Speech-to-Speech Translation

Lalaram Arya

Supervised by: Prof. S.R. Mahadeva Prasanna
Department of Electrical Engineering
Indian Institute of Technology (IIT) Dharwad

202021004@iitdh.ac.in

January 10, 2022

Table of contents

Speech to Speech Translation

- Speech translation is the process by which conversational spoken language are instantly translated in the target language.

Figure: General S2ST system

Need of Speech to speech translation system

- Many languages are there in the world.
- Not many people know multiple languages.
- S2S translation breaks down communication barriers among people who typically do not speak a common language.
- Enable the Spread of Ideas and Information
- Enables the Global Economy, Travel and Tourism, Health care, etc..

Traditional Speech to Speech Translation system

- Speech (acoustic signal) in – speech recognition (ASR)
- Speech (acoustic signal) out – speech sysnthesis (TTS)
- In between: text to text – Translation (MT)

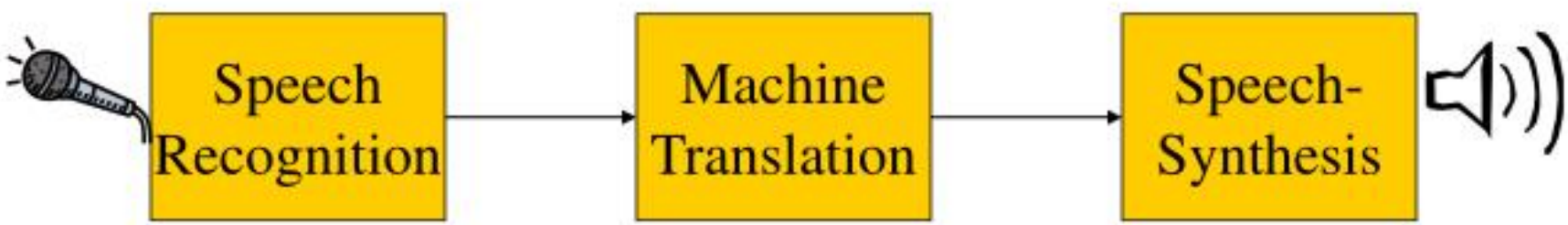


Figure: conventional S2ST system

Connecting the modules: ASR-MT

- What is propagated from recognizer to translator?
  - First best hypothesis, n-best list, entire lattice
  - Additional annotations: internal scores of recognizer, acoustic, language model, confidence scores
  - Prosodic information

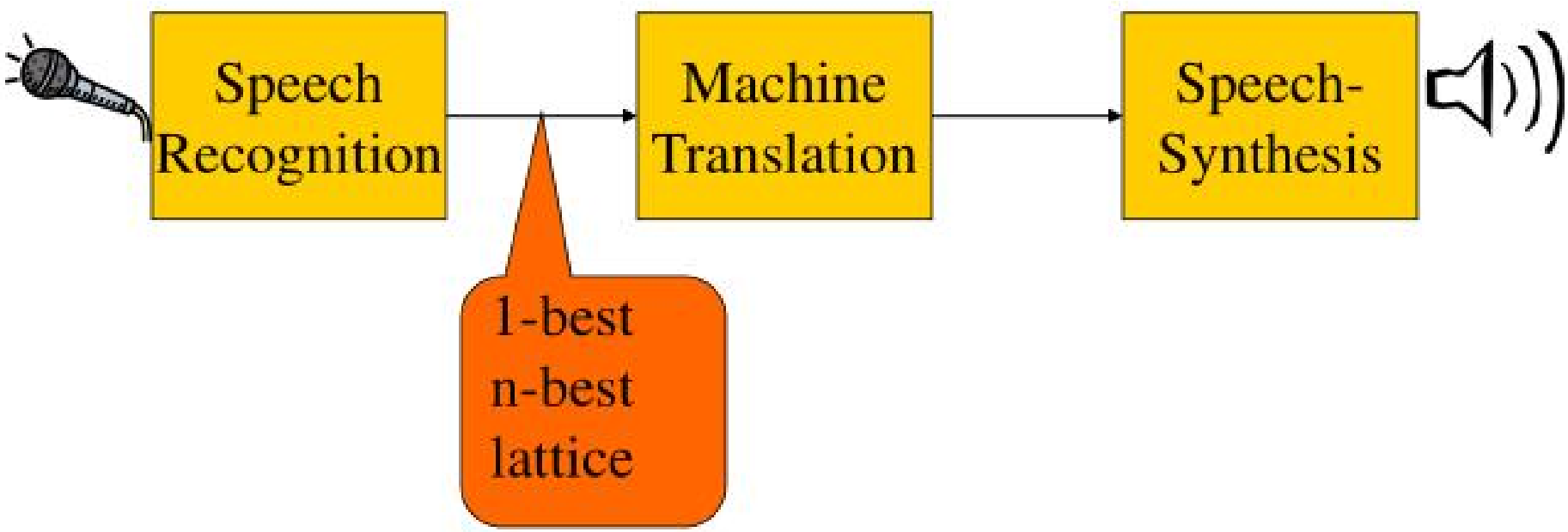


Figure: Information propagating from ASR to MT

Connecting the modules: MT-TTS

- What is propagated from translation to synthesis?
  - Sequence of words
  - How important is case information, punctuation?
  - Acoustic information: gender of speaker, f0, accents
  - Some of this information attached to specific words- alignment

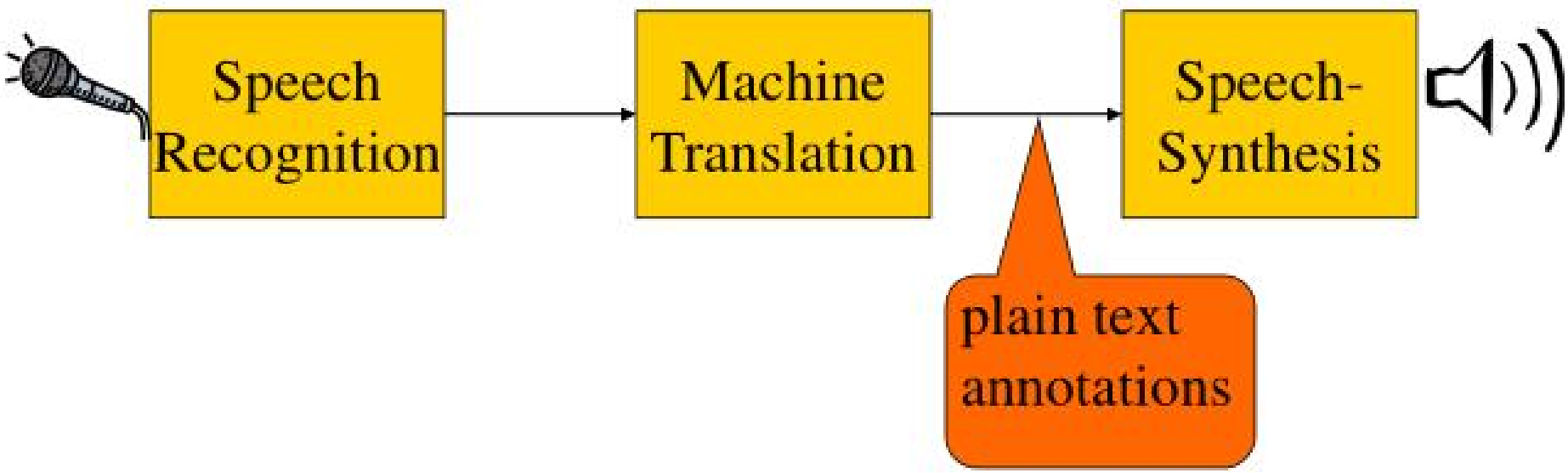


Figure: Information propagating from MT to TTS

GENERAL FLOW OF AN ASR

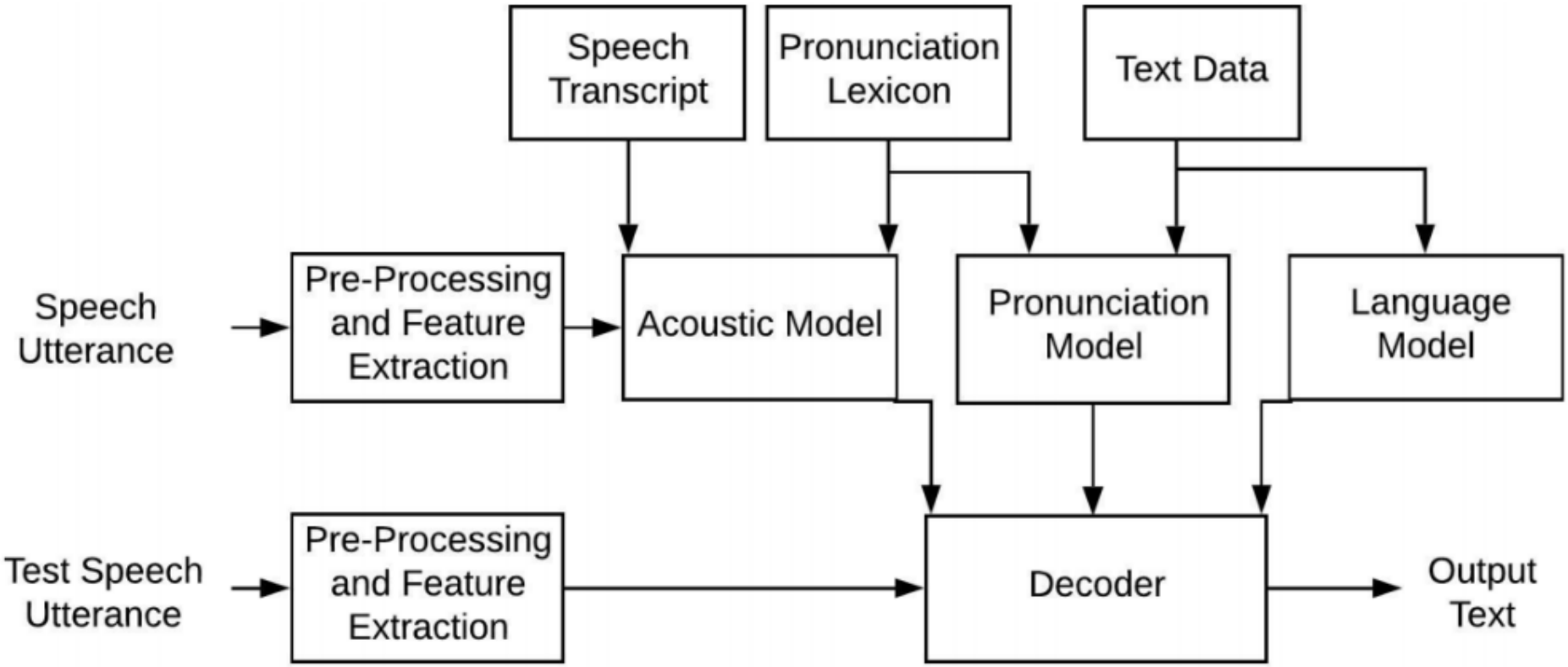


Figure: Flow diagram of ASR

Classification of ASR system

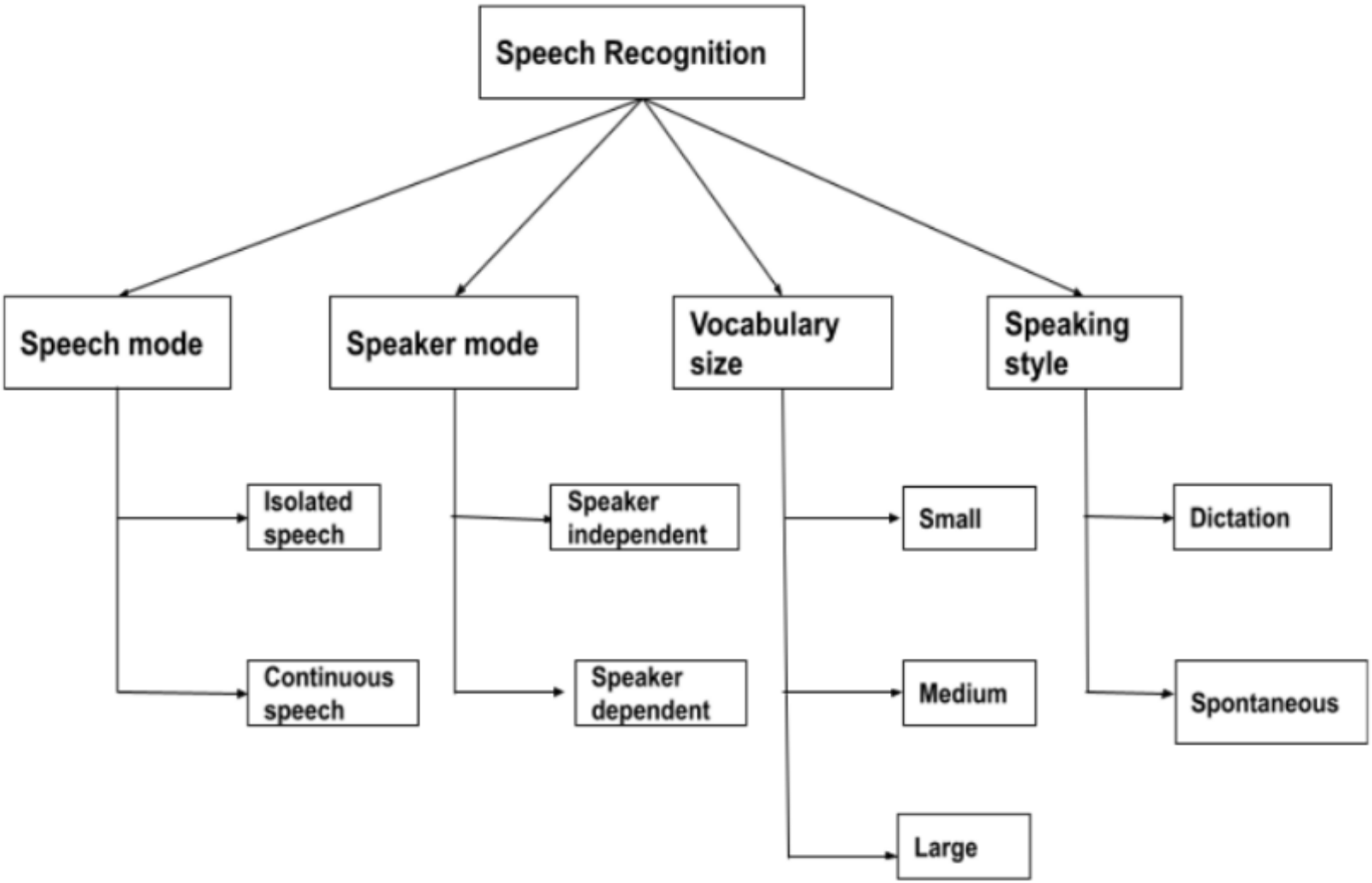


Figure: Classification of ASR system

Milestones in ASR

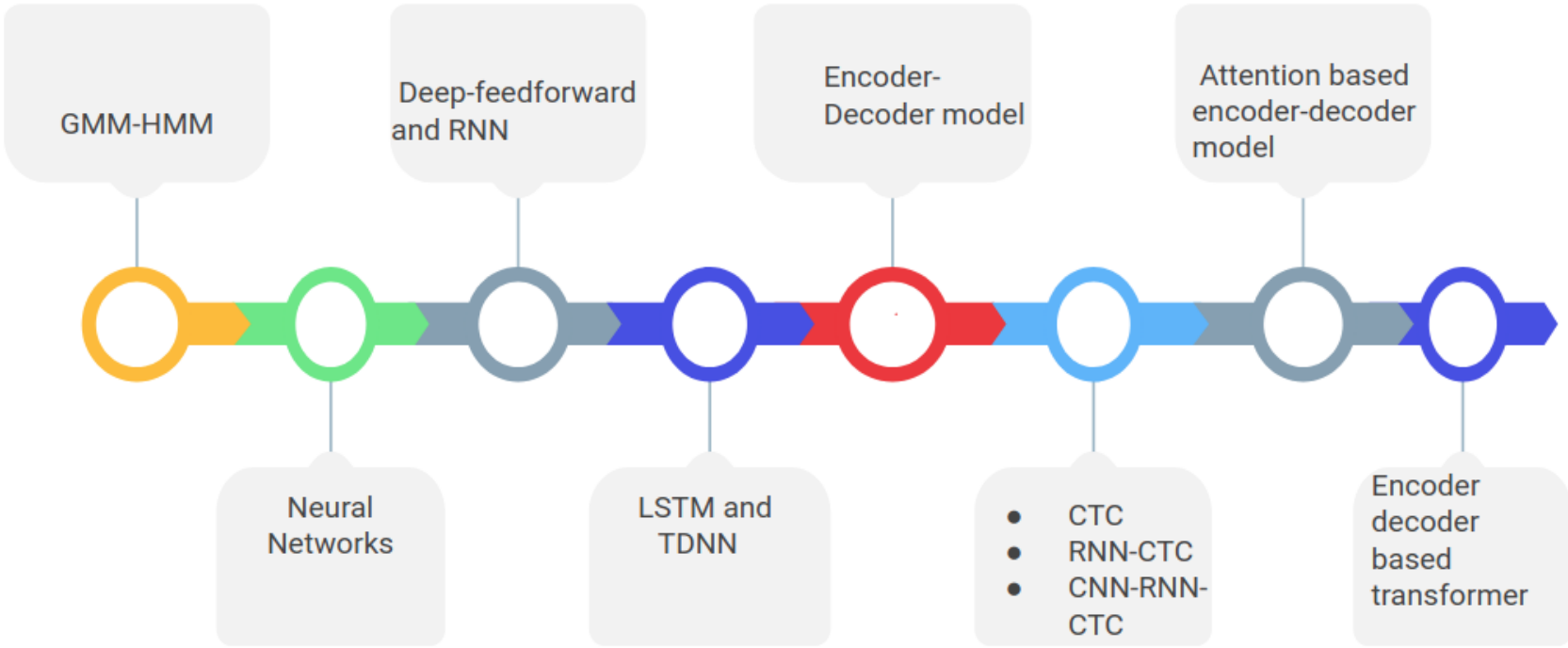


Figure: Milestones of ASR

Machine Translation

- The encoder is a series of LSTM layers where the input is fed so as to retain the structure of the sentence and summarizes the information in something called the context vector.
- The decoder is an LSTM whose initial states are initialized to the final states of the Encoder LSTM, i.e. the context vector of the encoder's final cell is input to the first cell of the decoder network.

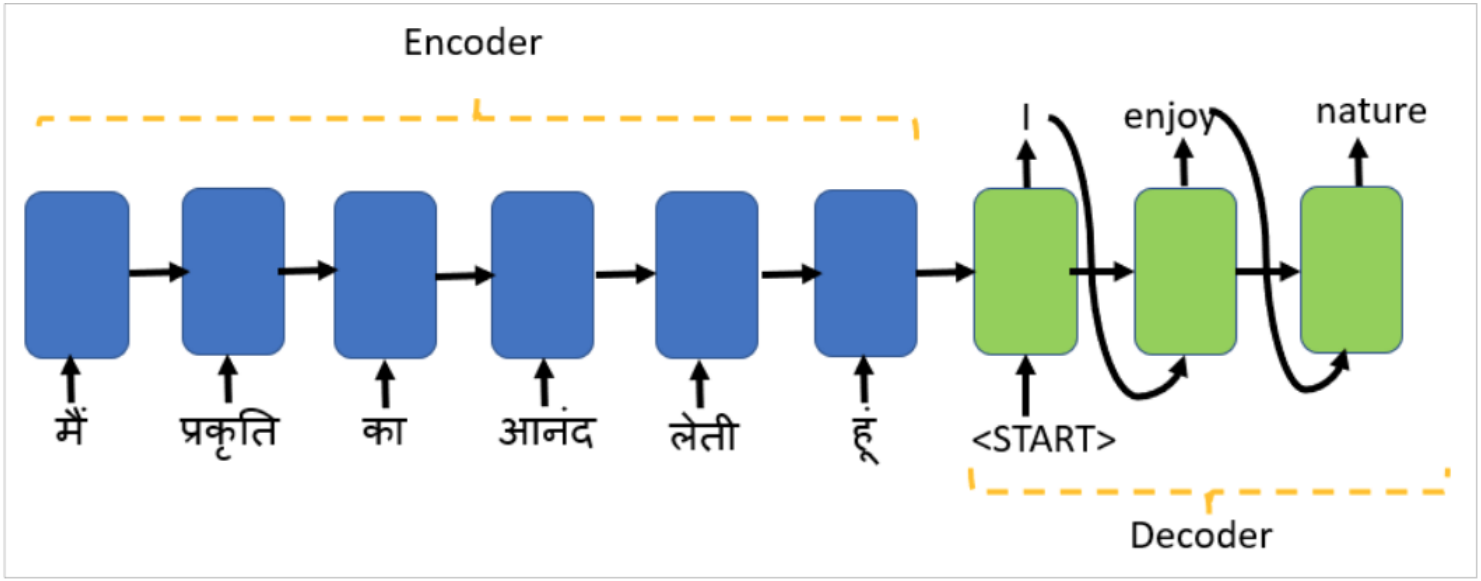


Figure: Seq-to-Seq model

Milestones in MT

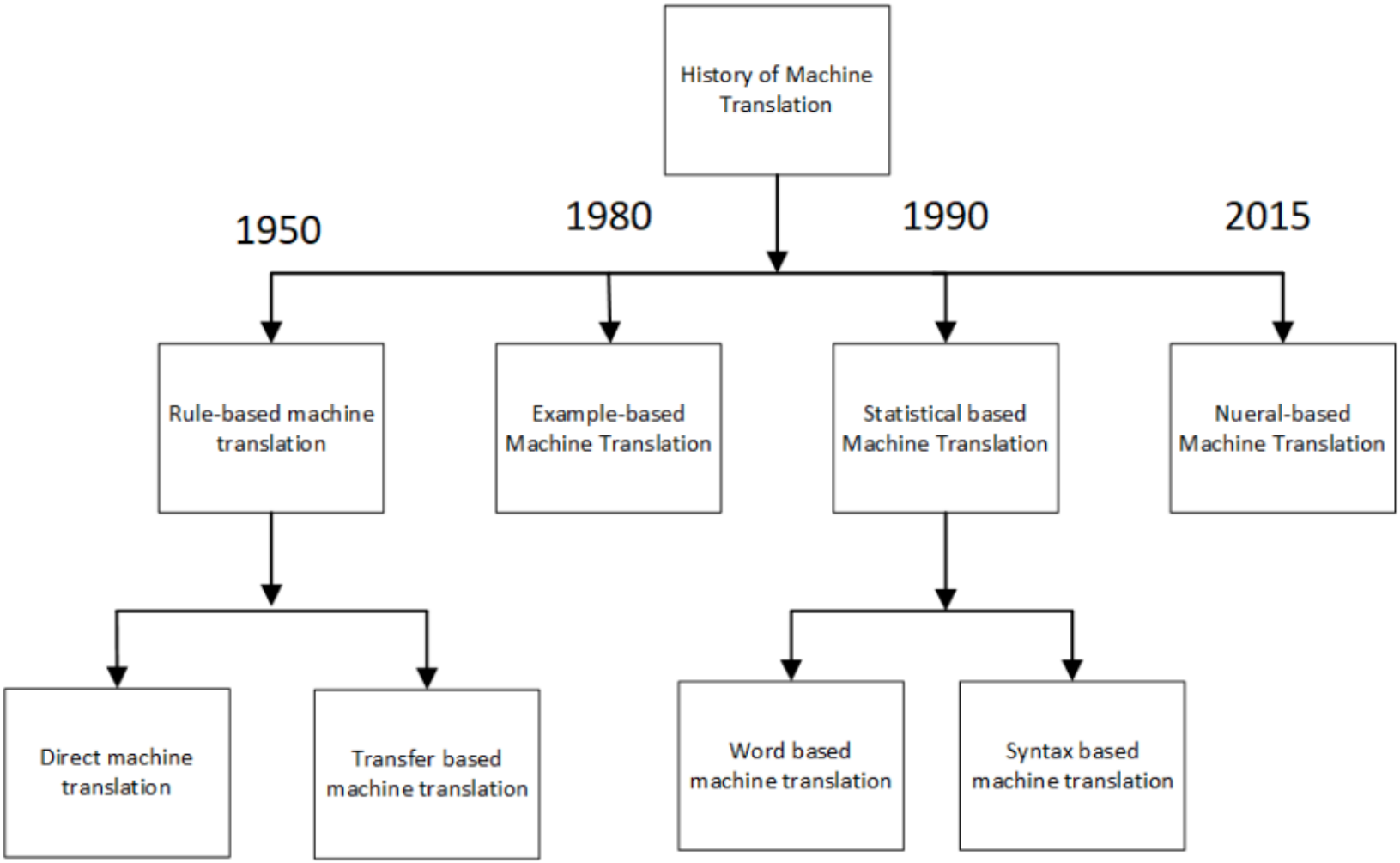


Figure: Different Models of MT grows with time

Current research trends in MT

- Cognate Detection in Machine Translation
- Undirected Sequence Models
- Parameter Sharing across Layers in Transformers

Text-to-speech

- The text-to-speech (TTS) synthesis procedure consists of two main phases.
  - Text analysis, where the input text is transformed into a phonetic or some other linguistic representation
  - Speech waveform, where the acoustic output is produced from this phonetic and prosodic information.

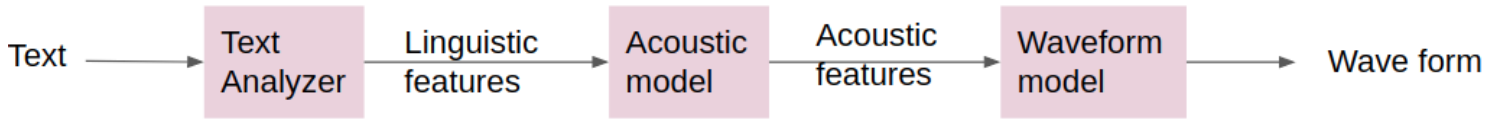


Figure: Flow diagram of ASR

Milestones in TTS

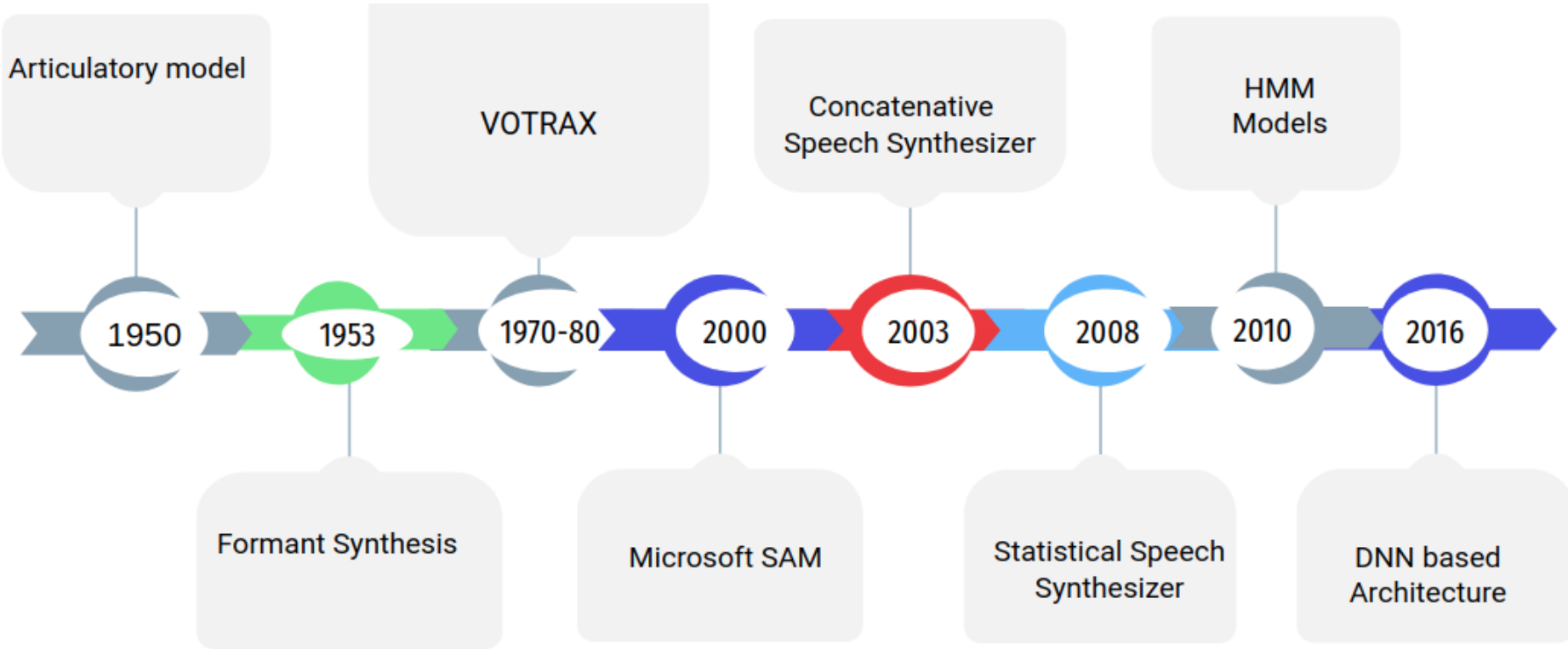


Figure: Different Models of TTS grows with time

Current research trends in TTS

- Front end text analysis,
- Supervised and semi supervised training of models
- The application of other speech related scenarios

Problems in spoken language translation

- Spoken language differs from written language
  - Vocabulary
  - Style
  - sentence structure
- How to deal with this
  - Specific translation models
  - Training on specific data-hardly available

Milestones in S2ST

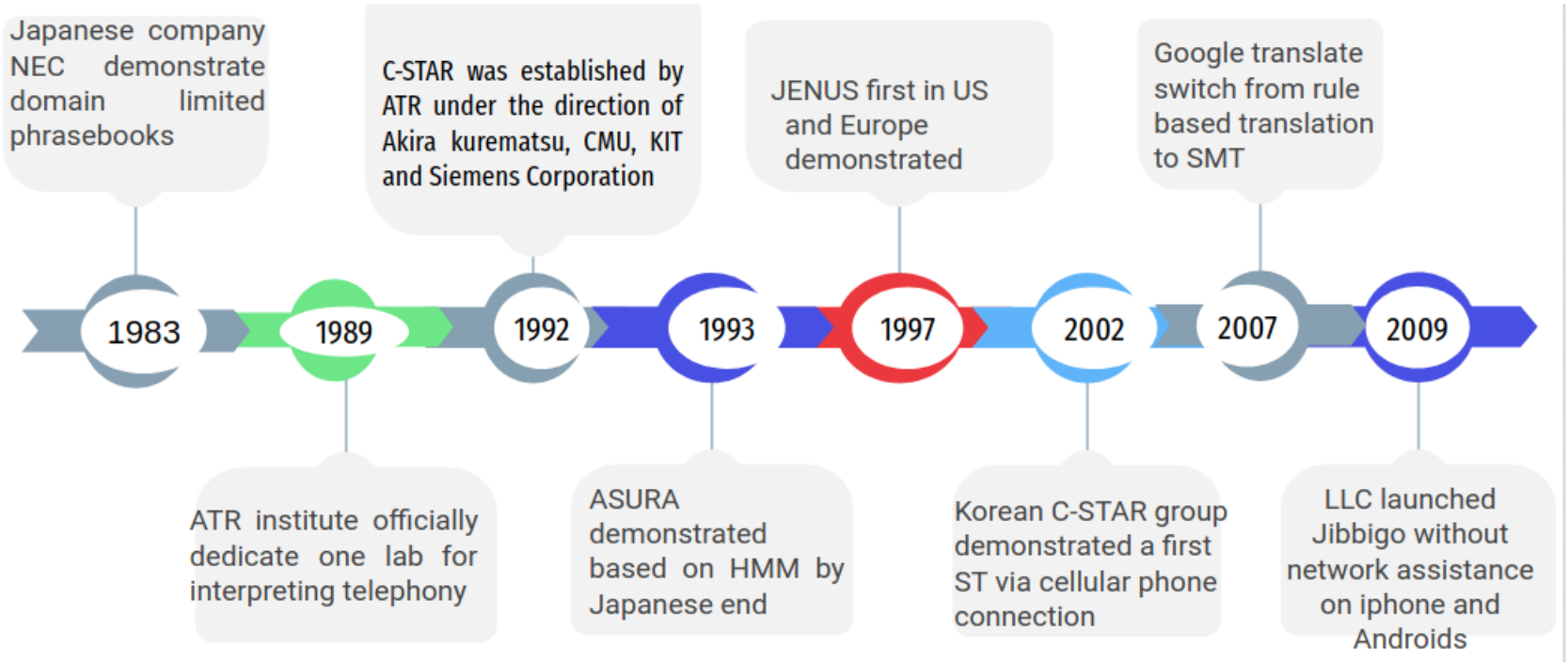


Figure: Different Models of TTS grows with time

Milestones in S2ST

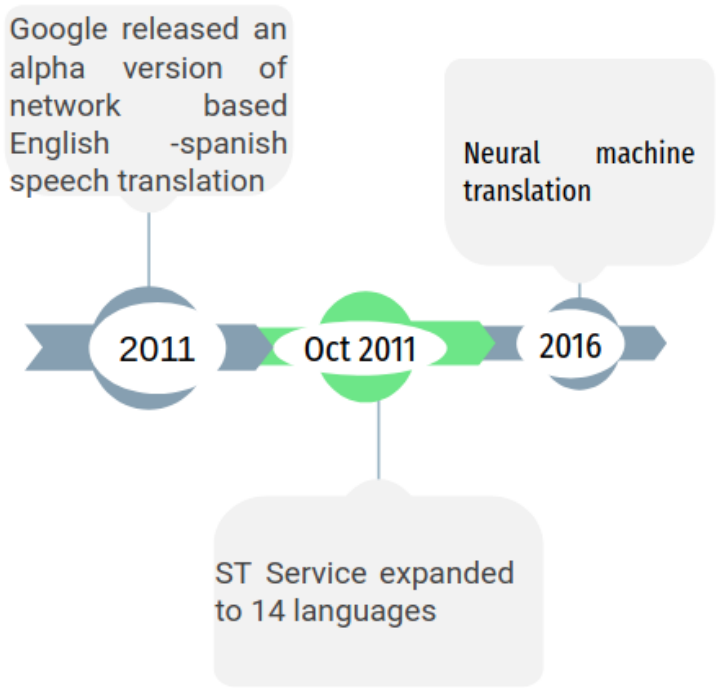


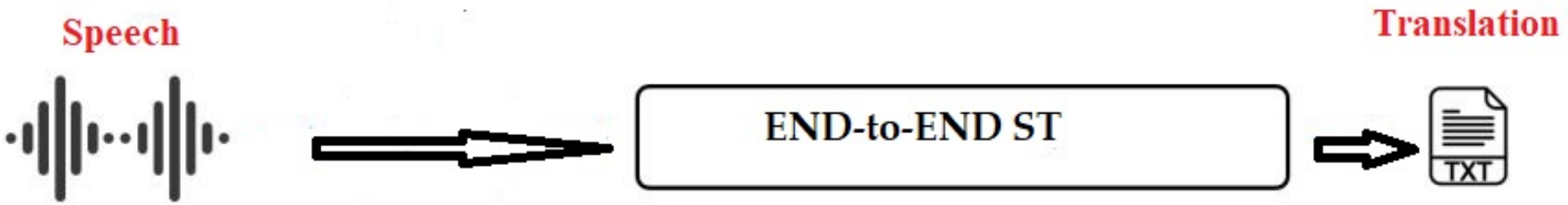
Figure: Different Models of TTS grows with time

Issues in Speech to Speech Translation system

- Cascaded systems have the potential problem of errors compounding between components, e.g. recognition errors leading to larger translation errors.
- Deep Learning models are data hungry so to train the model audio parallel corpus are hardly available.
- No specific evaluation measure for speech to speech translation.

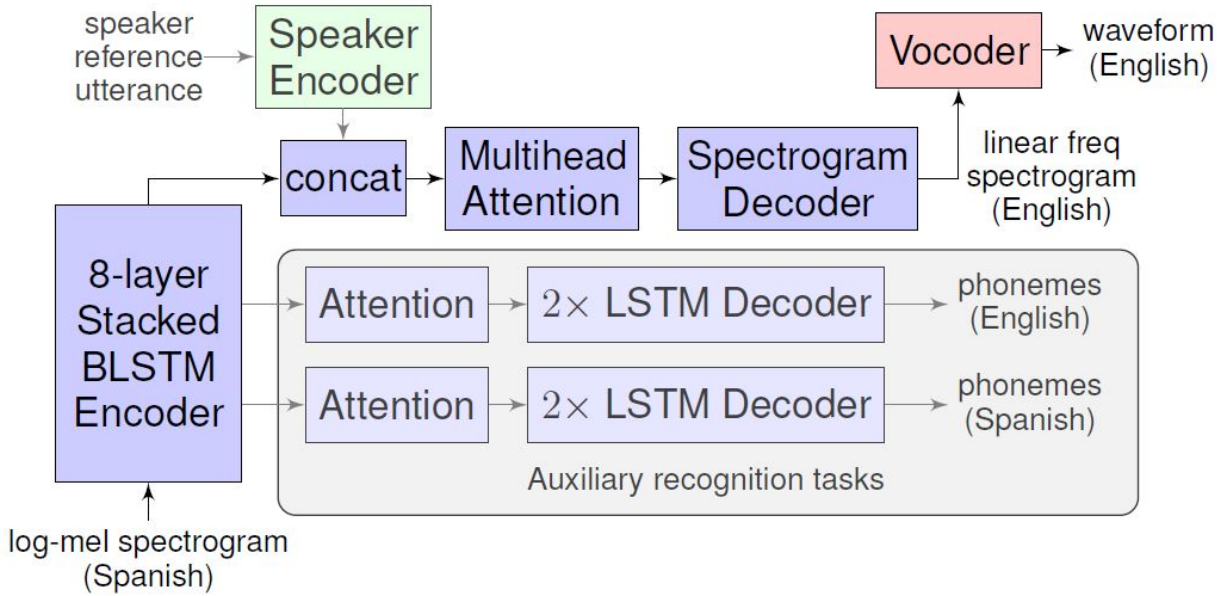
Development of Current Trends

- **END-to-END speech translation**



Direct speech to speech translation

- The network is trained end-to-end, learning to map speech spectrograms into target spectrograms in another language



Ye Jia et al. 2019 Direct speech-to-speech translation with a sequence-to-sequence model.

Thank You