Deep Learning in Speaker, Speech and Language Technologies

2020/05/27

Prof. Brian Mak

Department of Computer Science and Engineering

Hong Kong University of Science and Technology 香港科技大學HKUST

My Research Areas

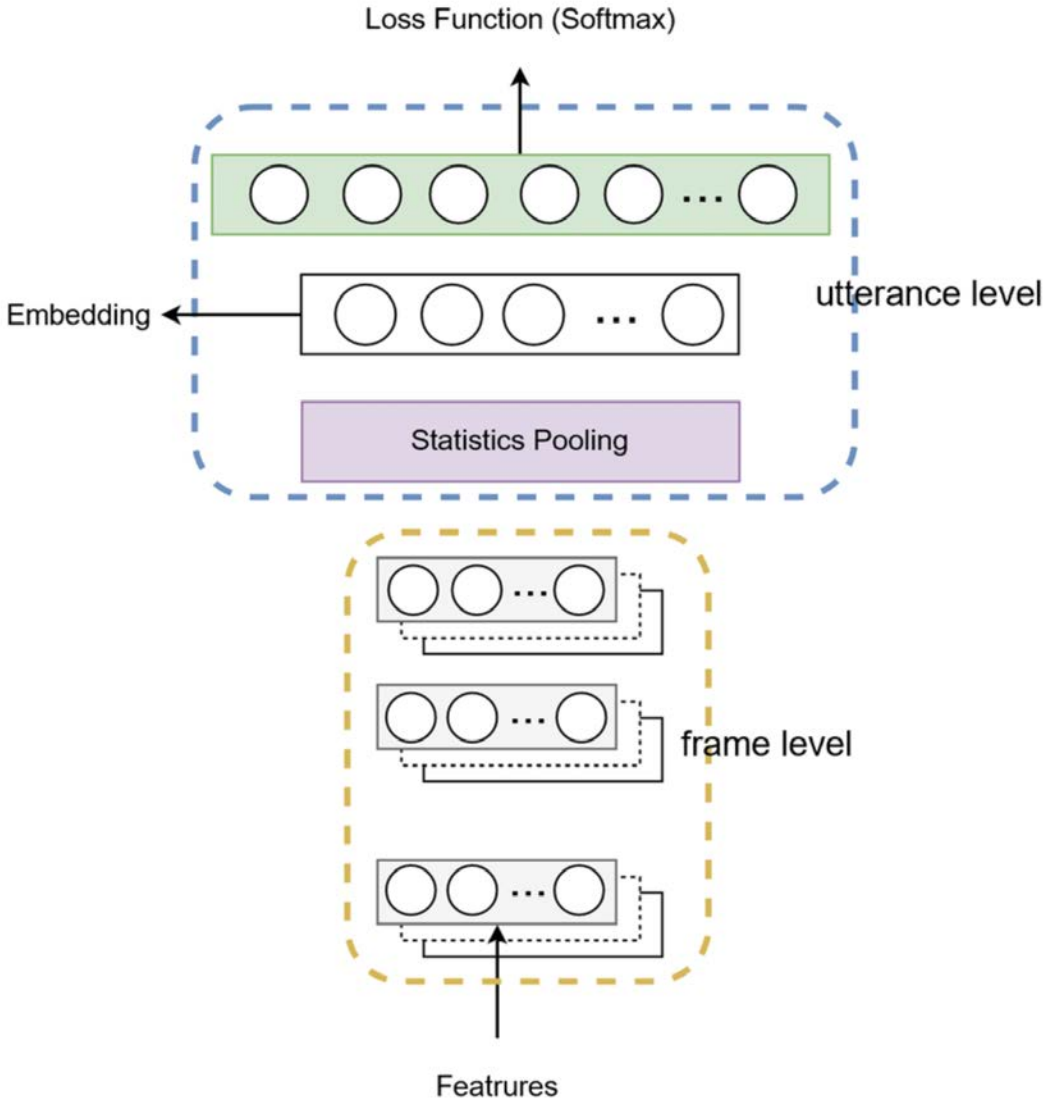
- automatic speech recognition
- sign language recognition, translation, generation
- speaker verification/recognition
- speaker diarization
- (multi-lingual, multi-speaker) speech synthesis
- voice conversion
- lip reading
- NLP: multi-lingual document representation

Recent Projects

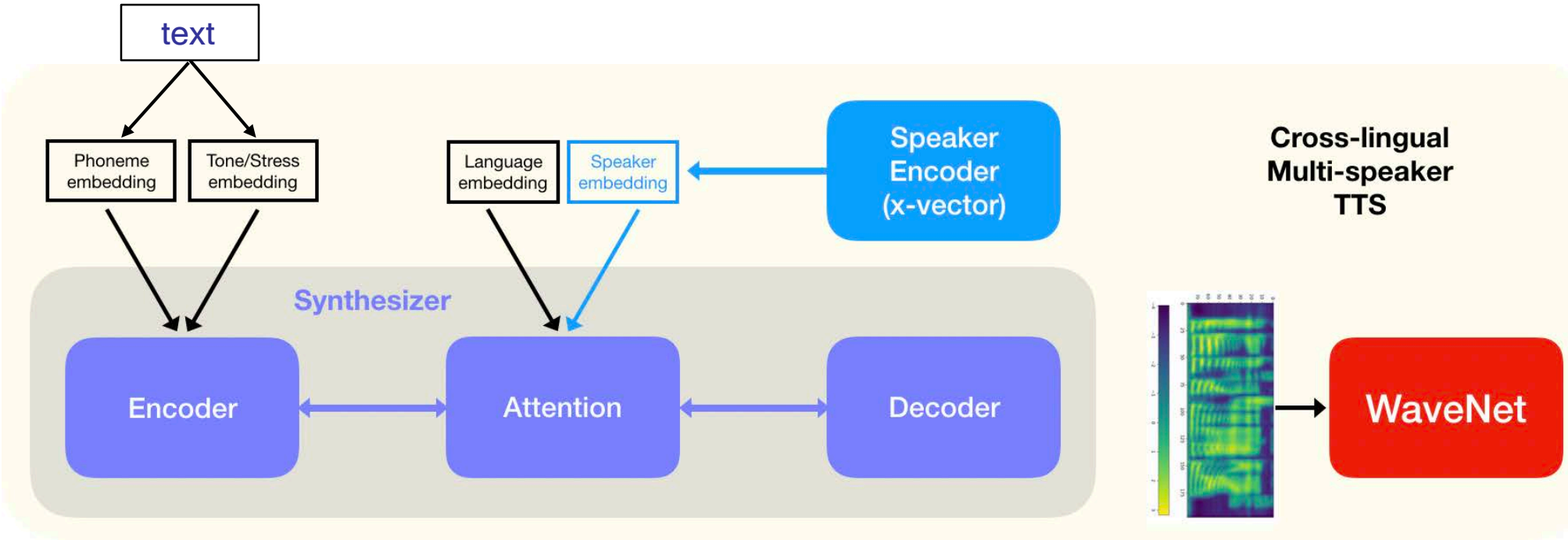
- RGC Theme-based: Research and Development of Artificial Intelligence in Extraction and Identification of Spoken Language Biomarkers for Screening and Monitoring of Neurocognitive Disorders (以人工智能提取和鑑定口語生物標誌物供神經認知障礙篩查和監測的研究及技術開發)
  - elderly speech recognition
  - speaker diarization
- LSCM: Elderly-friendly Text-to-speech Synthesis System
  - TTS

Speaker Verification

- Speaker embedding: x-vector + + multi-head self-attention
- Probabilistic LDA (PLDA) classifier
- NIST SRE2016: trained with English, test on Cantonese: EER = 2-3%



Multi-lingual Multi-speaker TTS



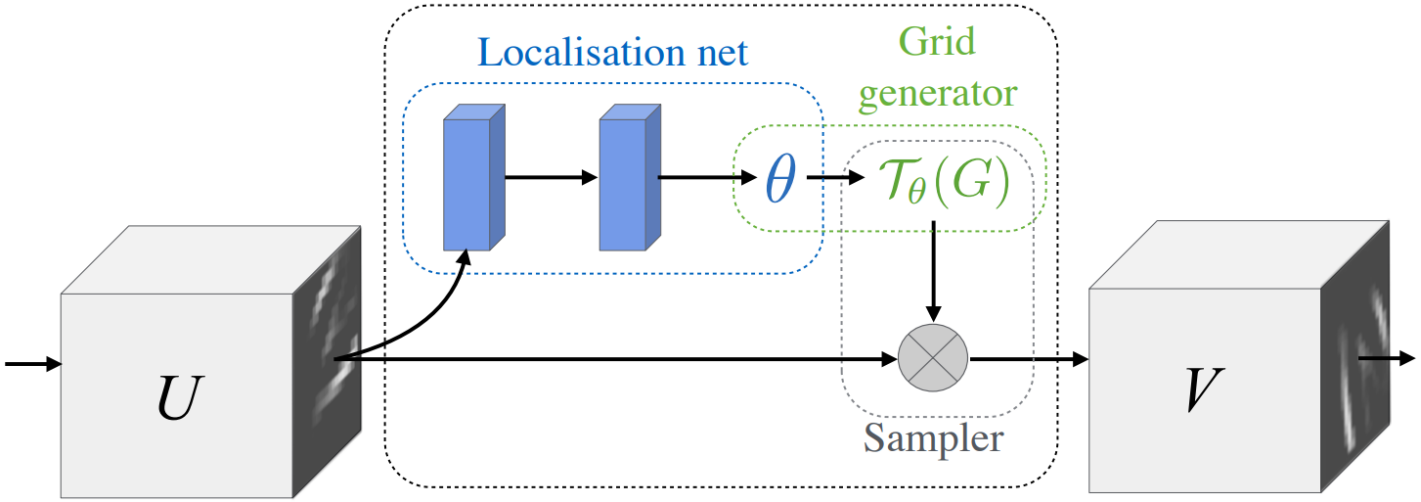
demo

Neural Voice Conversion

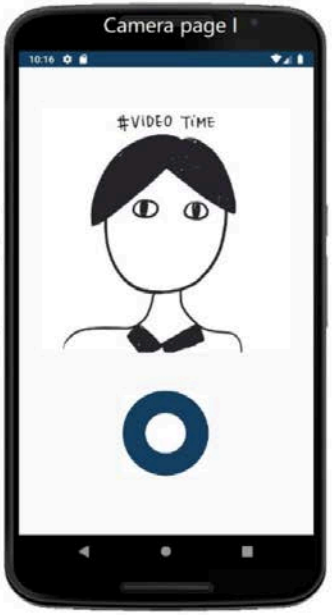
- No parallel data
- ~100 English speakers for training
- VC among the training speakers
- Make use of neural TTS
- demo

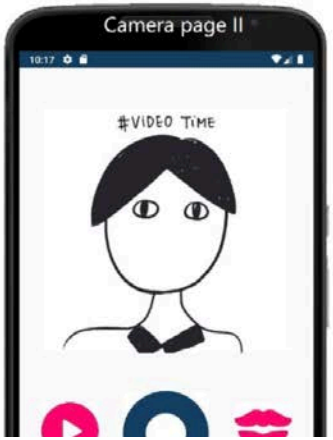
Lipreading

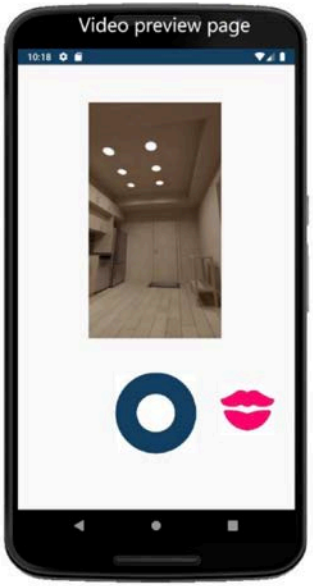
- Lip Reading in the Wild; 500 English words
- end-to-end with spatial transformer: 79.6%

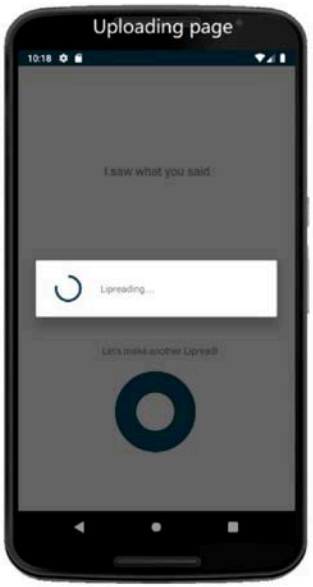


Cantonese Lipreading FYP









Multi-lingual Document Embeddings

- Ideally a document embedding is independent of its language.
- Need bilingual corpus between any 2 languages; otherwise, use NMT to get the translated texts.