DLP - Week - 8

L1 Recap of Previ. discussion

- speech is much better than text.
- chatgpt is old multi-modal LLMs.
- language identification
- speaker diarization
- automatic speech recogni"
  ↳ this is the half part. What if I want to listen to result? That is TTS (Text To Speech)

L2 Intro to TTS

- Text to Speech conversion
  - text → TTS → output is speech
    ↳ synthesized audio → natural and intelligible speech.
- Applica"
  1. speech based technologies
  2. help people with literacy difficulties
  3. aid the visually challenged.

- Frameworks.
  1. unit selection synthesis (USS)
  2. hidden markov model based (HTS)
  3. neural network based (convention)
  4. end to end (E2E)

- Phases
  1. Training phase
  2. synthesis / testing phase

L3. History of TTS

- long history
- in 1990, → first female voice synthesizer (successful)
- none very powerful to achieve TTS.
- speech has phonemes (they were considered as states) → to approx. the acoustic vectors.
- DB
  <text, audio> pairs → continuous speech.

L4. Analogous of ASR & TTS

Text

How are you?

characters

Speech

phonemes

in spoken language gets concatenated by them.

/iy/ → ε

/ih/ → ε/2

ASR

editable text

Today, we talk about the inverse problem.

Text

TTS

Speech

- maybe gender
- big actors
- deepfake (ethical issues)
- dubbing applications.

L5 Phoneme To Text

Lexicon / Dictionary

ঈশ্বর → /ah/ /m/
good → /g/ /uh/ /d/

} good to do for 1 lang. predict.

Text → (lexicon) → seq. of phonemes

ASR → [illegible] → [illegible] embeddings

Mel Filter Bank Features.

Neural Network

Ph1 Ph2 Ph3

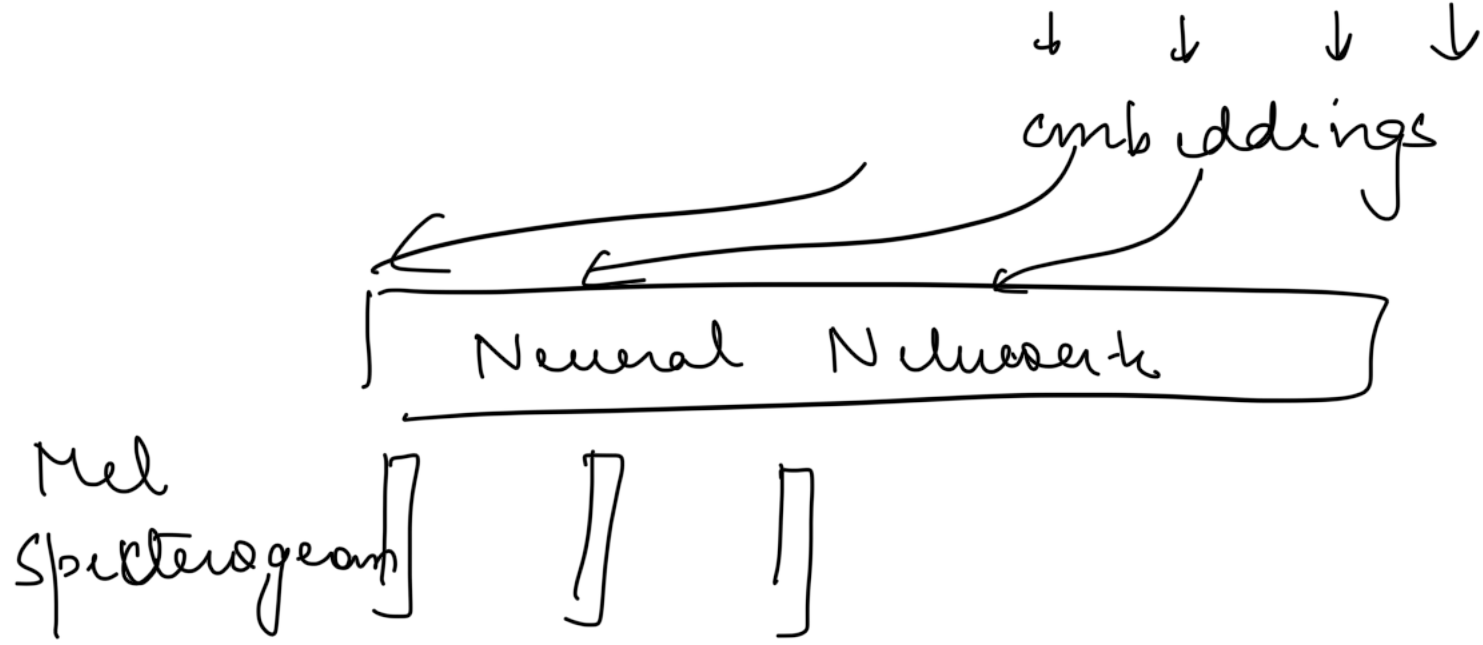
probabs. the text

Text

- In text to speech

How are you

(lexicon) → /h/ /ow/ /a/ /u/



Vocoder

↳ this synthesizes the speech.

+ Add. speaker (speaker embedding using speaker vector representation)

- L6 to L8 demo