# End-to-End Machine Learning for Speech Processing: Speech-to-Speech, Speech-to-Text and Text-to-Speech


Gérard Chollet (CNRS-SAMOVAR)

and Colleagues from IV, Zaion, CNRS/IMT and SMI

gerard.chollet@telecom-sudparis.eu


Intelligent Voice®

zaion TIME TO TALK

Une école Institut Mines-Télécom

TELECOM SudParis

cnrs

Depuis 80 ans, nos connaissances bâtissent de nouveaux mondes

speechmorphing

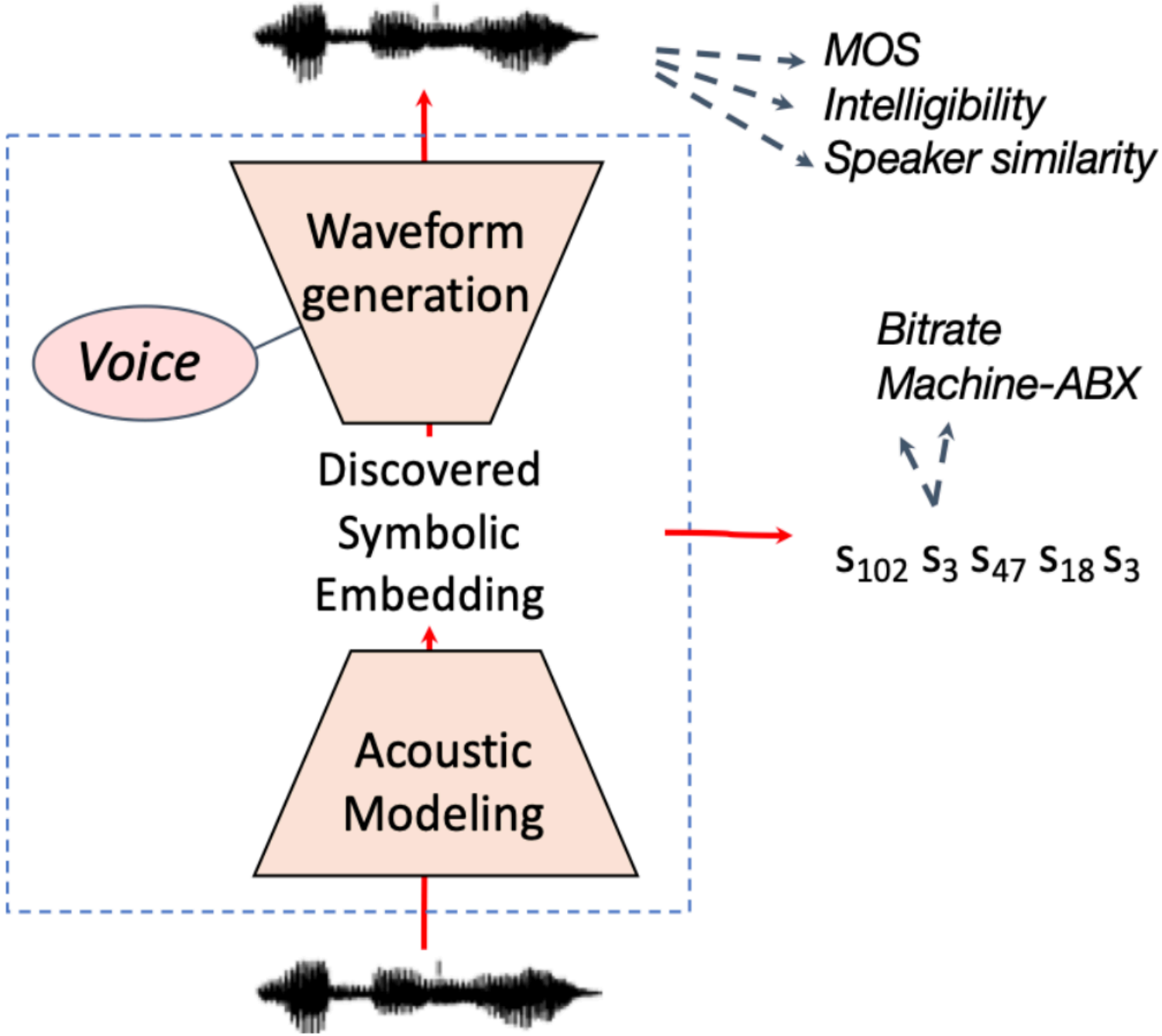
# What is End-to-End Machine Learning?

Deep Neural Networks (DNN, CNN, RNN,…) are designed (hand crafted/AutoML) and trained end-to-end to optimize a desired input-output association.

Large data sets (natural/augmented) are necessary to cover all possible (unknown) variations.

An example:

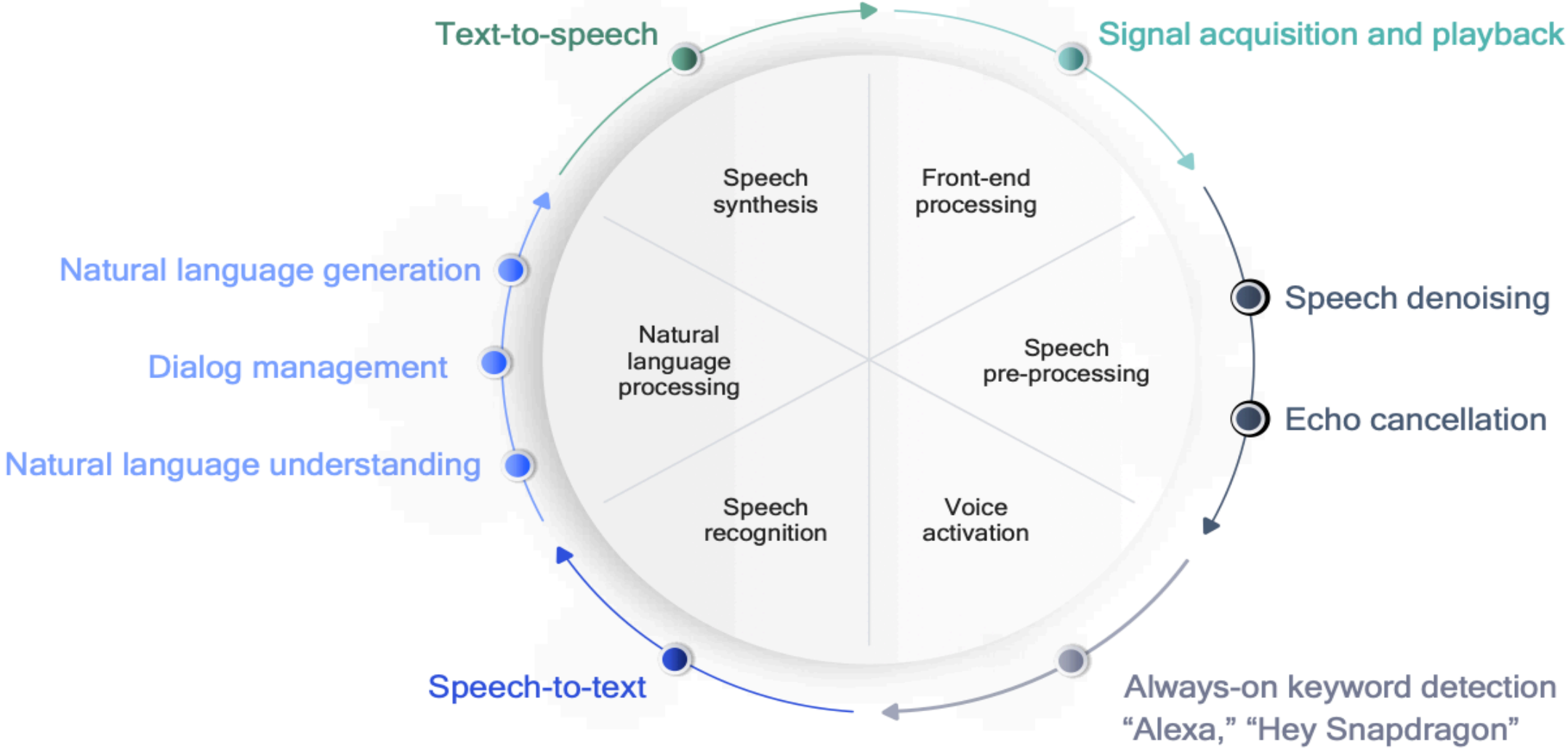
# ZeroSpeech 2019: TTS without T




https://zerospeech.com/2019/

# Voice UI components required for an end-to-end solution

## Machine speech chain: listener and speaker



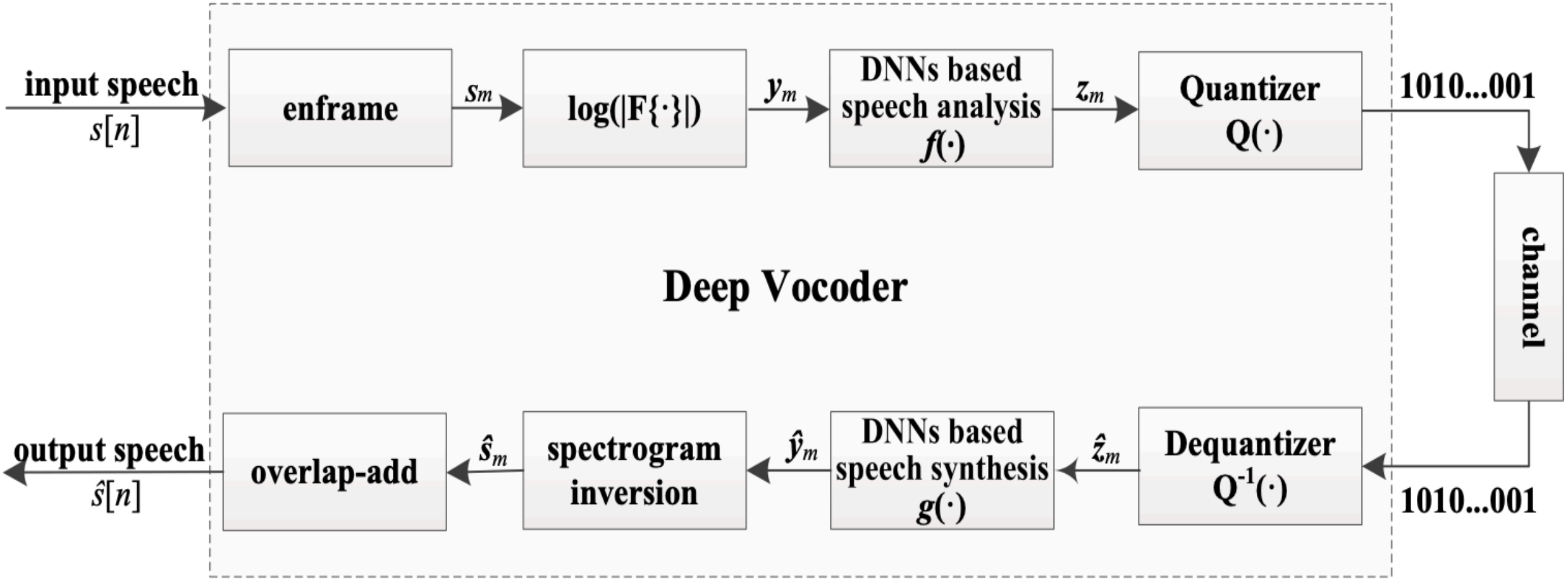

Qualcomm Snapdragon is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.

https://www.qualcomm.com/media/documents/files/making-an-on-device-personal-assistant-a-reality.pdf

# Speech-to-Speech processing

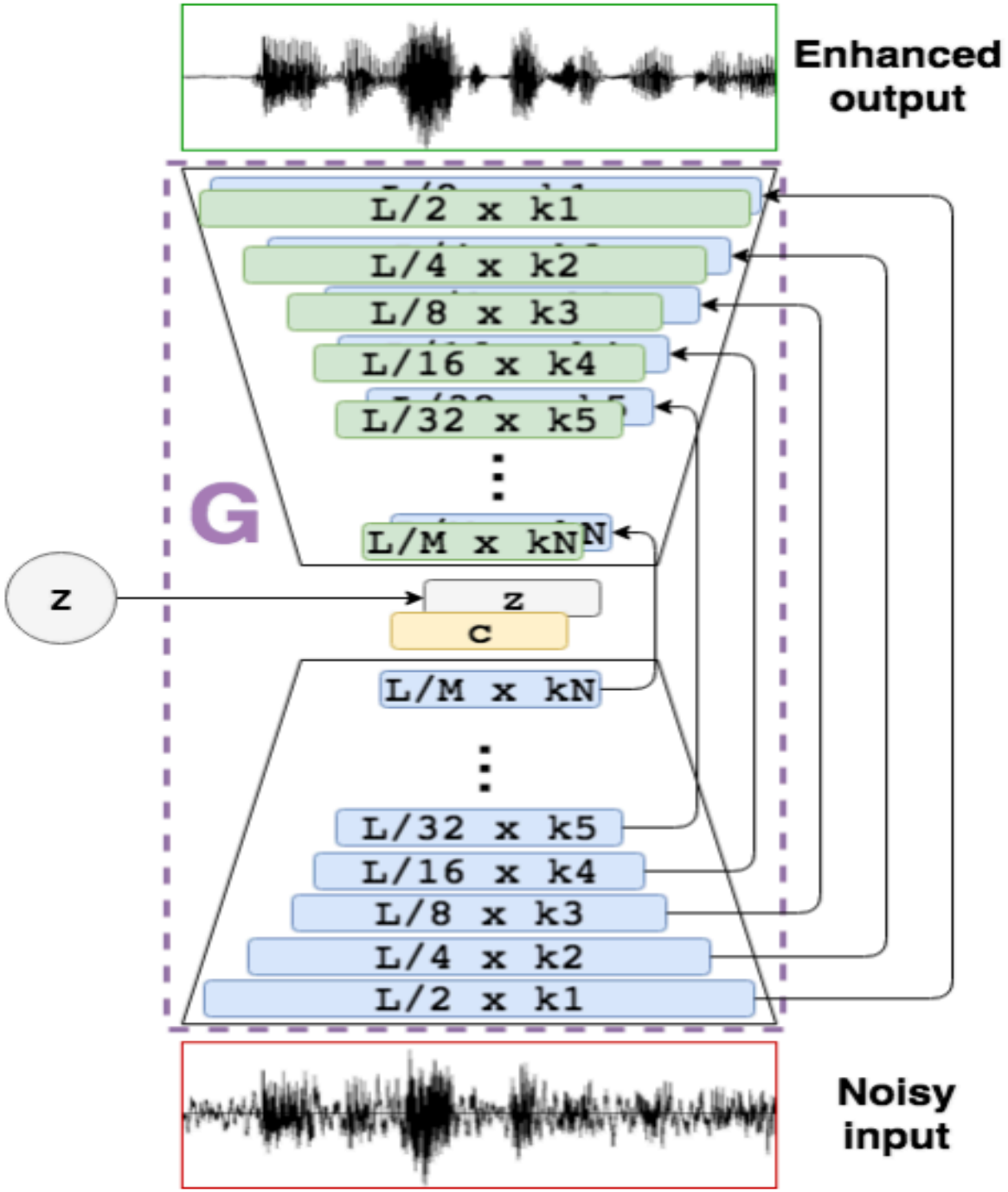
- Speech coding, compression,
- Speech enhancement, denoising,
- Voice conversion,
- Speech-to-Speech Translation,
- Spoken Dialogue System,
- …

# DEEP VOCODER: LOW BIT RATE COMPRESSION OF SPEECH WITH DEEP AUTOENCODER




https://arxiv.org/pdf/1905.04709.pdf

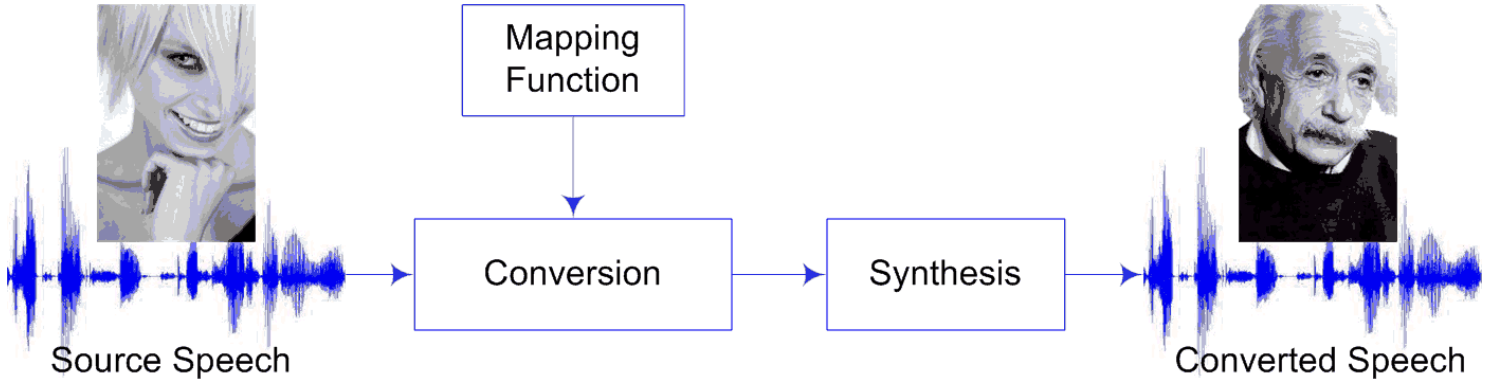
# SEGAN: Speech Enhancement Generative Adversarial Network



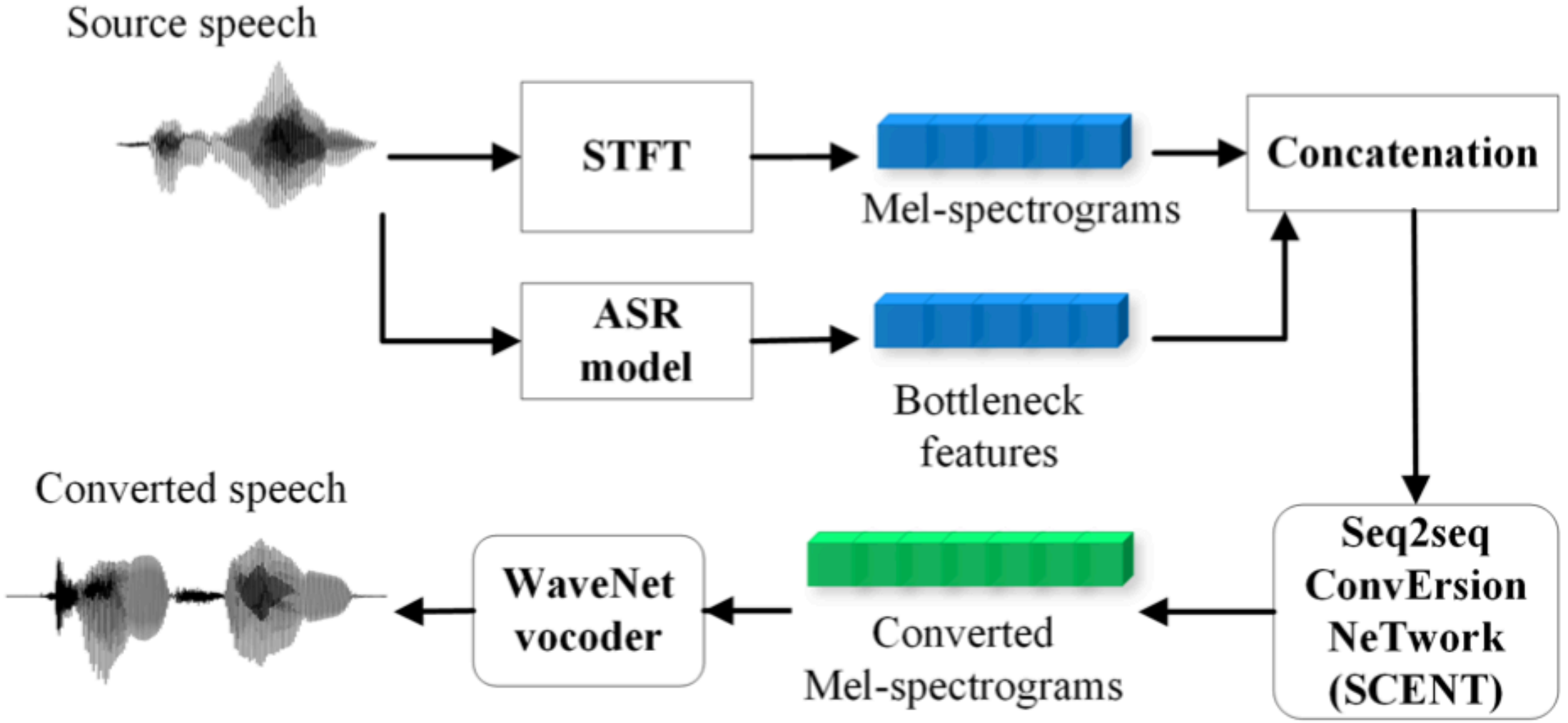

https://arxiv.org/pdf/1703.09452v1.pdf

# Speaker Transformation (3/3)

- The estimated transformation rule is applied to an original speech pronounced by the source speaker
- The new utterance sounds like the same speech pronounced by the target speaker
- The last step is the re-synthesis of the signal to reconstruct the “forged” source speech voice

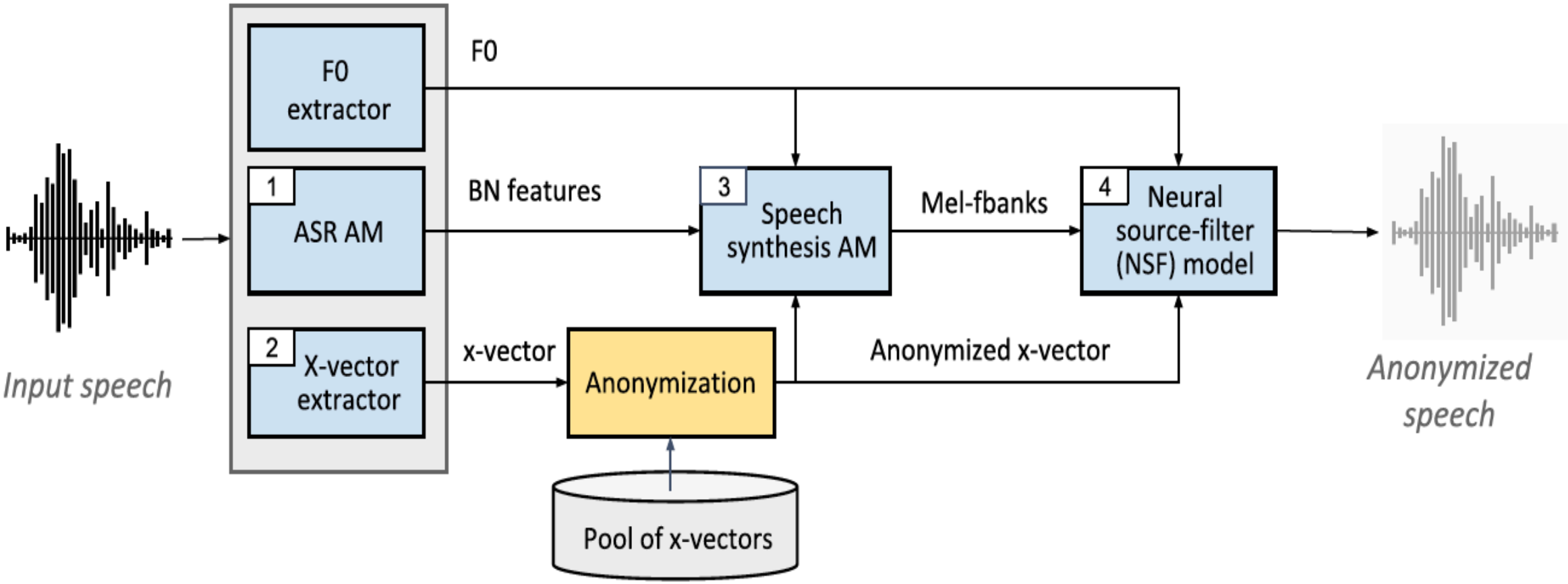


# Sequence-to-Sequence Acoustic Modeling for Voice Conversion



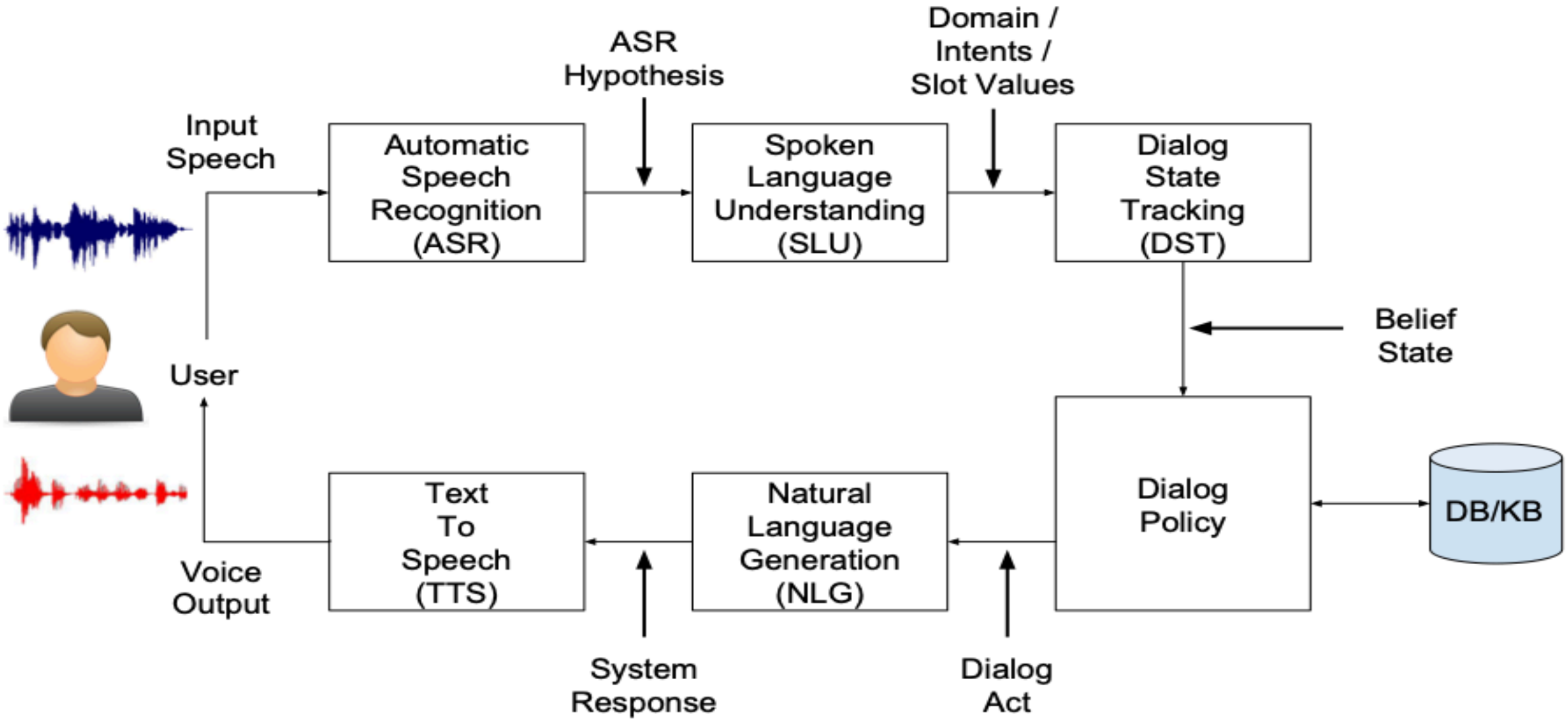

https://arxiv.org/pdf/1810.06865.pdf

# The VoicePrivacy 2020 Challenge



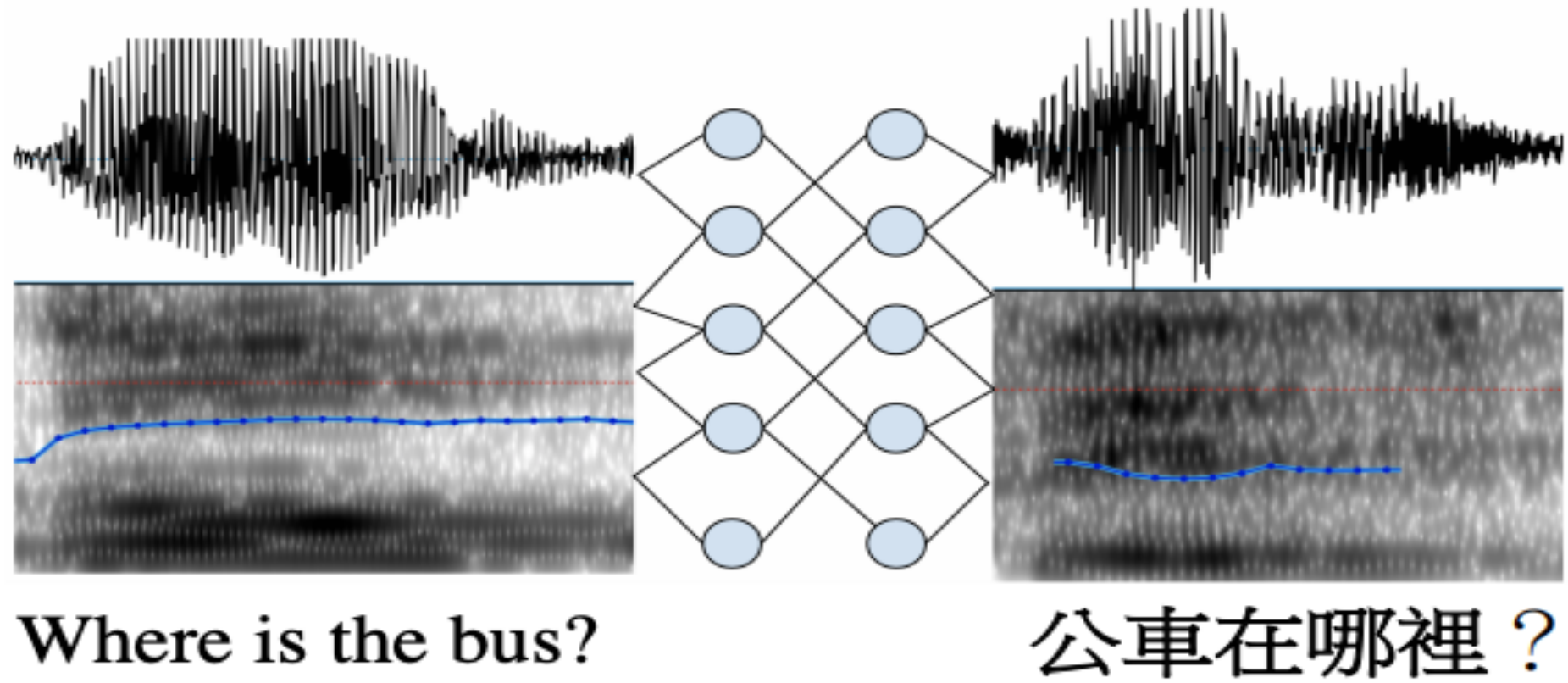

https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020

# End-to-End Learning of Task-Oriented Dialogs




https://www.aclweb.org/anthology/N18-4010.pdf

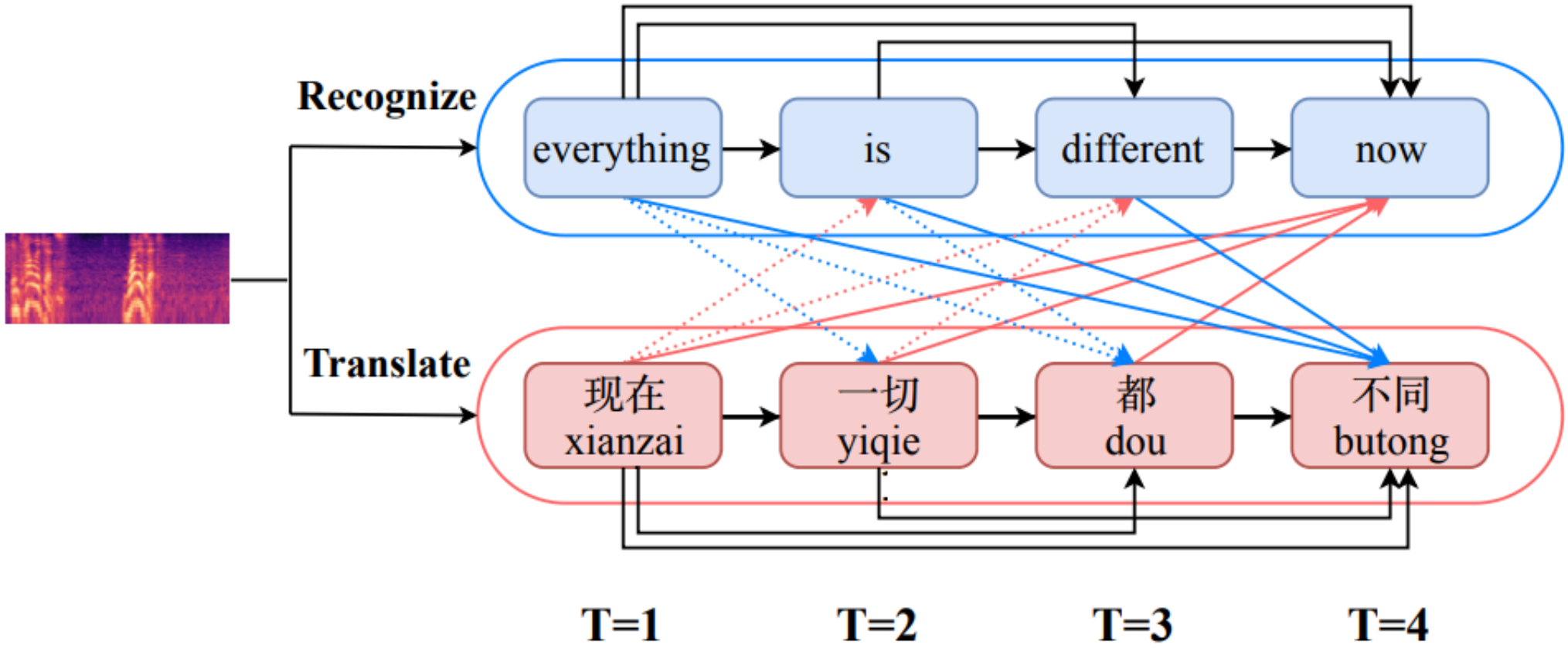
# Speech-To-Speech Translation




http://web.stanford.edu/class/cs224s/reports/Michelle_Guo.pdf

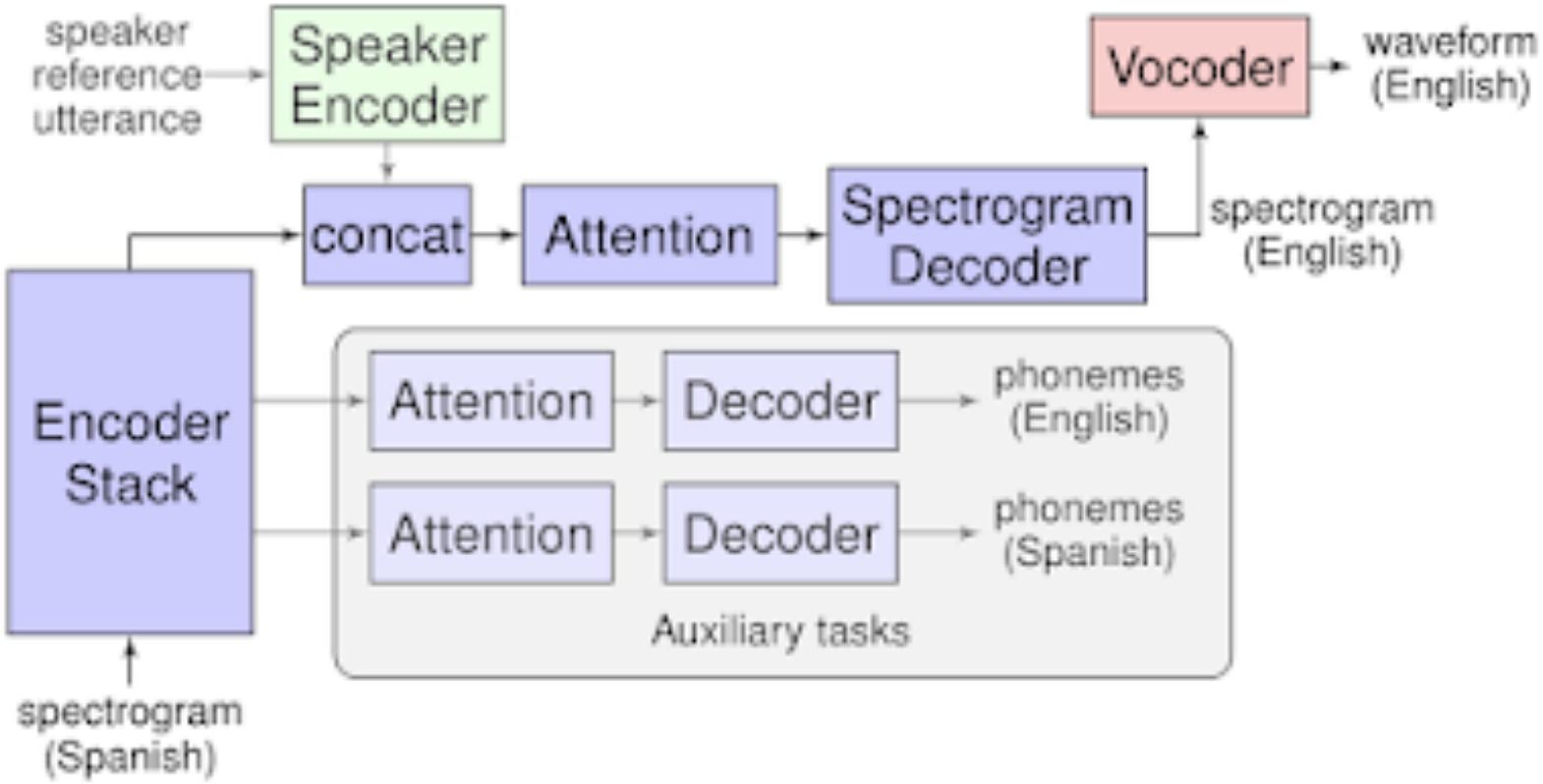
# Synchronous Speech Recognition and Speech-to-Text Translation with Interactive Decoding

https://arxiv.org/pdf/1912.07240.pdf



# Translatotron: An End-to-End Speech-to-Speech Translation Model

https://ai.googleblog.com/2019/05/introducing-translatotron-end-to-end.html



# Speech-to-Text/Symbols

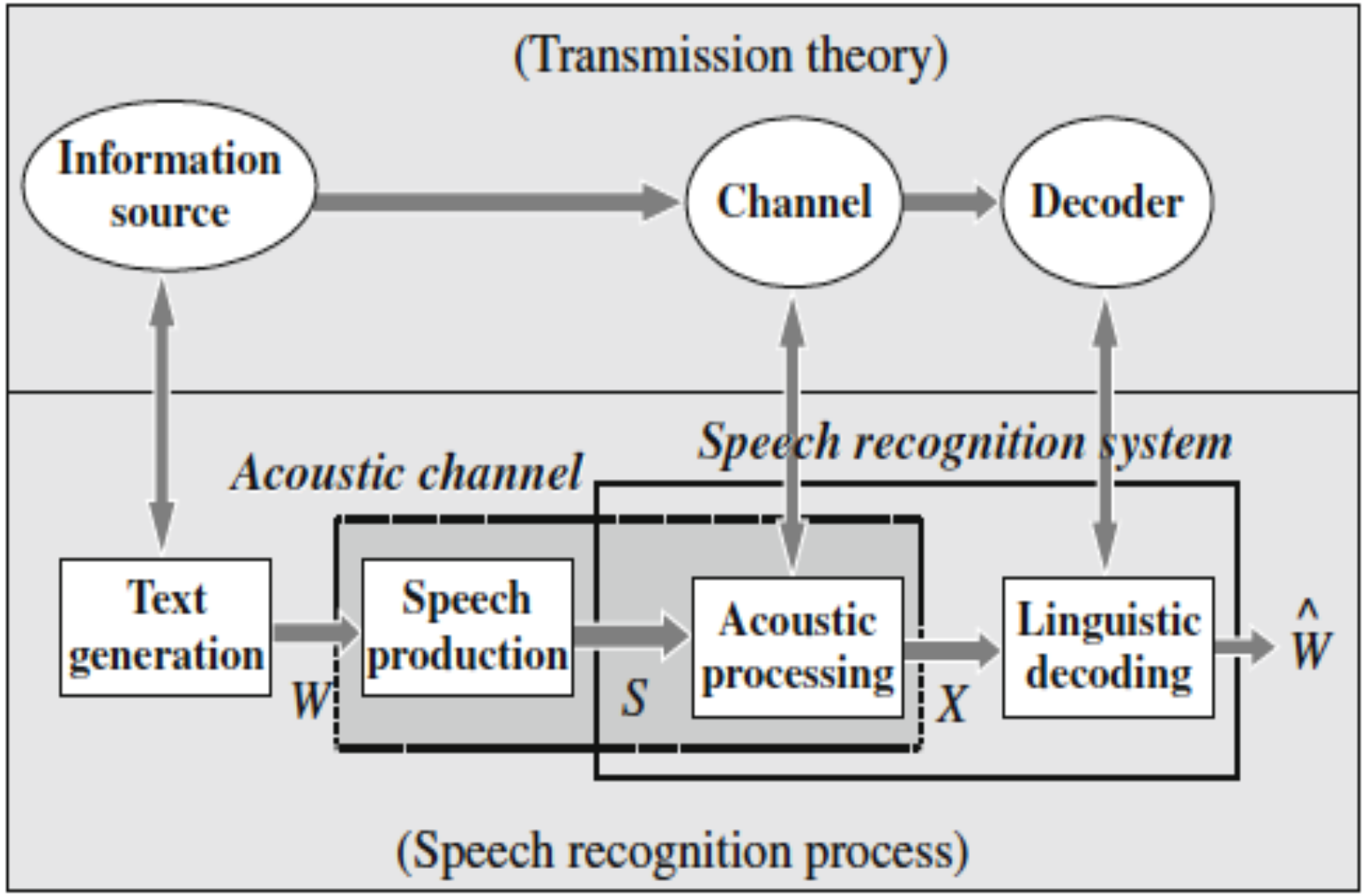
- Automatic Speech Recognition, Wav2Letters, phone alignment,
- Embeddings (acoustic, word, semantic,…)
- Key word spotting,
- Paralinguistic Pattern Recognition:
  - Gender recognition,
  - Age estimation,
  - Emotional state detection,
  - …
- Speaker identification/verification/diarization,
- Language, dialect, accent recognition,
- …

Fang Chen
Kristiina Jokinen
*Editors*

# Speech Technology

Theory and Applications

Springer




$$\hat{W} = \arg\max_{W} P(W|X) = \arg\max_{W} \frac{P(X|W)\,P(W)}{P(X)}$$

# Towards End-to-End Speech Recognition with Recurrent Neural Networks

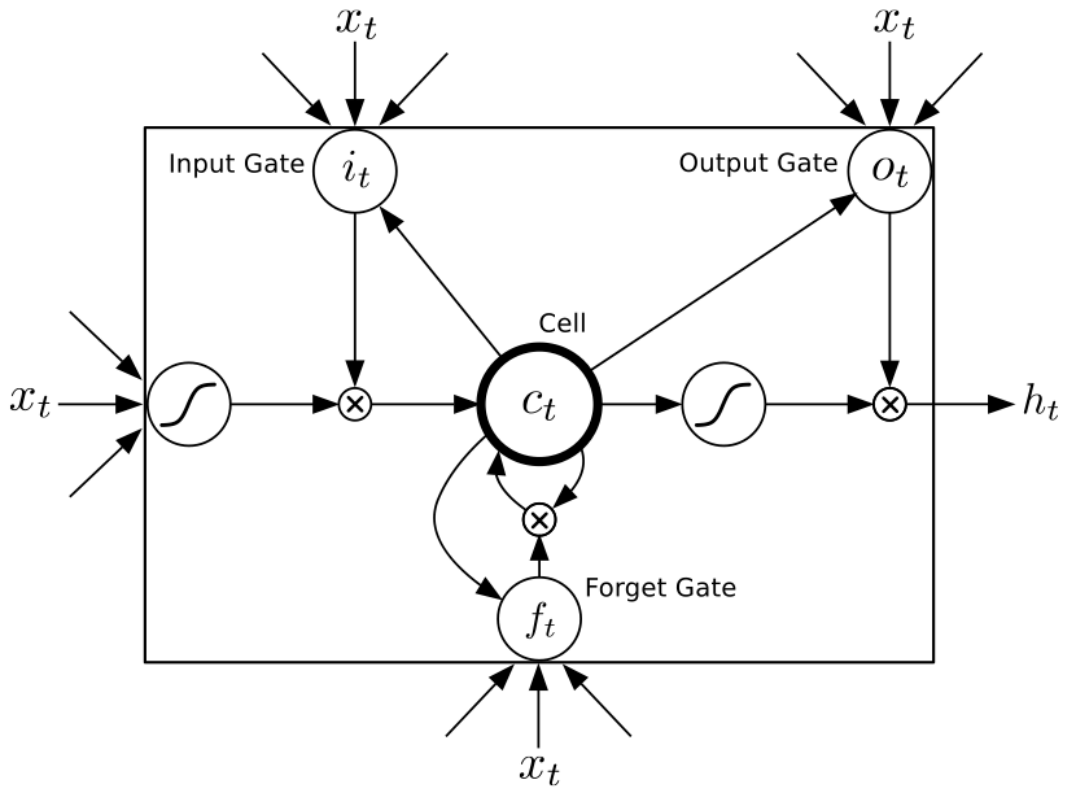



*Figure 1.* **Long Short-term Memory Cell.**

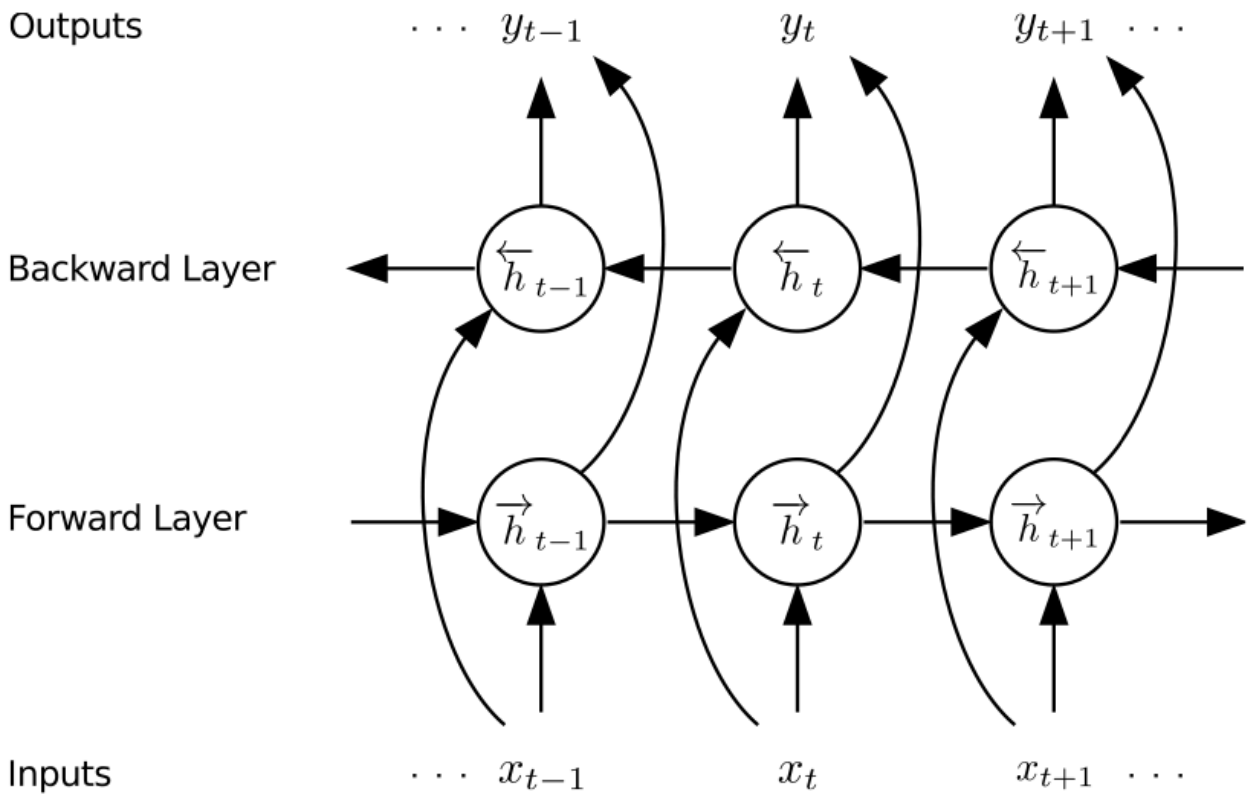
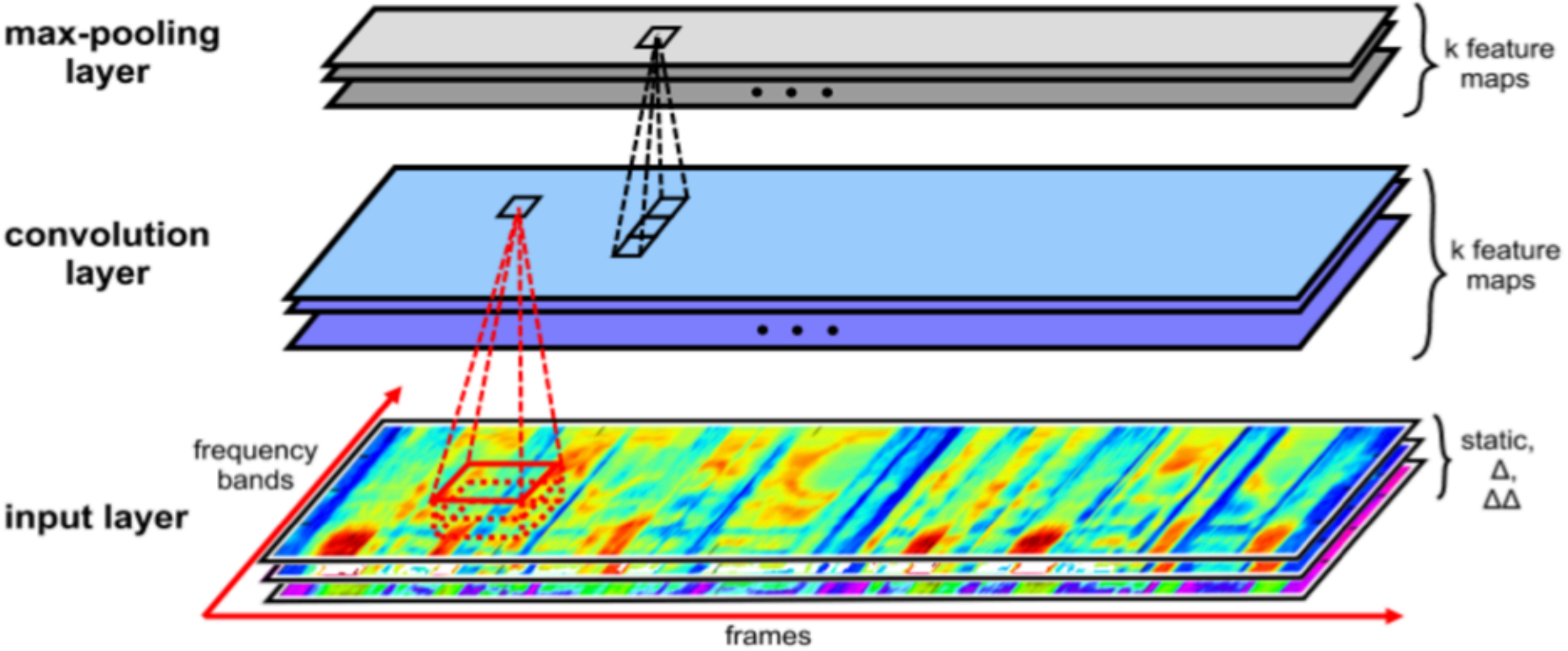



*Figure 2.* **Bidirectional Recurrent Neural Network.**

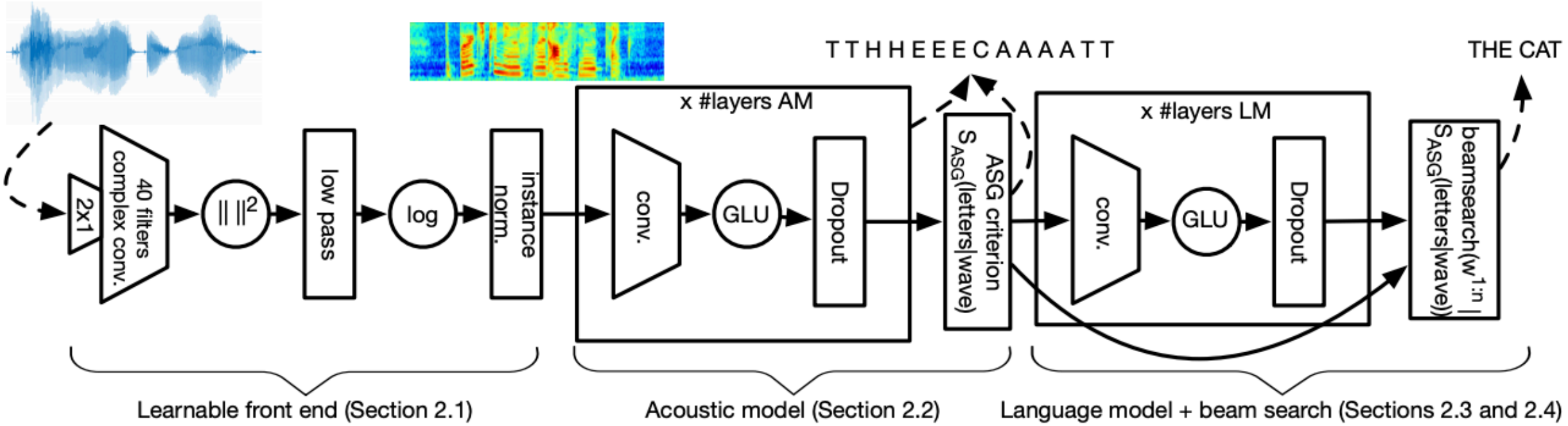
http://proceedings.mlr.press/v32/graves14.pdf

# Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks



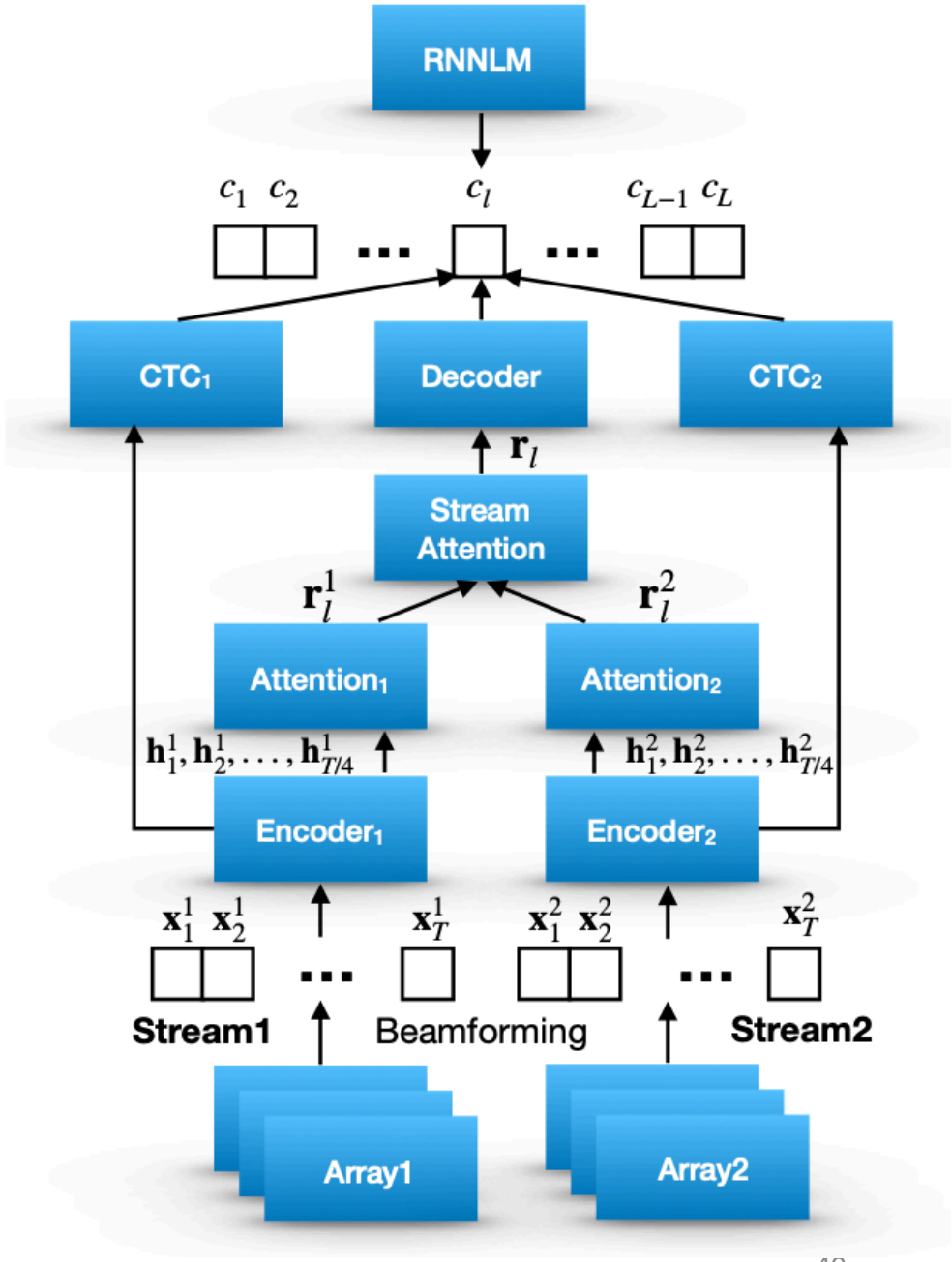

https://arxiv.org/pdf/1701.02720.pdf

# Facebook Fully Convolutional Speech Recognition




https://arxiv.org/pdf/1812.06864.pdf

# STREAM ATTENTION-BASED MULTI-ARRAY END-TO-END SPEECH RECOGNITION

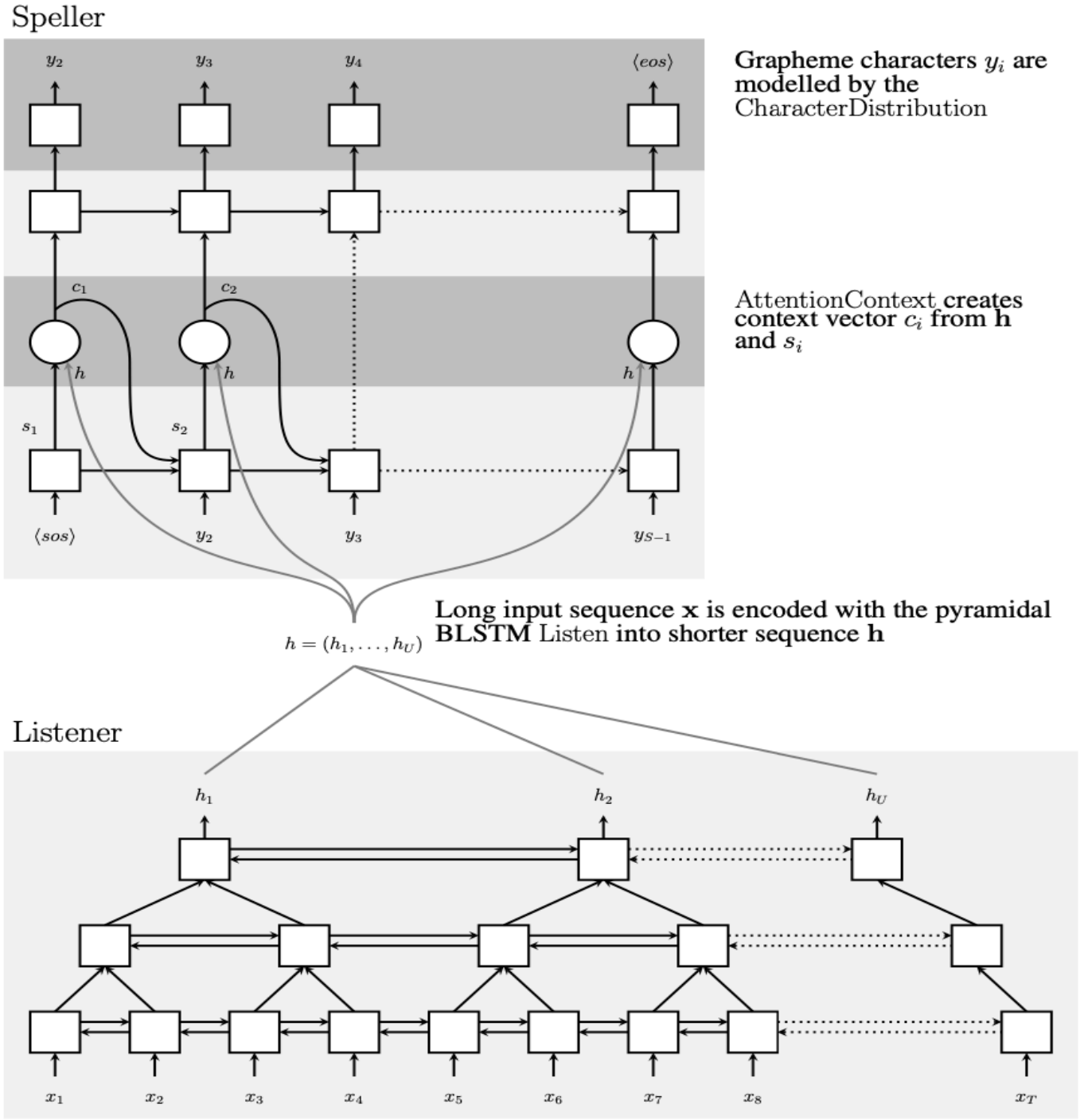



http://150.162.46.34:8080/icassp2019/ICASSP2019/pdfs/0007105.pdf

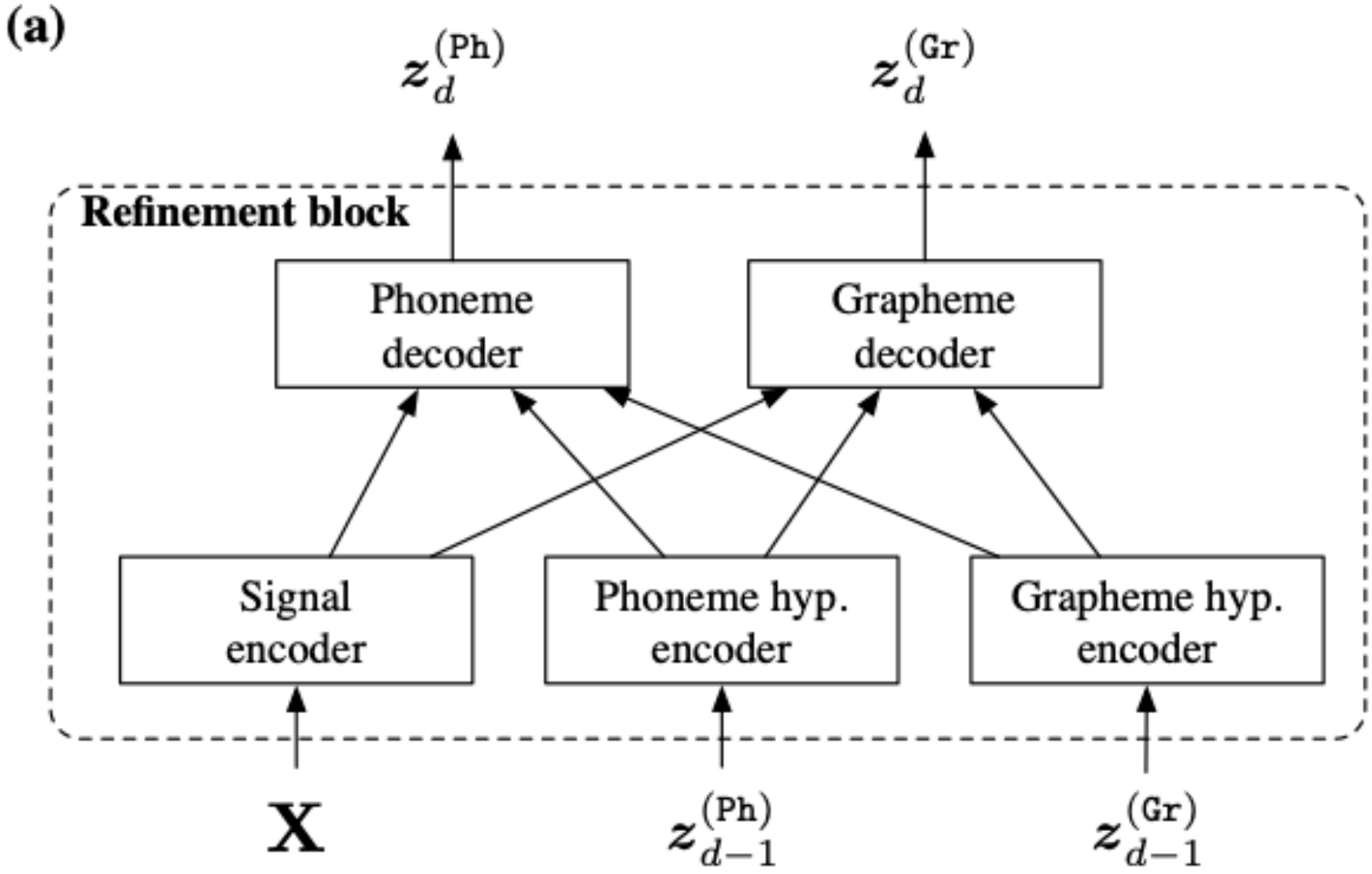
# Listen, Attend and Spell

Google Brain

https://arxiv.org/pdf/1508.01211.pdf



# JOINT PHONEME-GRAPHEME MODEL FOR END-TO-END SPEECH RECOGNITION



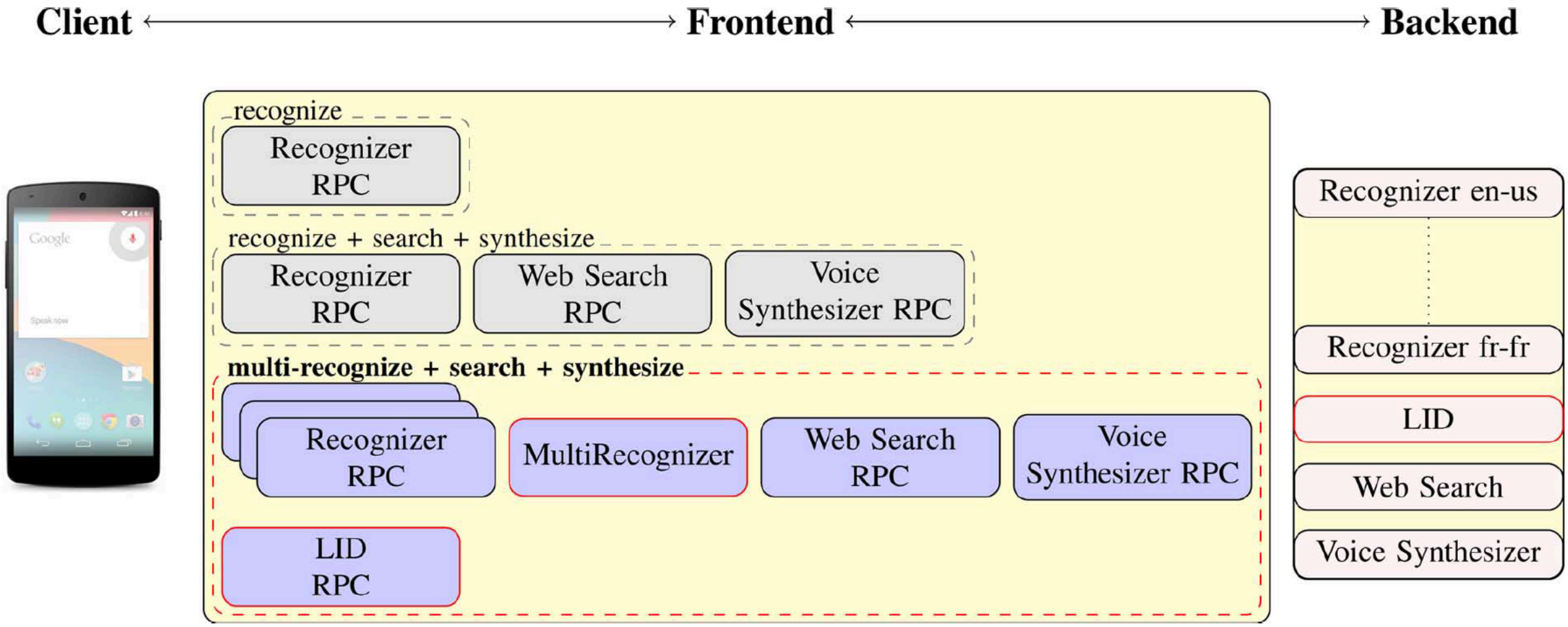

https://storage.googleapis.com/pub-tools-public-publication-data/pdf/45885a679869bf40528dcf46cfd795a3003eeb5e.pdf

# End-to-End Multilingual Multi-Speaker Speech Recognition



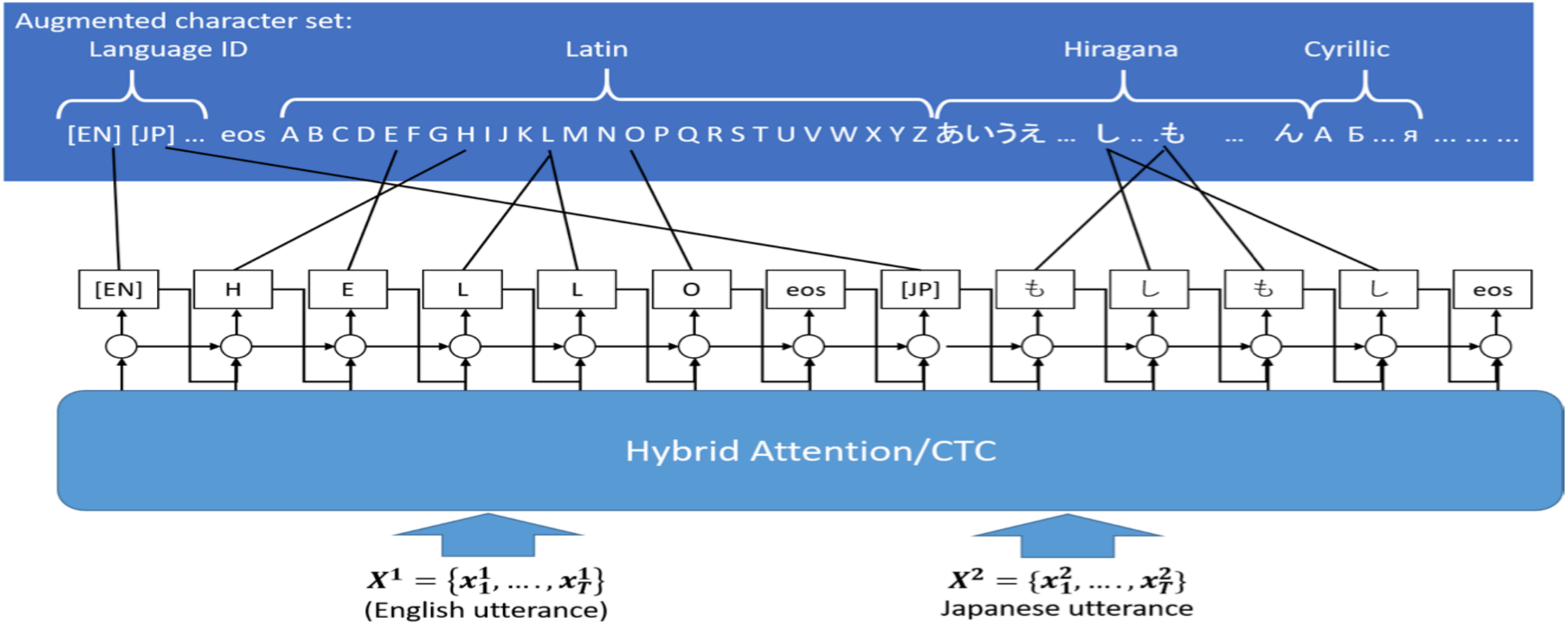
https://www.merl.com/publications/docs/TR2019-101.pdf

# A Real-Time End-to-End Multilingual Speech Recognition Architecture



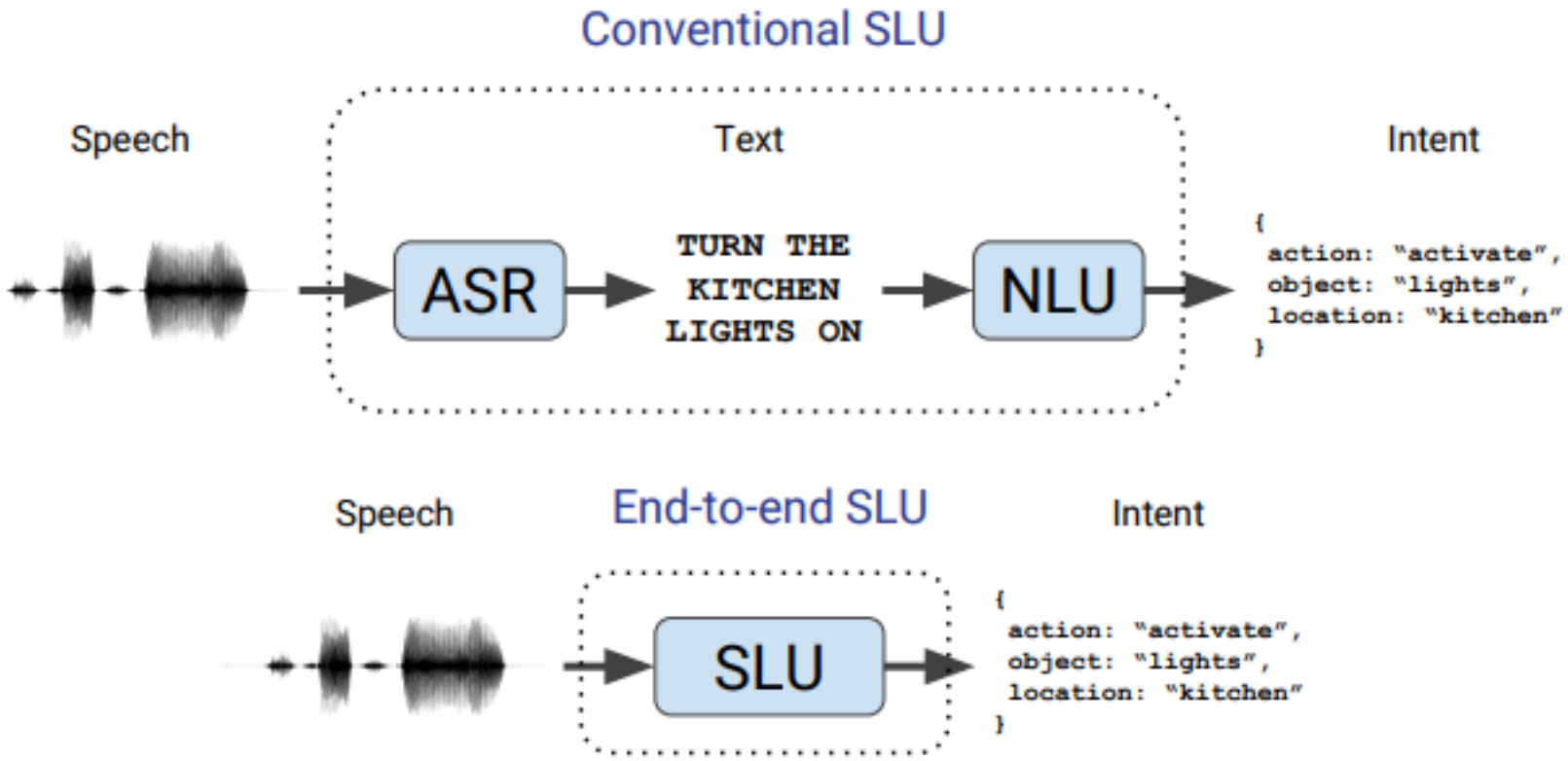

https://www.researchgate.net/publication/276298472_A_Real-Time_End-to-End_Multilingual_Speech_Recognition_Architecture

# Language Independent End-to-End Architecture For Joint Language and Speech Recognition




https://www.merl.com/publications/docs/TR2017-182.pdf

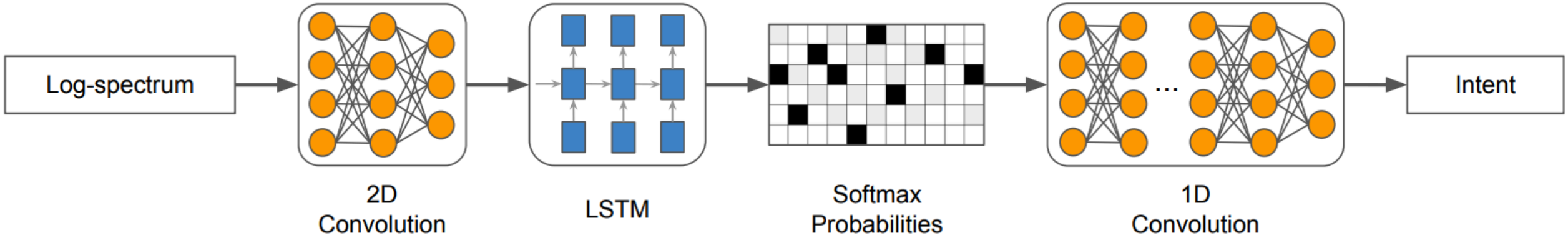
# Speech Model Pre-training for End-to-End Spoken Language Understanding




https://mila.quebec/wp-content/uploads/2019/08/1904.03670.pdf

https://arxiv.org/pdf/1904.03670.pdf

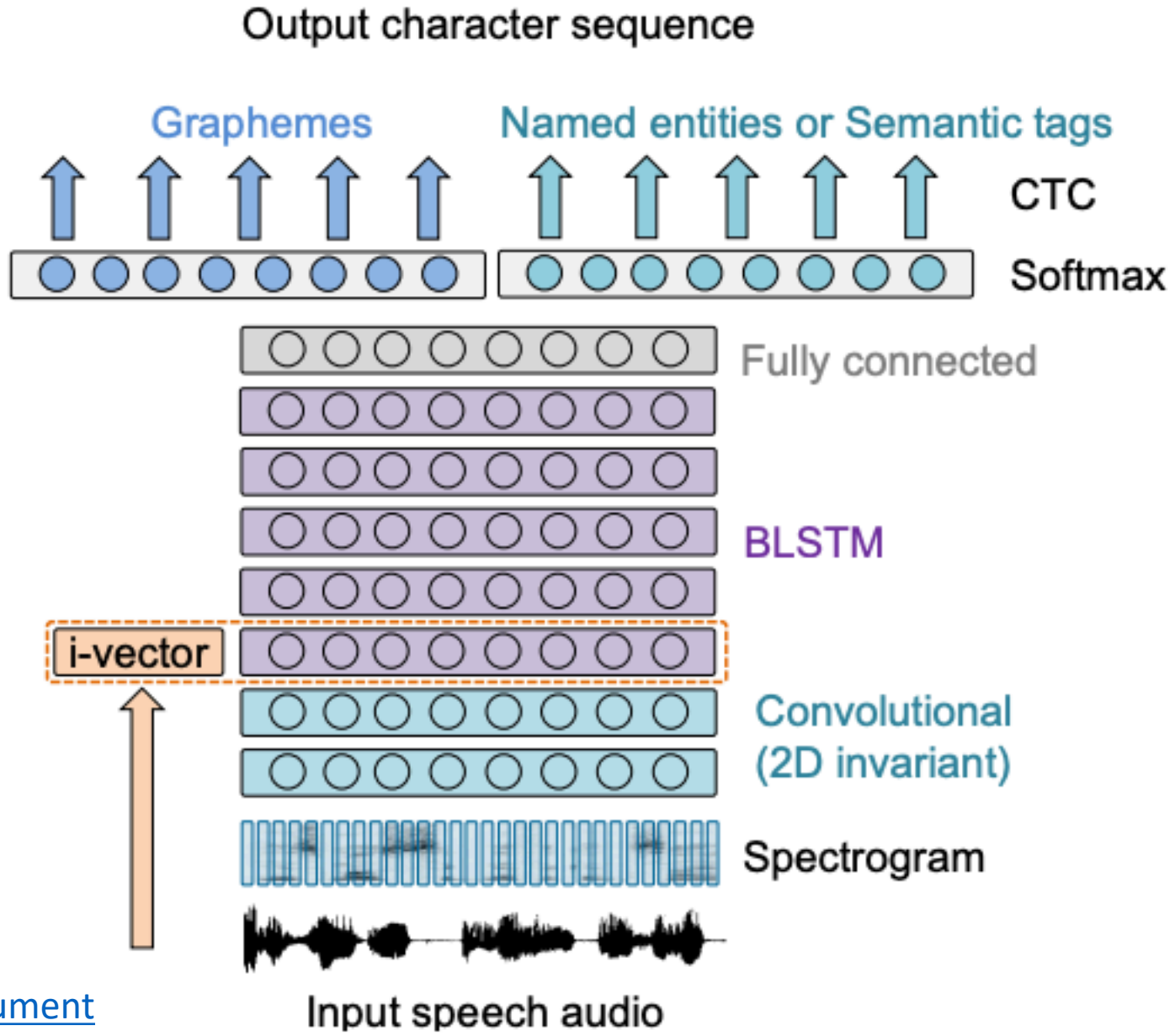
# FROM AUDIO TO SEMANTICS: APPROACHES TO END-TO-END SPOKEN LANGUAGE UNDERSTANDING




https://arxiv.org/pdf/1809.09190.pdf

http://yp-chen.com/static/YPChen18ICASSP_Spoken-Language-Understanding.pdf

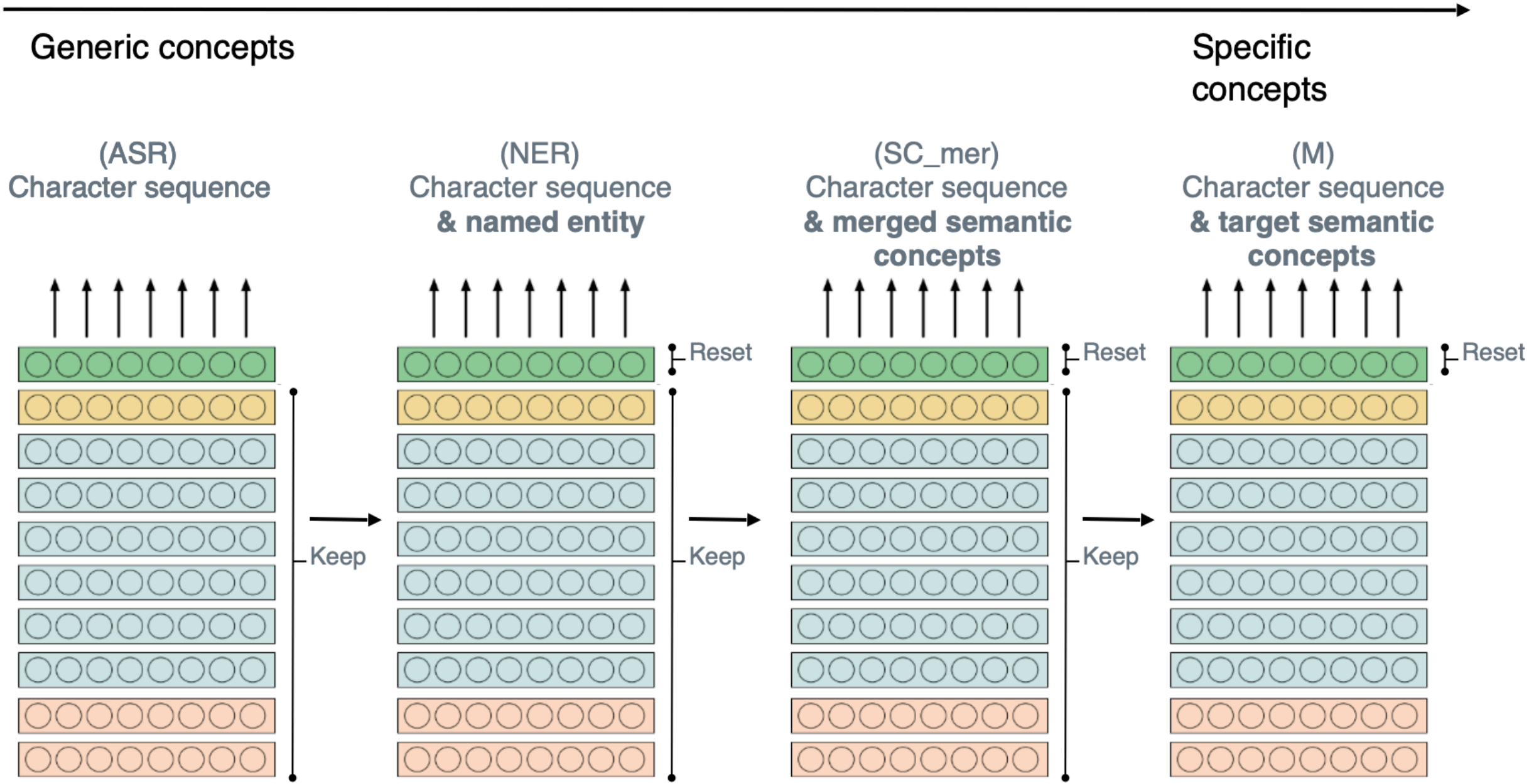
# Recent Advances in End-to-End Spoken Language Understanding



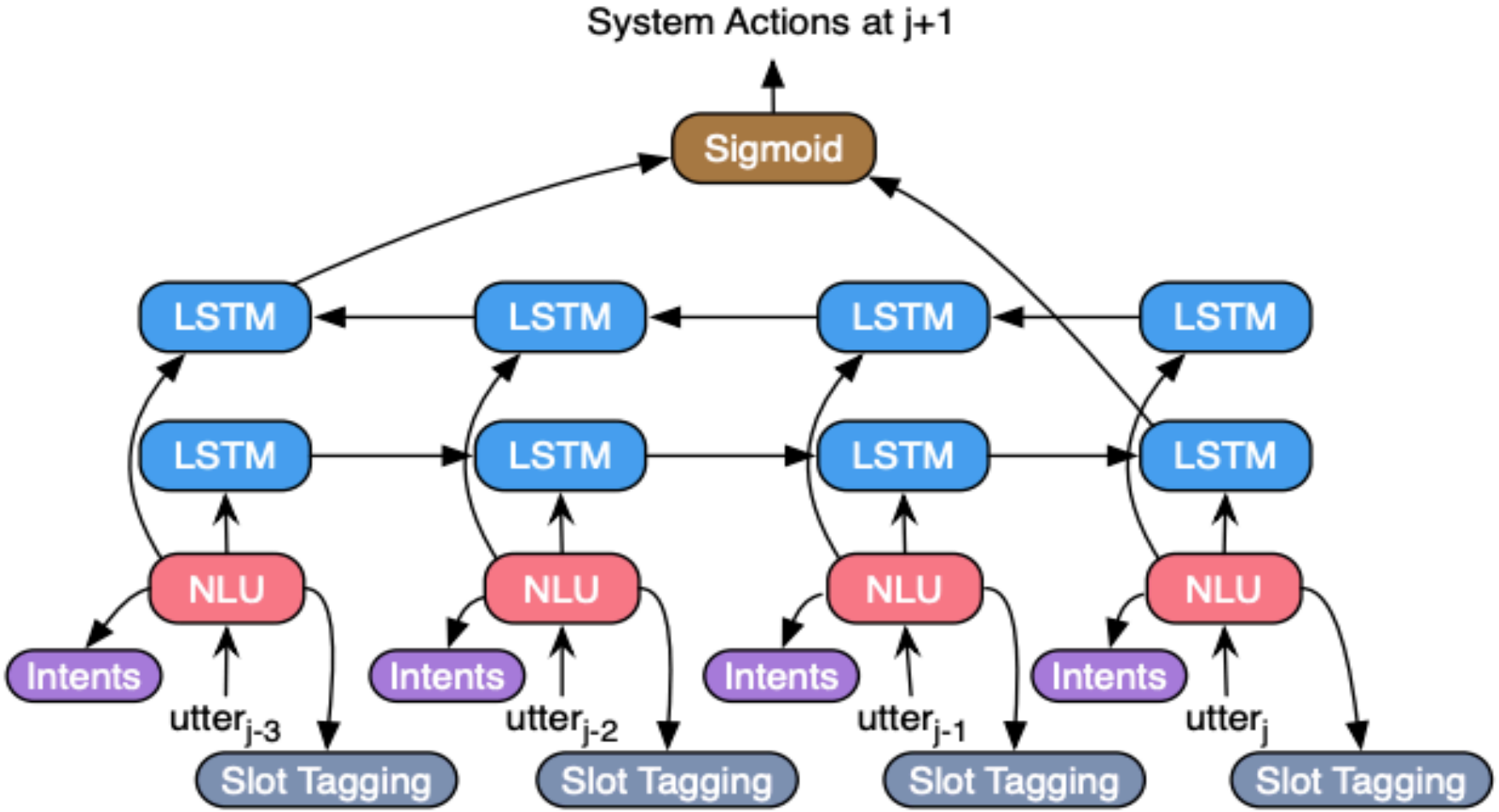

https://arxiv.org/pdf/1909.13332.pdf

https://hal.archives-ouvertes.fr/hal-02307811/document

http://csit.lp.edu.ua/presentation/esteve.pdf



# END-TO-END JOINT LEARNING OF NATURAL LANGUAGE UNDERSTANDING AND DIALOGUE MANAGER



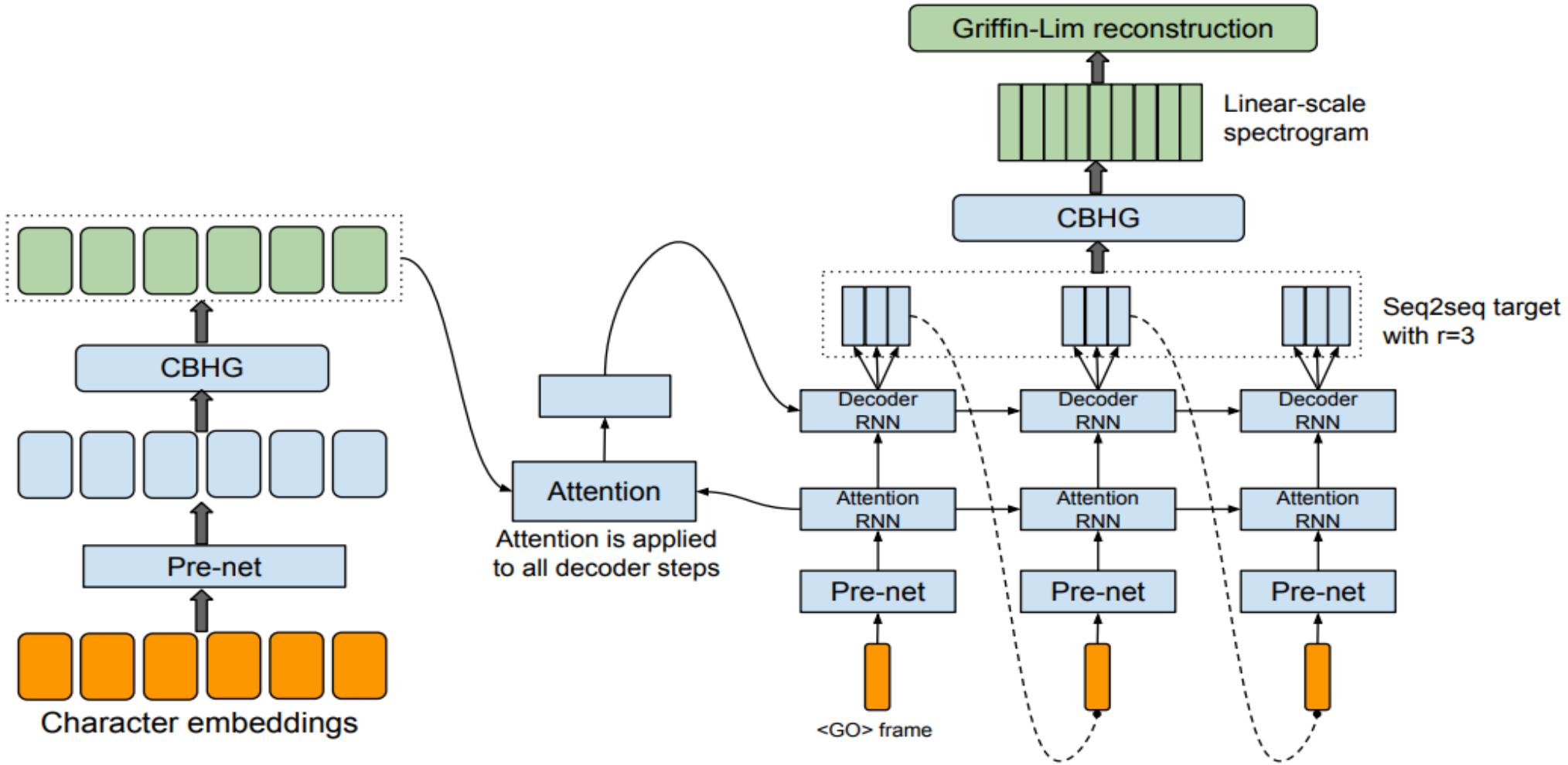

https://www.csie.ntu.edu.tw/~yvchen/doc/ICASSP17_E2E.pdf

# Text-to-Speech

- End-to-End TTS,
- Text normalisation,
- Lexical ambiguities/Heterophone Homographs,
- Graphemes-to-Phonemes,
- Prosody prediction,
- Emotional Speech,
- Personalisation,
- …

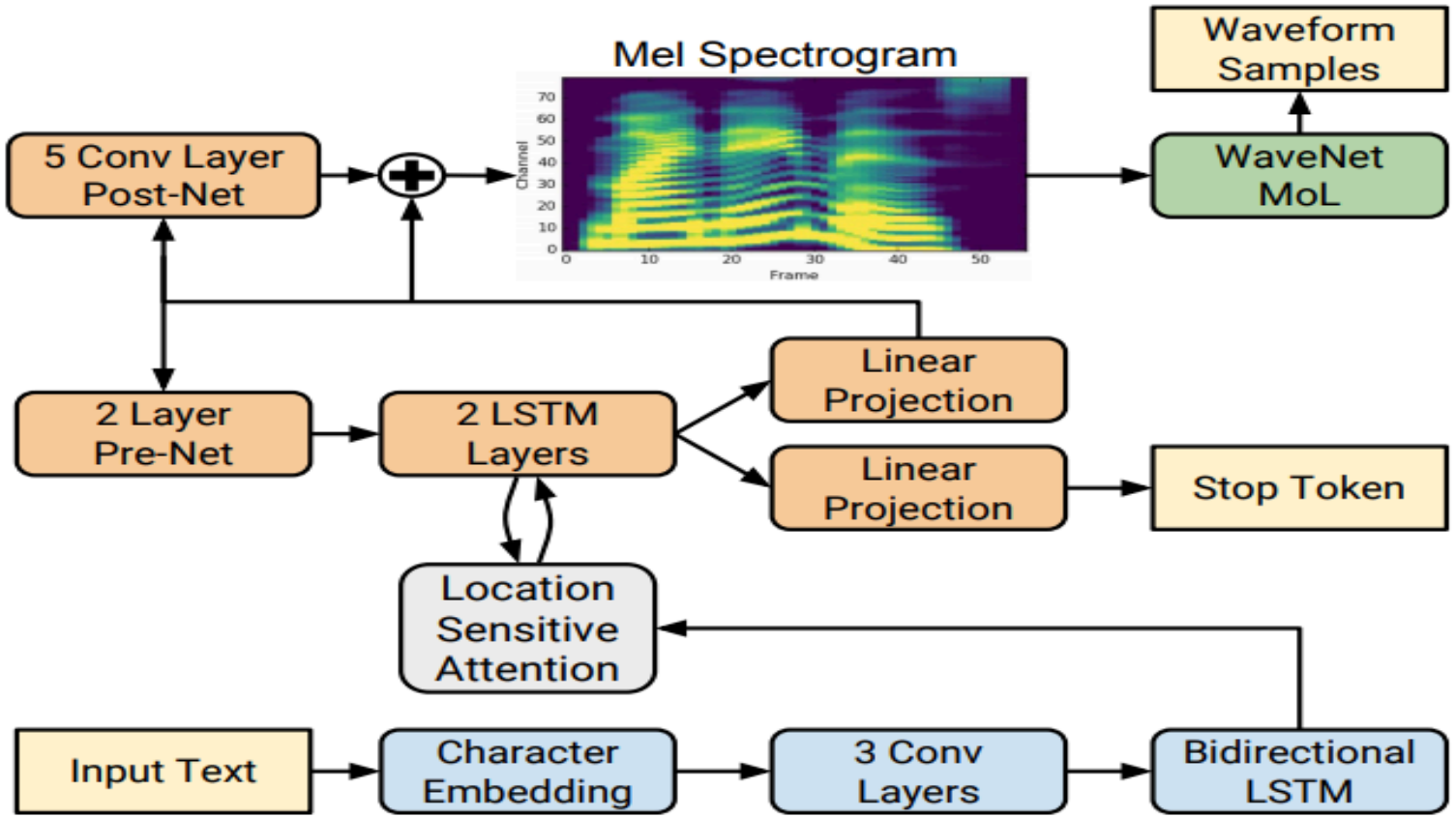
# https://google.github.io/tacotron/

https://arxiv.org/pdf/1703.10135.pdf



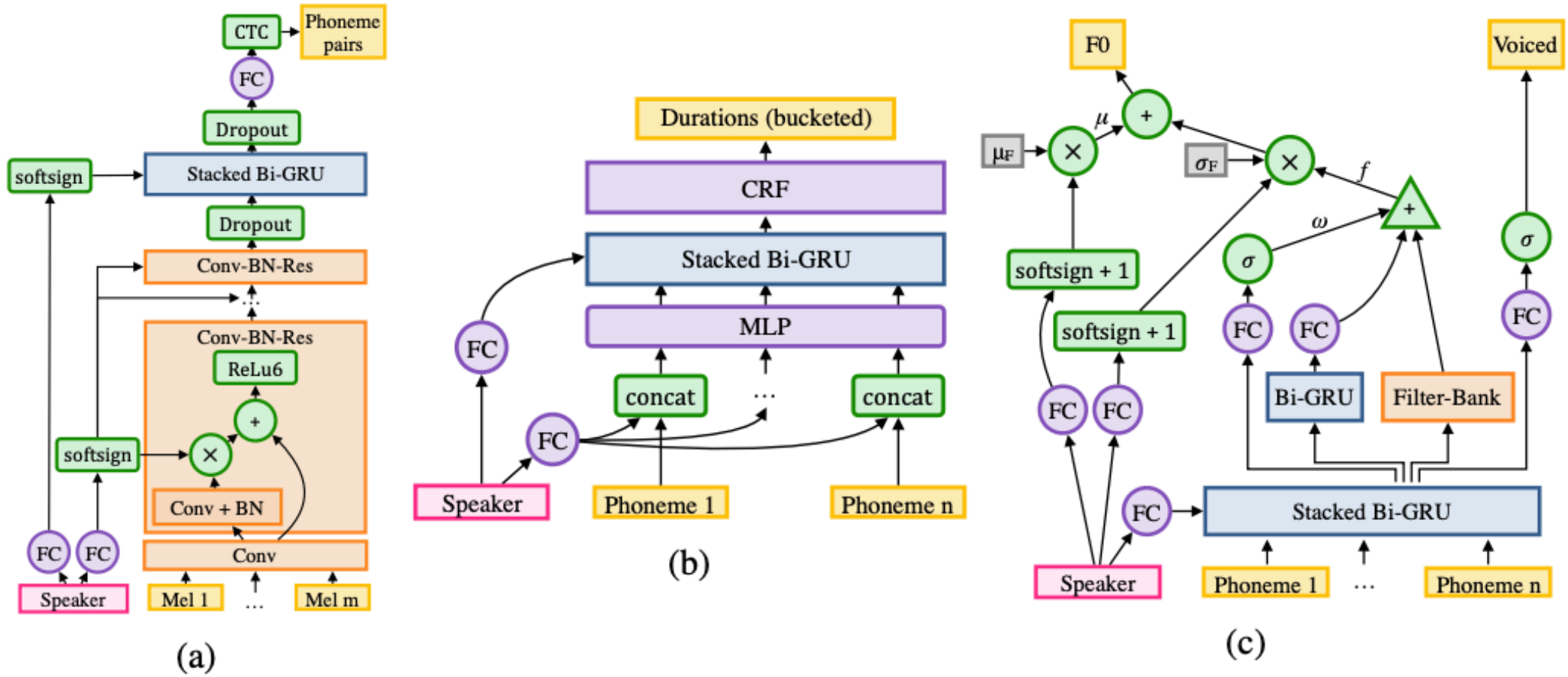
# NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS

https://arxiv.org/pdf/1712.05884.pdf



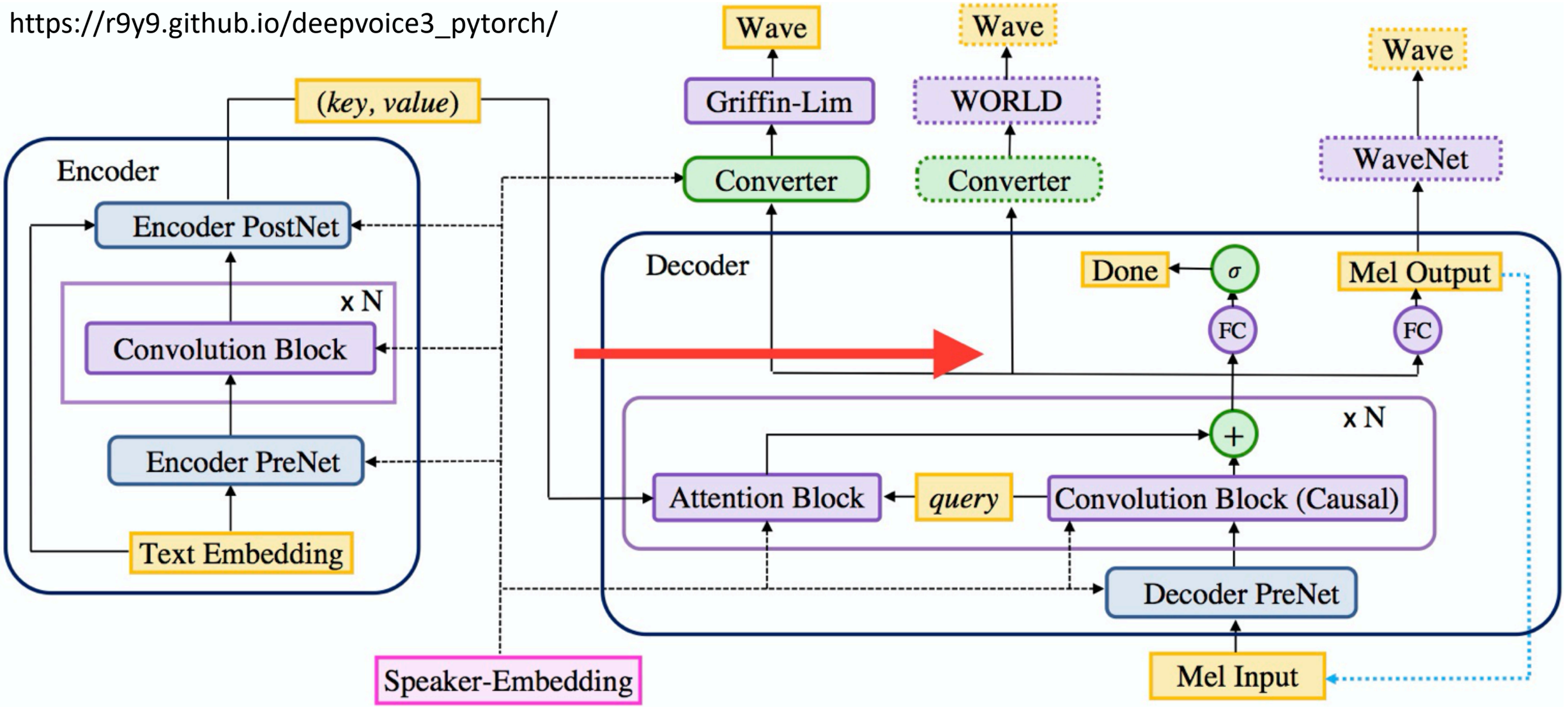
# Deep Voice 2: Multi-Speaker Neural Text-to-Speech

https://arxiv.org/pdf/1705.08947.pdf



# Deep Voice 3: 2000-Speaker Neural Text-to-Speech

https://r9y9.github.io/deepvoice3_pytorch/

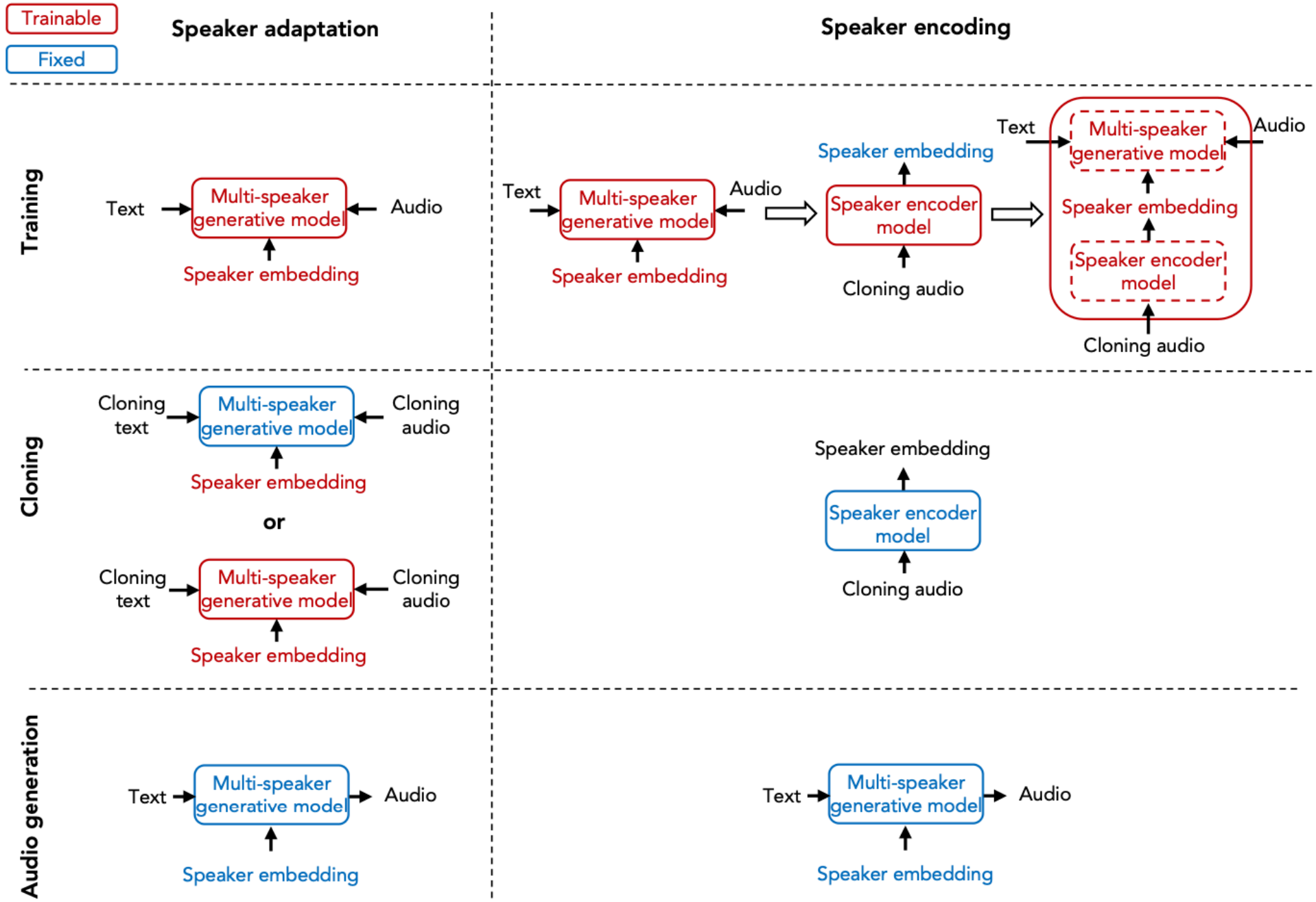



http://research.baidu.com/Blog/index-view?id=91



https://arxiv.org/pdf/1710.07654.pdf

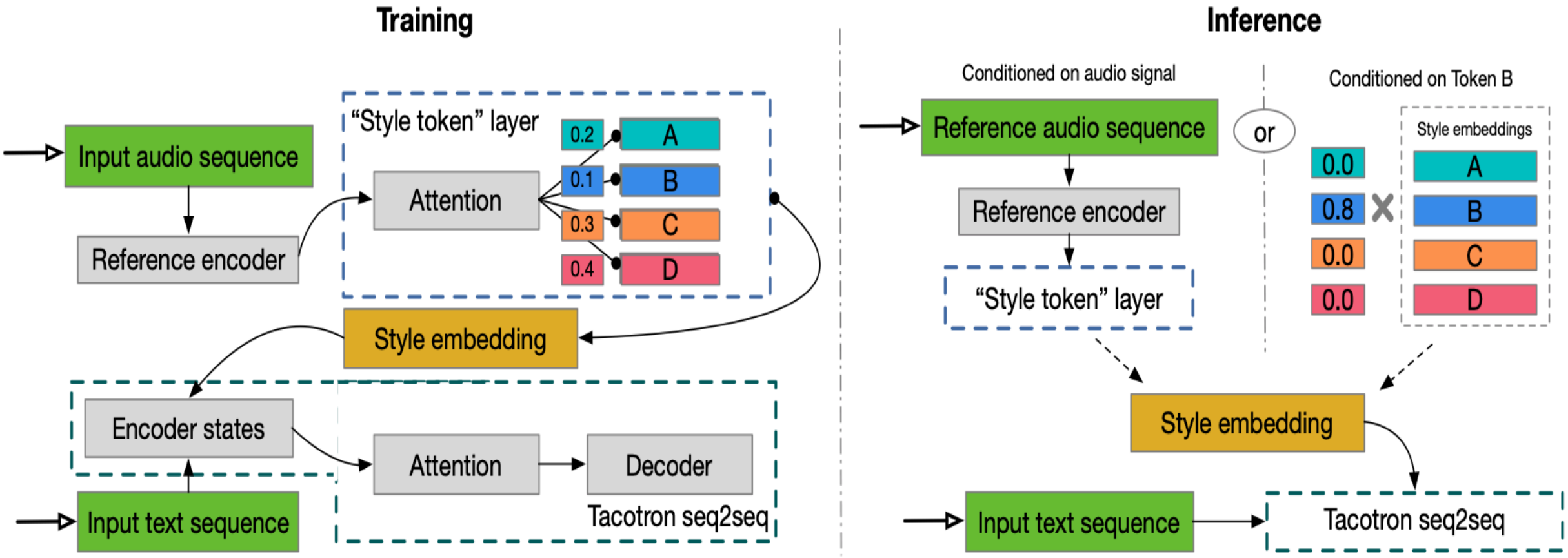
# Neural Voice Cloning with a Few Samples




https://arxiv.org/pdf/1802.06006.pdf

https://audiodemos.github.io

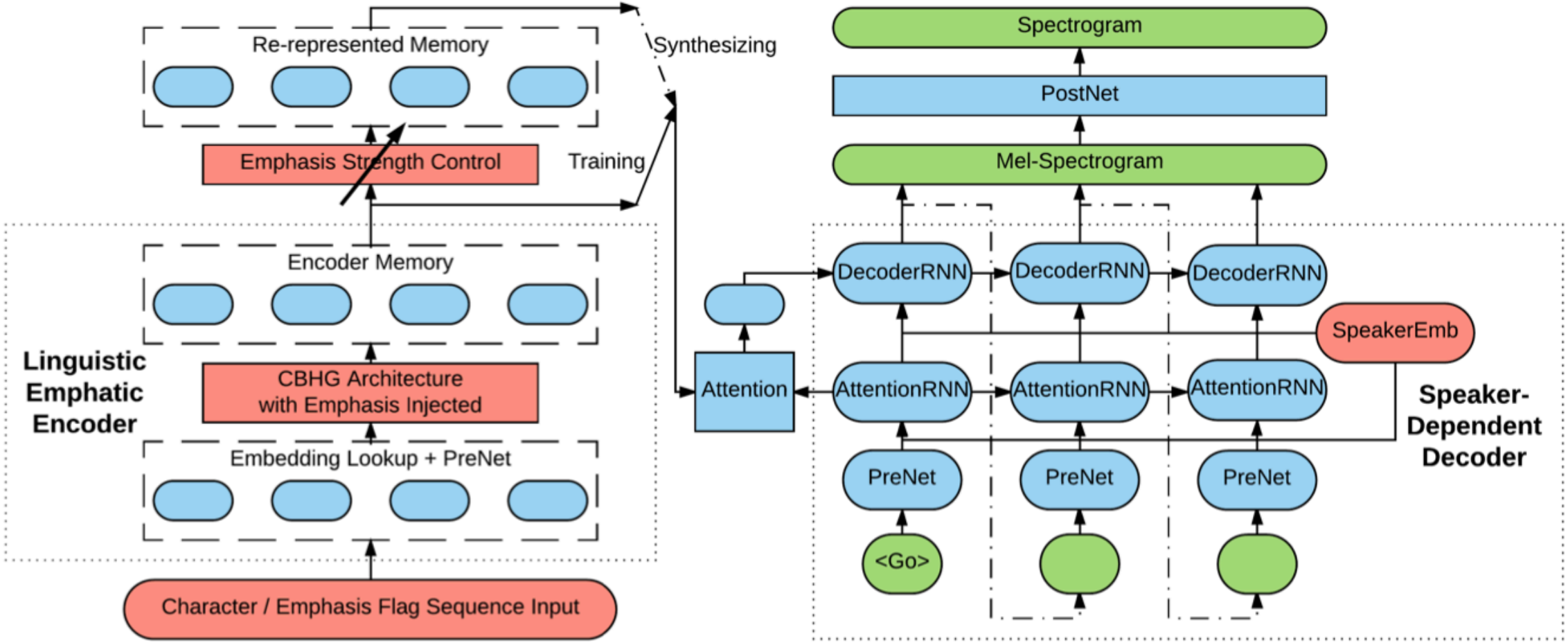
# Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis




https://arxiv.org/pdf/1803.09017.pdf

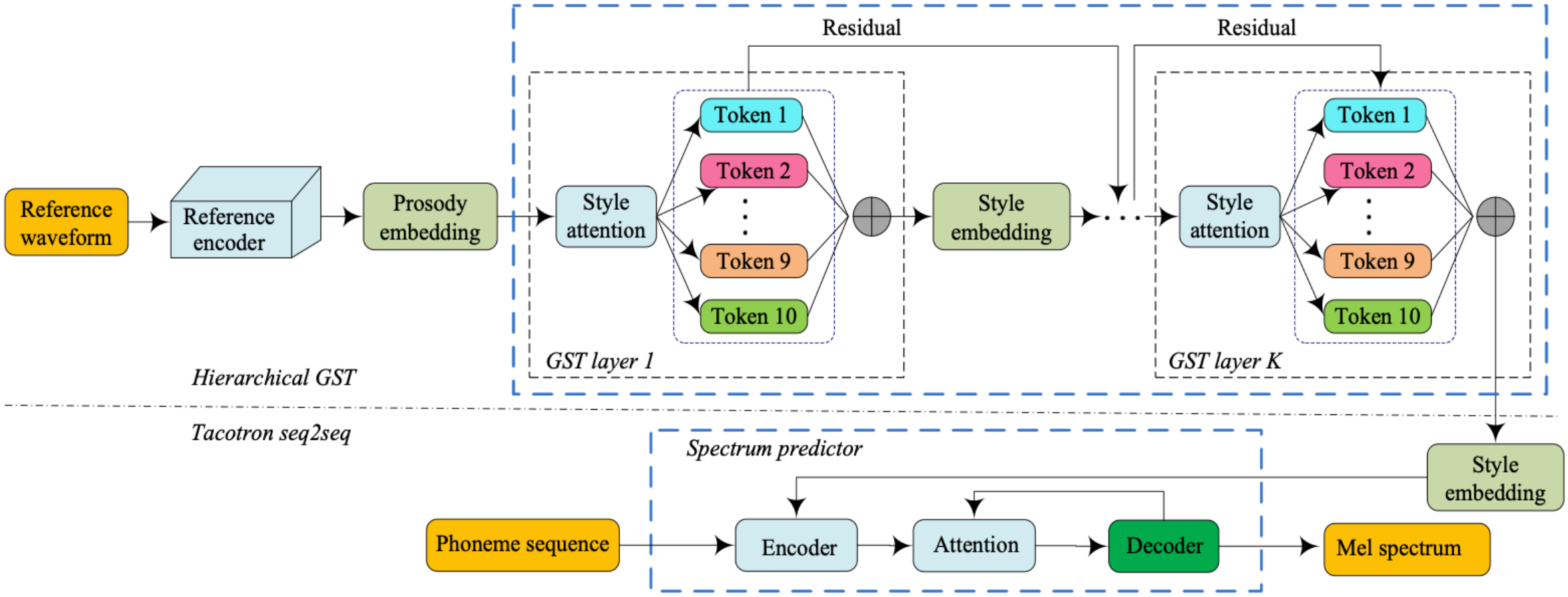
https://google.github.io/tacotron/publications/global_style_tokens

# Emphatic and Control Based on Characteristic Transferring in End-to-End Speech Synthesis




http://www1.se.cuhk.edu.hk/~hccl/publications/pub/2018_201805_ACII_Asia_WangMu.pdf

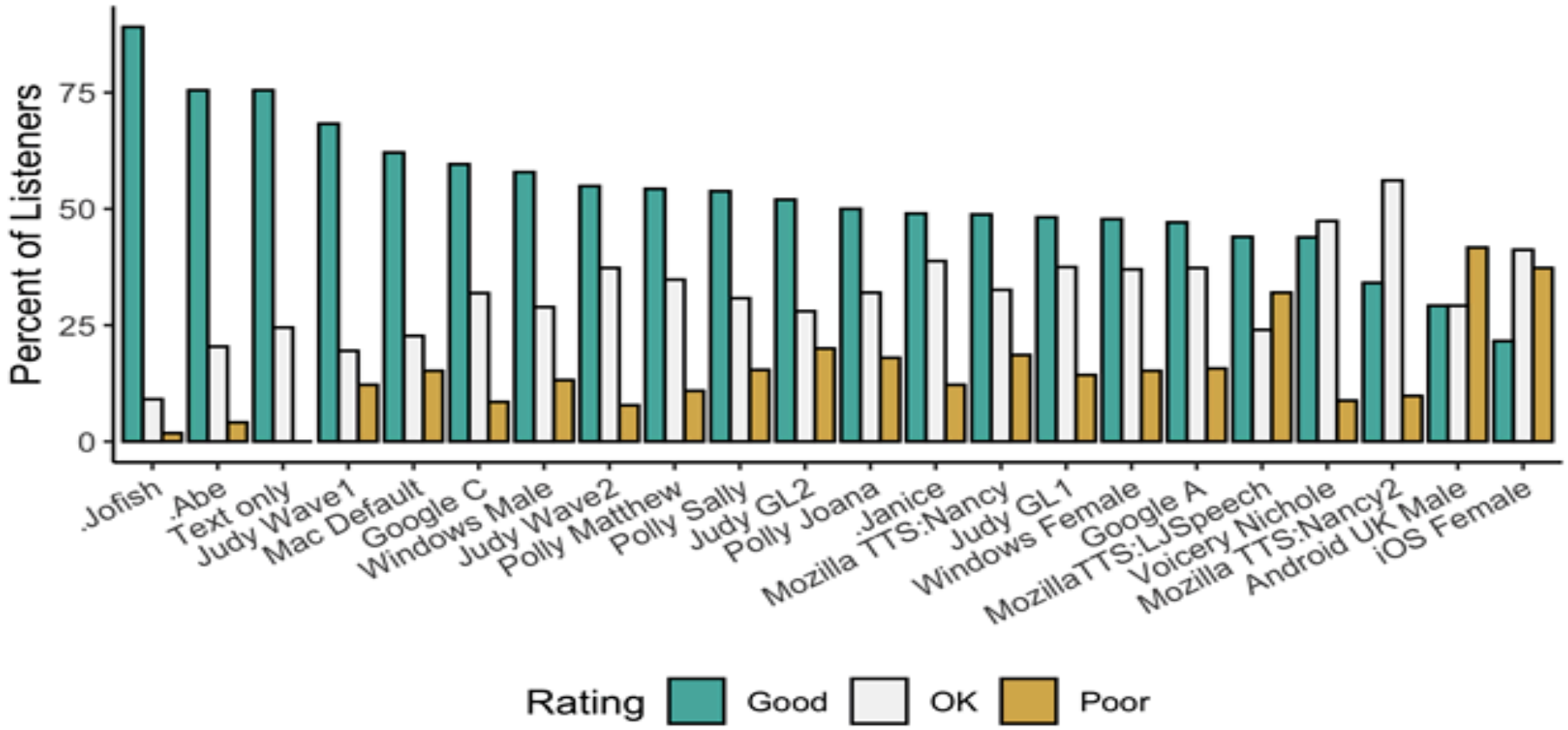
# LEARNING HIERARCHICAL REPRESENTATIONS FOR EXPRESSIVE SPEAKING STYLE IN END-TO-END SPEECH SYNTHESIS



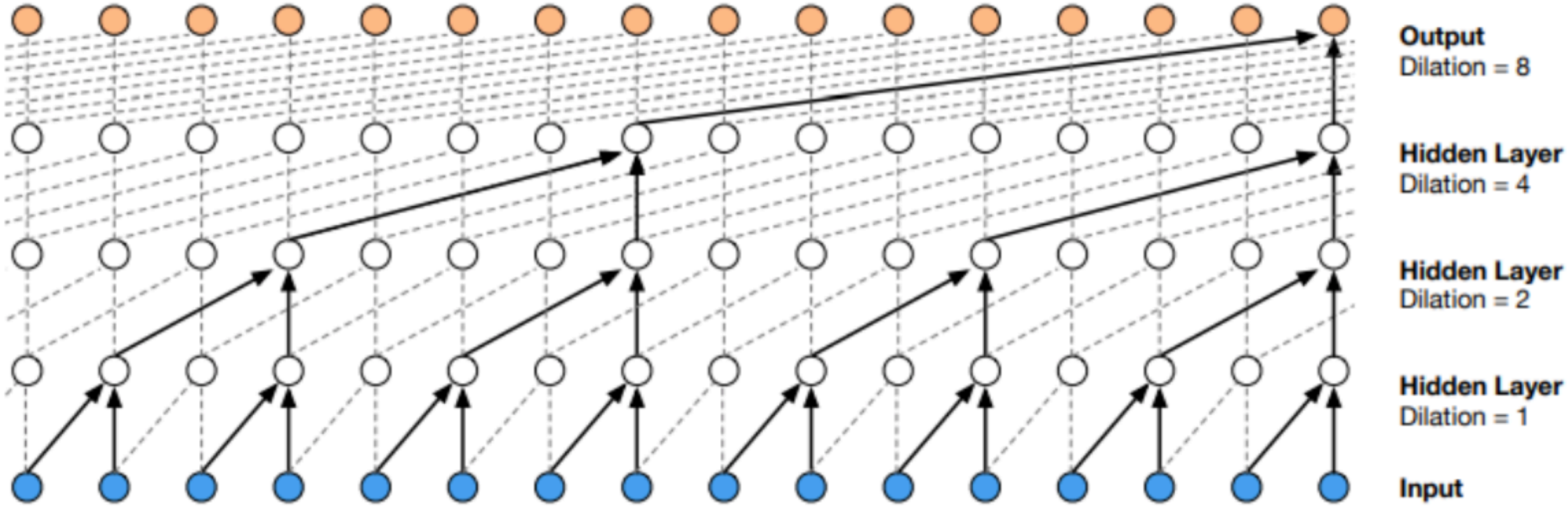

http://lxie.npu-aslp.org/papers/2019ASRU_AXC.pdf

# Mozilla Deep Learning based Text2Speech model

https://github.com/mozilla/TTS



# WAVENET: A GENERATIVE MODEL FOR RAW AUDIO




https://lh3.googleusercontent.com/XMJlFDtiFU7WtZYnPDMDhT6jTFEXvTpY7c7sY70PtTkzEIDm6DvUv5nvHHwbOrUrDckS34aIu5fiZx0615p1-nf3rFxzhza387ud=w2048-rw-v1

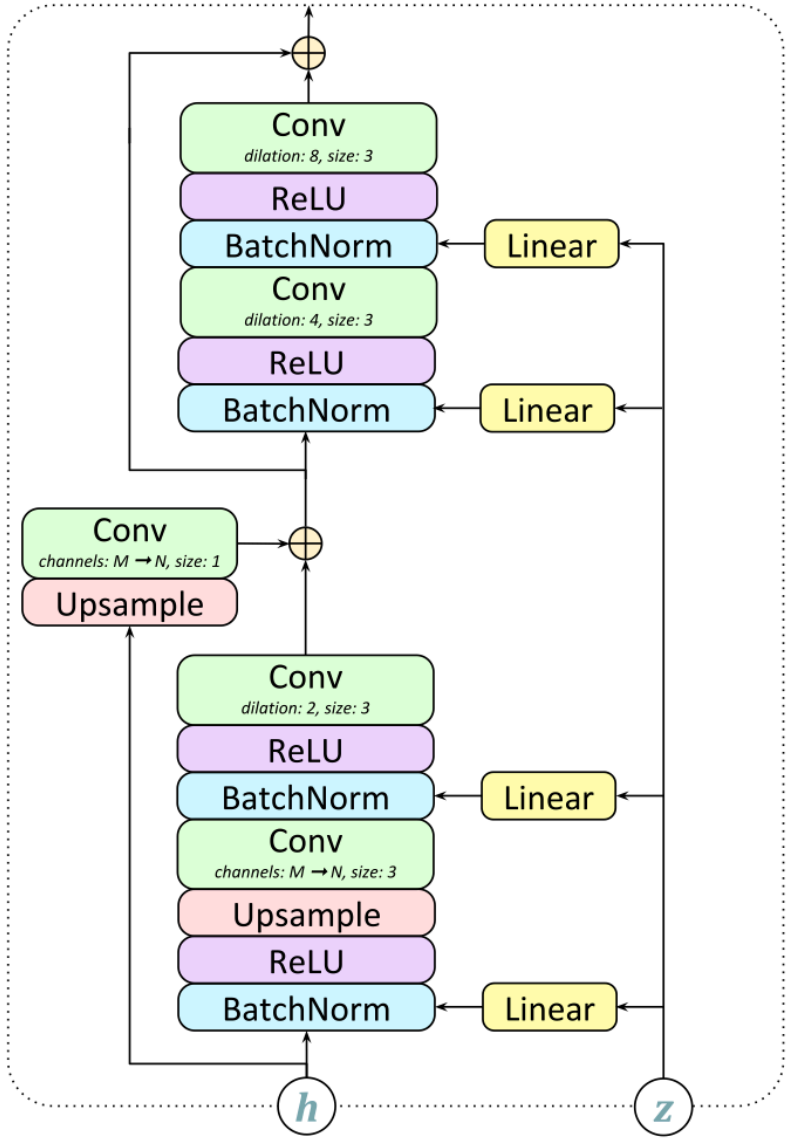
https://www.deepmind.com/blog/article/wavenet-generative-model-raw-audio

https://arxiv.org/pdf/1609.03499.pdf

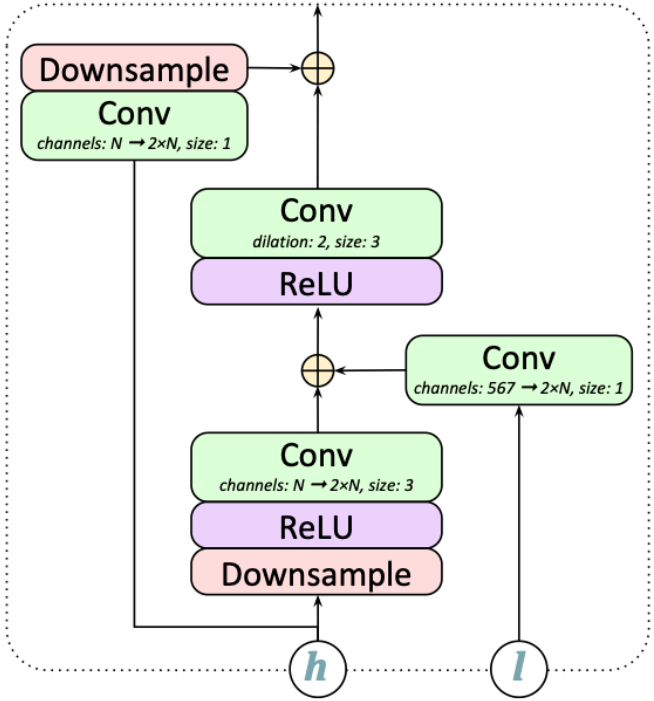
# HIGH FIDELITY SPEECH SYNTHESIS WITH ADVERSARIAL NETWORKS

DeepMind

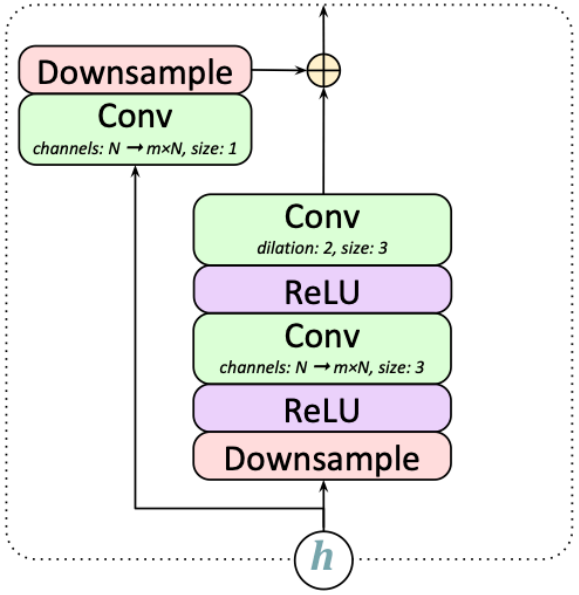
https://arxiv.org/pdf/1909.11646v2.pdf




(a) GBlock



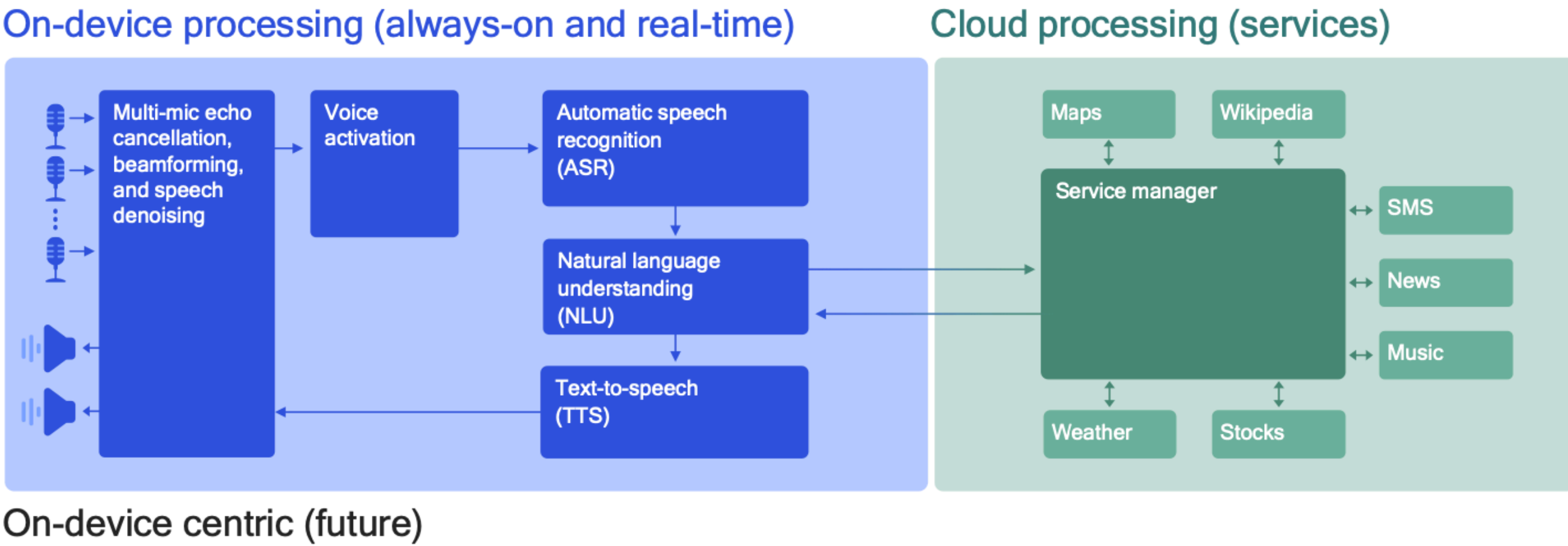

(b) Conditional DBlock




(c) DBlock

# Moving voice UI functionality to the end device

## An end-to-end solution powered by machine learning




On-device centric (future)

https://www.qualcomm.com/media/documents/files/making-an-on-device-personal-assistant-a-reality.pdf

# Some conclusions

- End-To-End Machine Learning (ETEML) is a simple solution to train models which are able to solve problems for which one is lacking (or do not trust) a-priori knowledge.
- ETEML is particularly suitable in (un-supervised/semi-supervised) situations for which large amounts of non-annotated data is available.
- ETEML should be bundled (fused) in hybrid solutions with existing (traditional, modular) and symbolic AI techniques.

# End-to-End Machine Learning for Speech Processing: Speech-to-Speech, Speech-to-Text and Text-to-Speech


Gérard Chollet (CNRS-SAMOVAR)

and Colleagues from IV, Zaion, CNRS/IMT and SMI

gerard.chollet@telecom-sudparis.eu


Intelligent Voice®

zaion TIME TO TALK

Une école Institut Mines-Télécom

TELECOM SudParis

cnrs

Depuis 80 ans, nos connaissances bâtissent de nouveaux mondes

speechmorphing