frame shift $\sim 10ms$

Applying Bayes' Theorem:

$$P(W|X) = \frac{p(X|W)\,P(W)}{p(X)}$$

$$\propto p(X|W)\,P(W)$$

$$W^* = \arg\max_W \underbrace{p(X|W)}_{\text{Acoustic model}} \underbrace{P(W)}_{\text{Language model}}$$

Implementasi ASR

- Dimulai sejak 1950 an oleh Bell Labs dan IBM
- Opensource Kaldi 2009 untuk ASR Generik Semua Bahasa *“Low Development Cost, High Quality Speech Recognition for New Languages and Domains"* oleh Prof. Dan Povey (Microsoft, John Hopkins dan Xiaomi)

Referensi:

https://kaldi-asr.org/doc/history.html
https://www.topionetworks.com/people/daniel-povey-5a9a6805105eb53d502a272fg

Implementasi ASR

- Dimulai sejak 1990 an oleh BPPT
- Universitas sejak 2015 ITB dan UI
- Kerjasama Google bersama UGM - ?

- Perusahaan Komersial
  - Bahasakita – eks BPPT 2014
  - Prosa – ITB 2018
  - Botika – UGM 2020
  - Widya Wicara - 2020

Tantangan ASR - 2021

- On Device ASR
- Interest to work on On Device ASR with Prasimax
- Tools: https://fosspost.org/open-source-speech-recognition/
- https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-predictions-for-2021-and-beyond/
- https://www.ai-startups.org/top/speech_recognition/
- https://www.iotworldtoday.com/2021/01/19/edge-nlp-is-about-doing-more-with-less/

Implementasi ASR – From Scratch - Diskusi

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016 9-12 May 2016, Yogyakarta, Indonesia

Towards Robust Indonesian Speech Recognition with Spontaneous-Speech Adapted Acoustic Models

Devin Hoesen* , Cil Hardianto Satriawan, Dessi Puji Lestari, Masayu Leylia Khodra - Institut Teknologi Bandung, Jl. Ganeca No. 10, Bandung 40115, Indonesia

Implementasi ASR – From Scratch - Diskusi

; http://www.inf.ed.ac.uk/teaching/courses/asr/labs-2019.html

Implementasi ASR – From Scratch - Diskusi

Building a Speech and Text Corpus of Turkish: Large Corpus Collection with Initial Speech Recognition Results

Huseyin Polat * and Saadin Oyucu Department of Computer Engineering, Faculty of Technology, Gazi University, 06560 Ankara, Turkey;