HỌC MÁY TRONG XỬ LÝ TIẾNG NÓI

Đỗ Văn Hải

haidovan@gmail.com

Trung tâm Không gian mạng Viettel

Khoa CNTT – ĐH Thủy Lợi

Các sản phẩm trên thế giới

Google Assistant

amazon echo

SIRI

Microsoft

Hi. I'm Cortana.

Ask me a question!

Ứng dụng tiếng nói tiếng Việt

Virtual Assistant

Smart Home

Voice Messaging

Viettel Voice Note
Viettel Group Application Music & Audio
Everyone

Interview
Meeting Note

BIZ
Reputa
Viettel Group Application Business

Social Listening

Supervise Call Centers
Telesales

Call-Bot

Audio NewsPaper

Các công nghệ xử lý tiếng nói

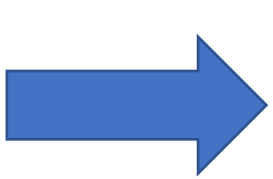
1. Text-to-Speech (Tổng hợp tiếng nói)
2. Speech-to-Text (Nhận dạng tiếng nói)
3. Speech Emotion Recognition (Nhận dạng ngữ điệu)
4. Speech Accent Recogntion (Nhận dạng phương ngữ)
5. Speech Quality Assessment (Đánh giá chất lượng tiếng nói)
6. Speaker Verification/ Recognition (Nhận dạng/ xác thực người nói)
7. Speaker Diarization (Phân biệt người nói trong cuộc họp)


TTS
Text To Speech

Tổng hợp tiếng nói là gì?

text

Author of the…



waveform

Ứng dụng của TTS

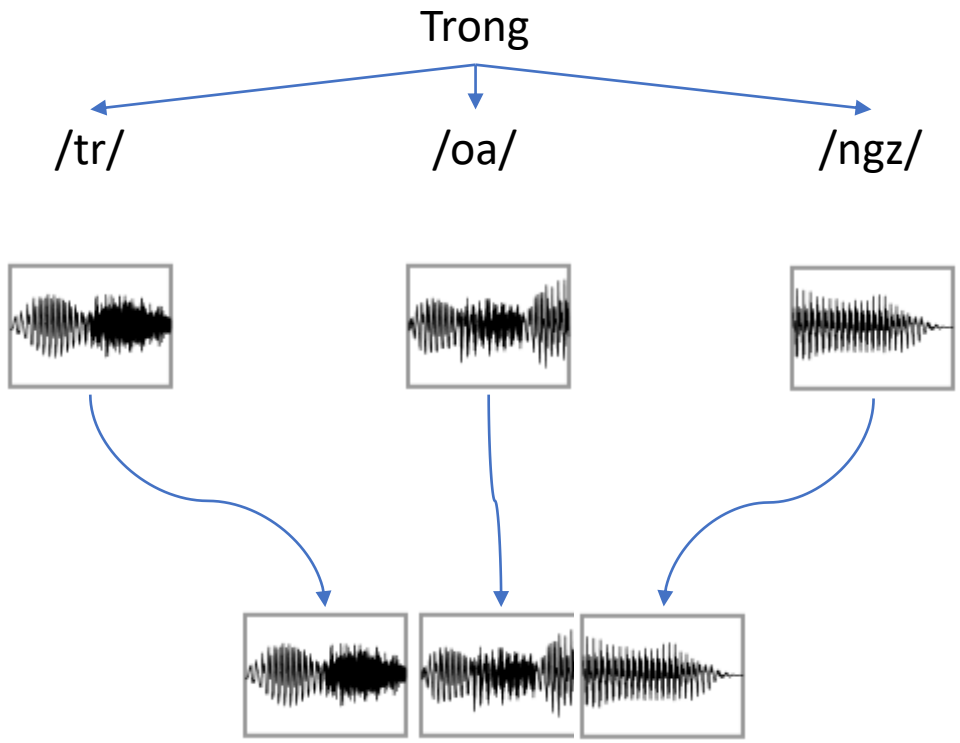
- Trợ lý ảo/ robot/ callbot
- Báo nói
- Sách nói
- Thuyết minh phim
- Tạo video bài giảng tự động

Các phương pháp của TTS

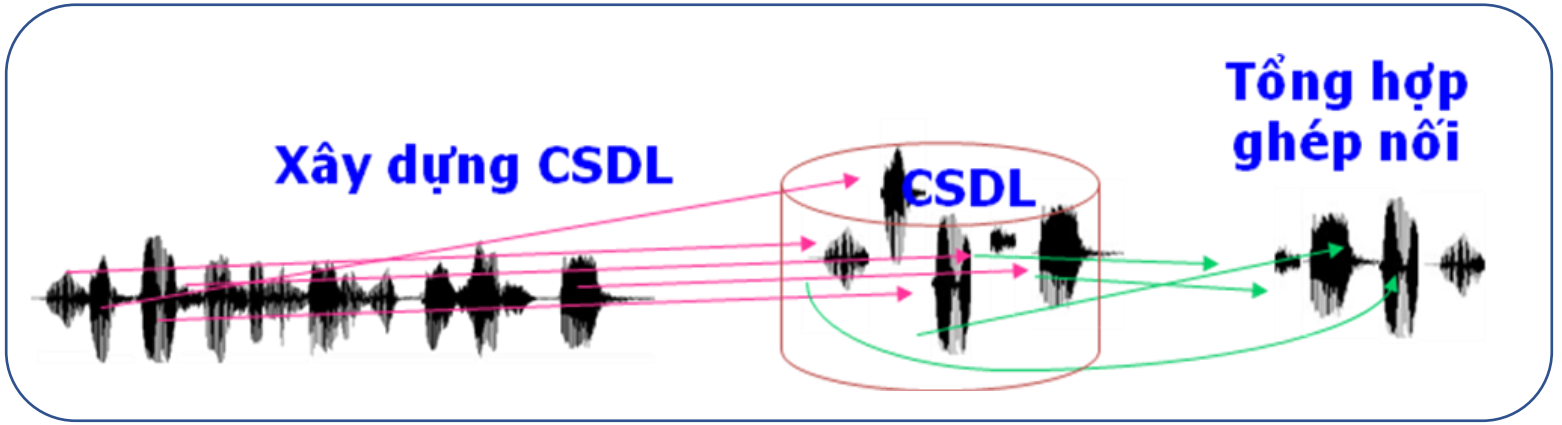
- Formant synthesis
- Concatenation synthesis
- HMM-based synthesis
- Deep learning: Merlin, Wavenet, Tacotron, ...

  Nguyen Van Thinh, Nguyen Quoc Bao, Phan Huy Kinh, Do Van Hai, "Development of Vietnamese Speech Synthesis System using Deep Neural Networks", *Journal of Computer Science and Cybernetics, V.34, N.4 (2018).*

Tổng hợp ghép nối



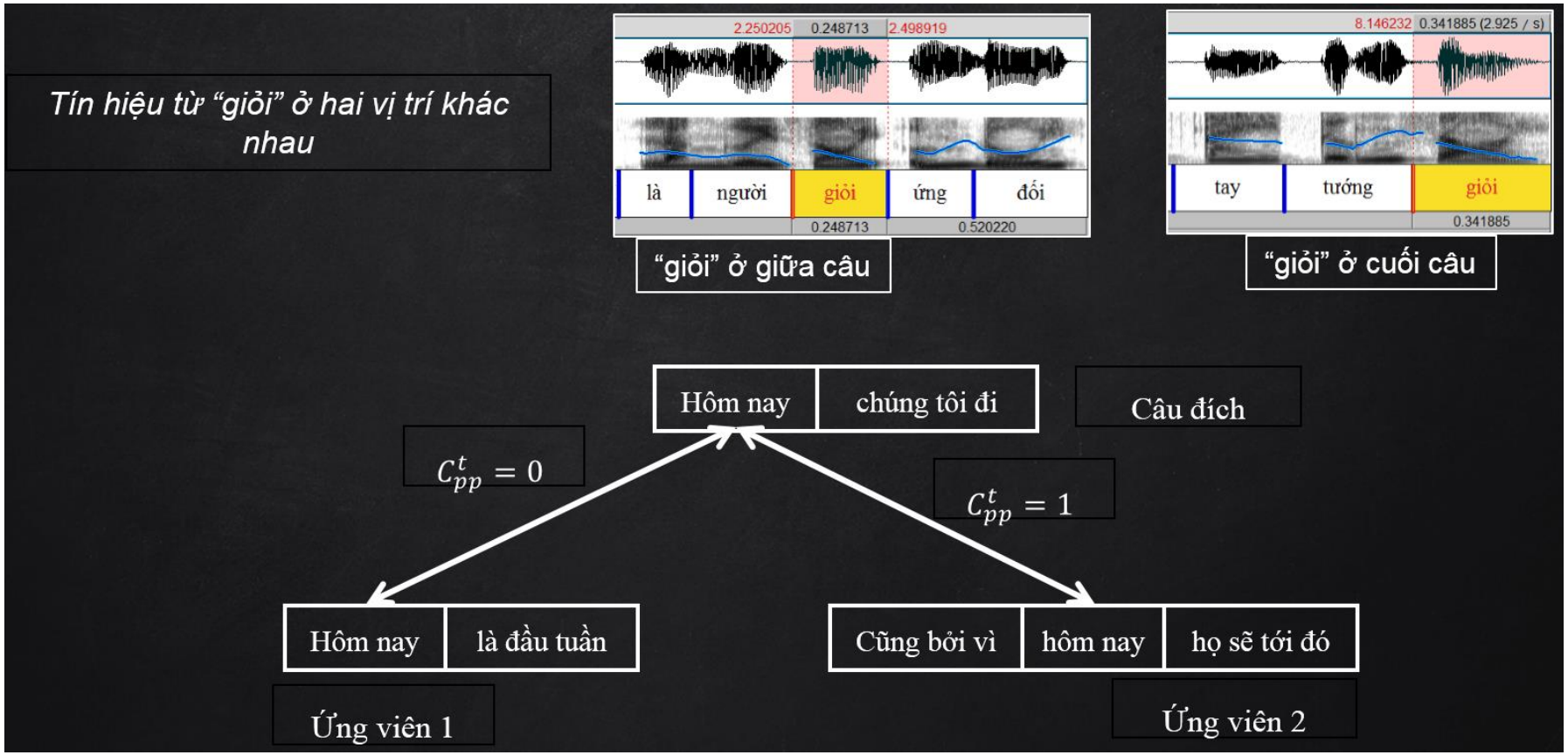
Tổng hợp ghép nối



- CSDL cần bao nhiêu đơn vị âm?
- Các loại đơn vị âm.

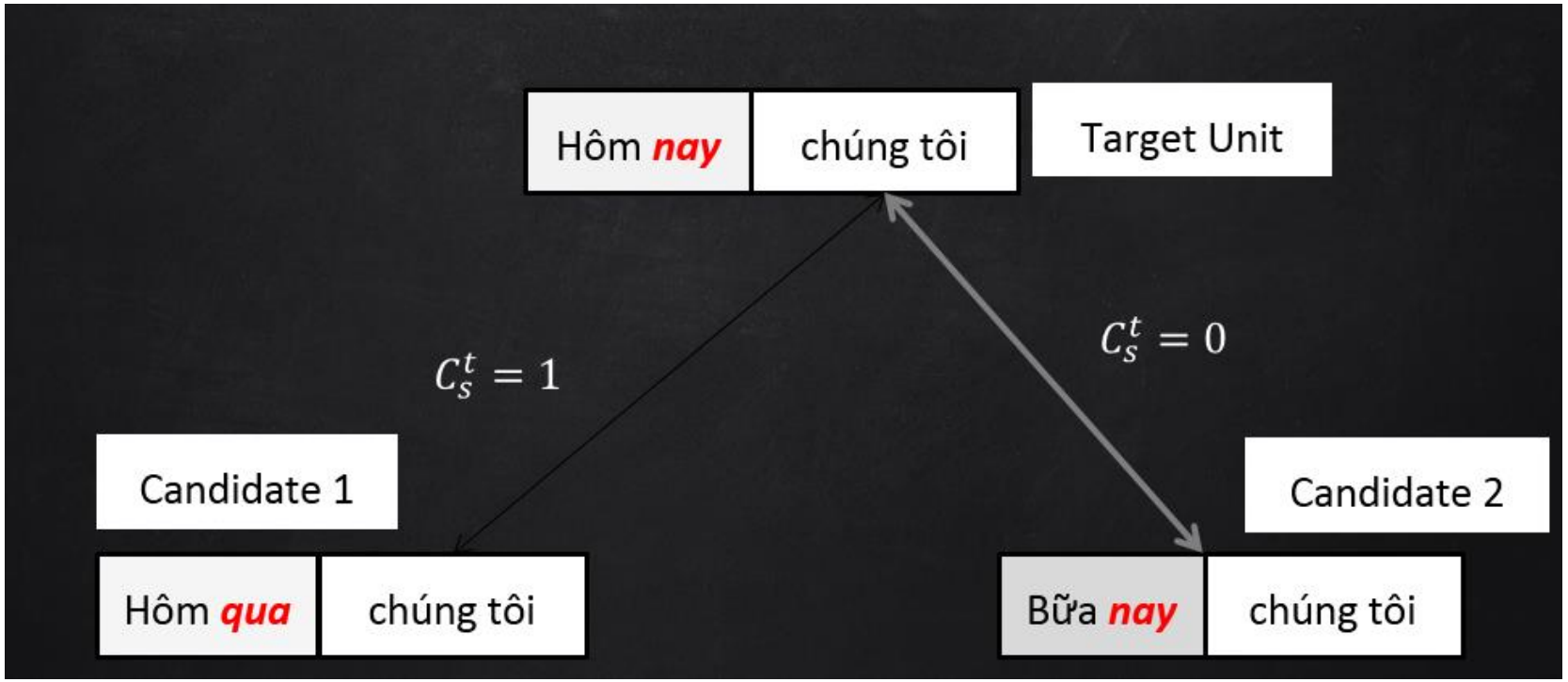
Tổng hợp ghép nối

phụ thuộc vào vị trí từ trong câu



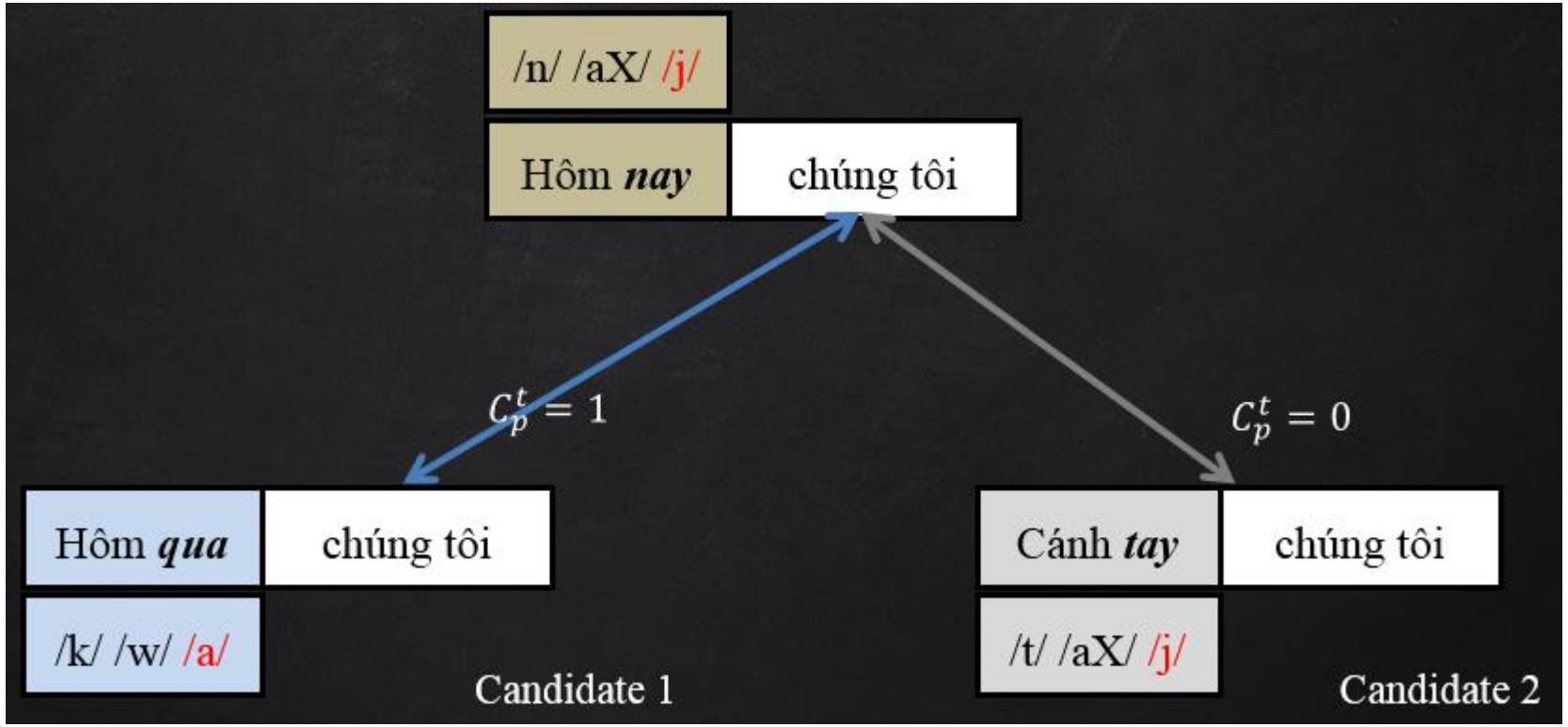
Tổng hợp ghép nối

phụ thuộc vào âm tiết đứng liền trước

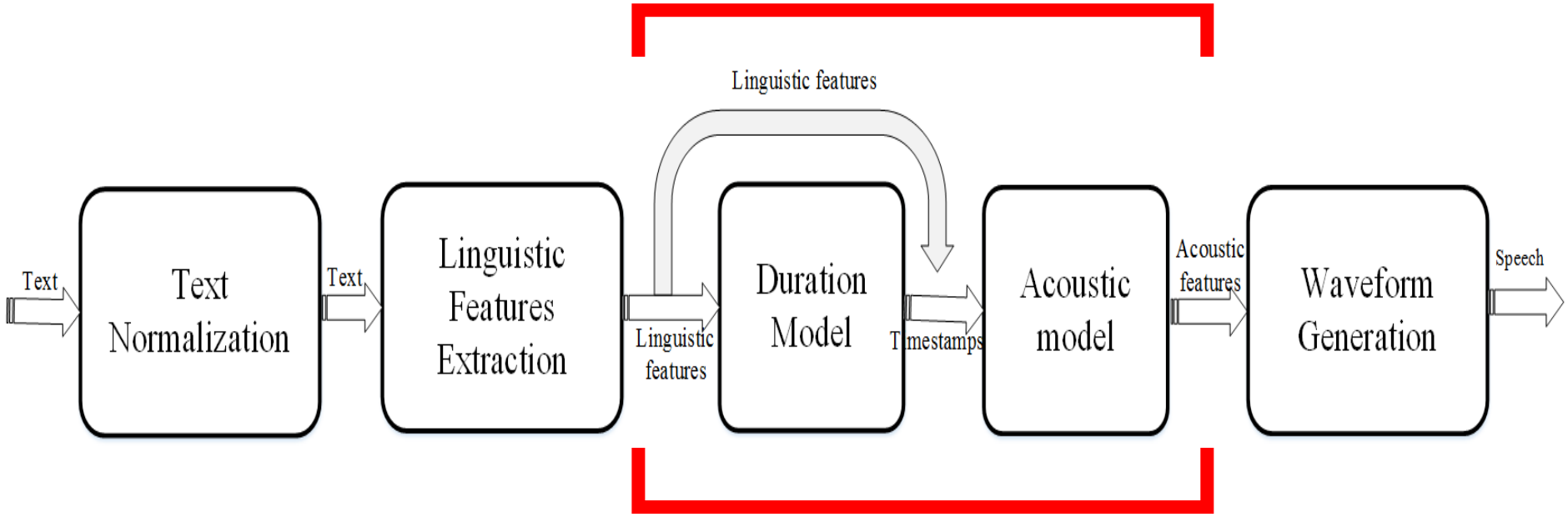


Tổng hợp ghép nối

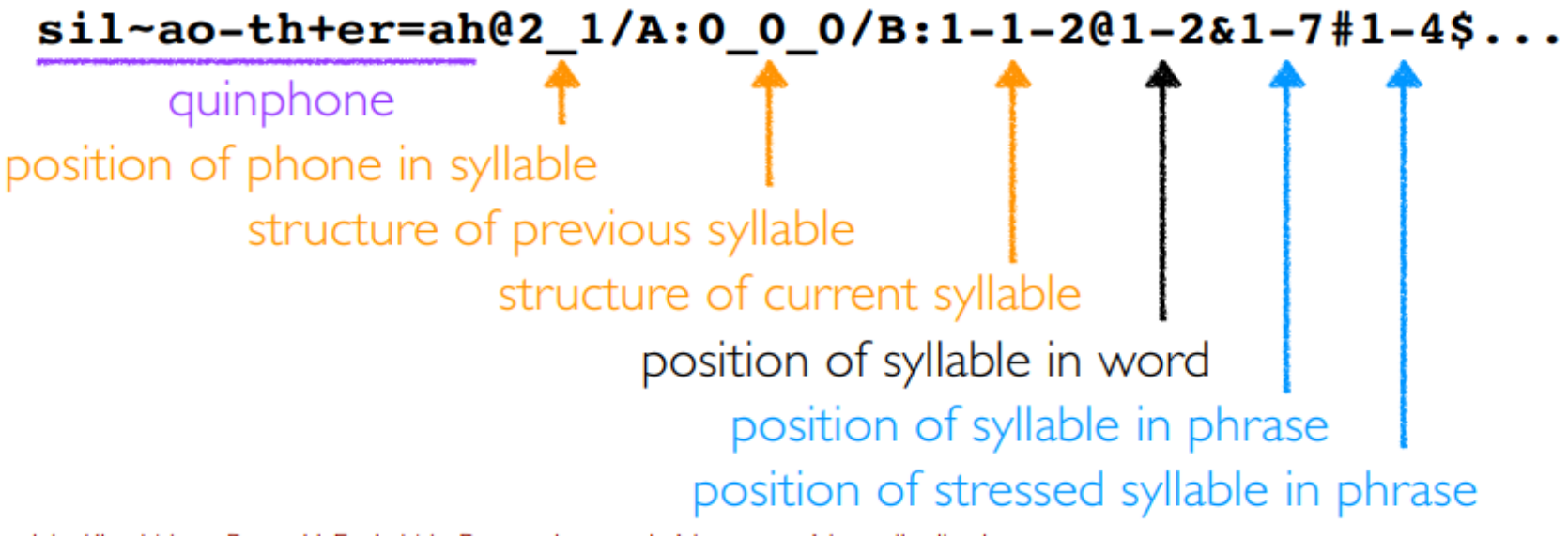
phụ thuộc vào âm vị đứng liền trước



Tổng hợp dựa trên mạng nơ-ron

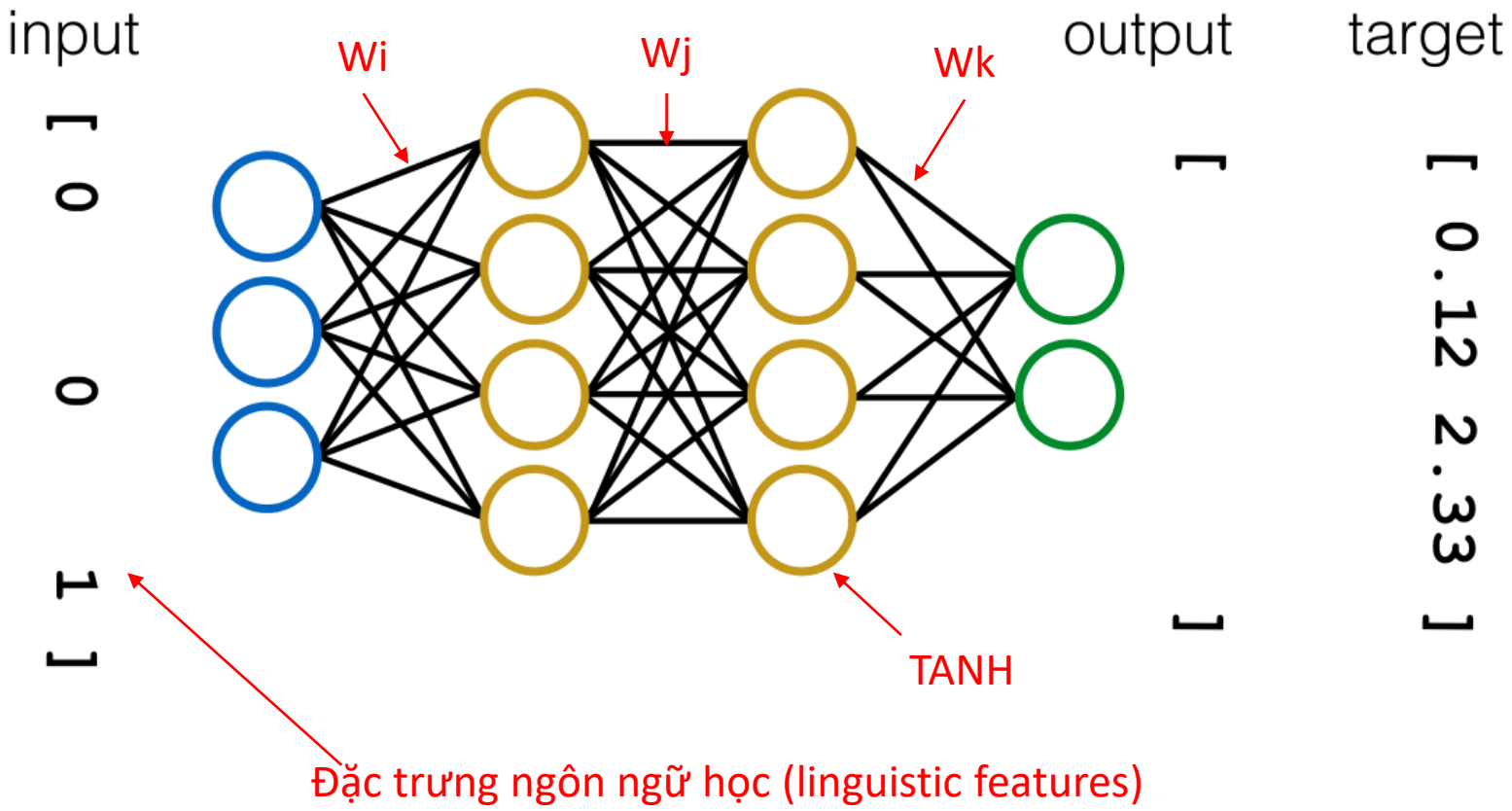


Trích chọn đặc trưng ngôn ngữ

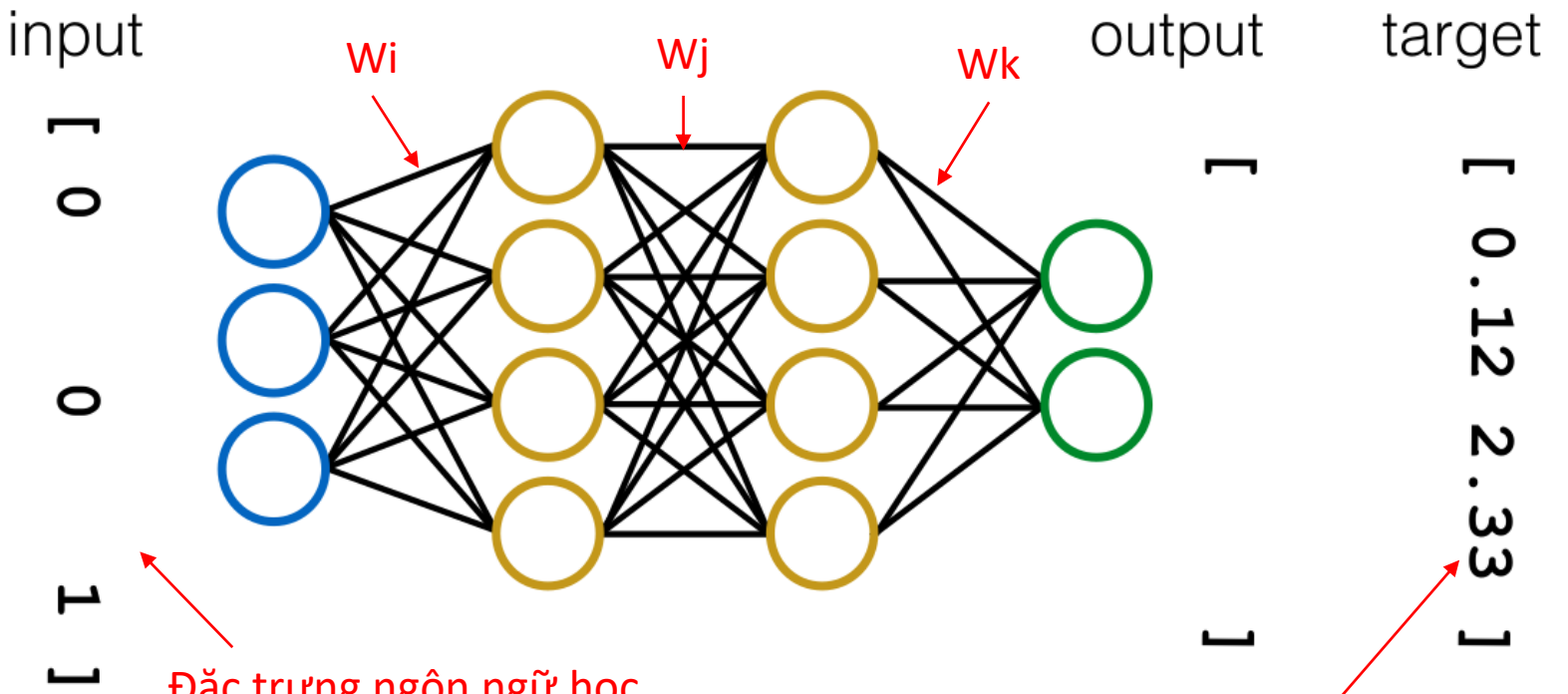


Biểu diễn đặc trưng ngôn ngữ

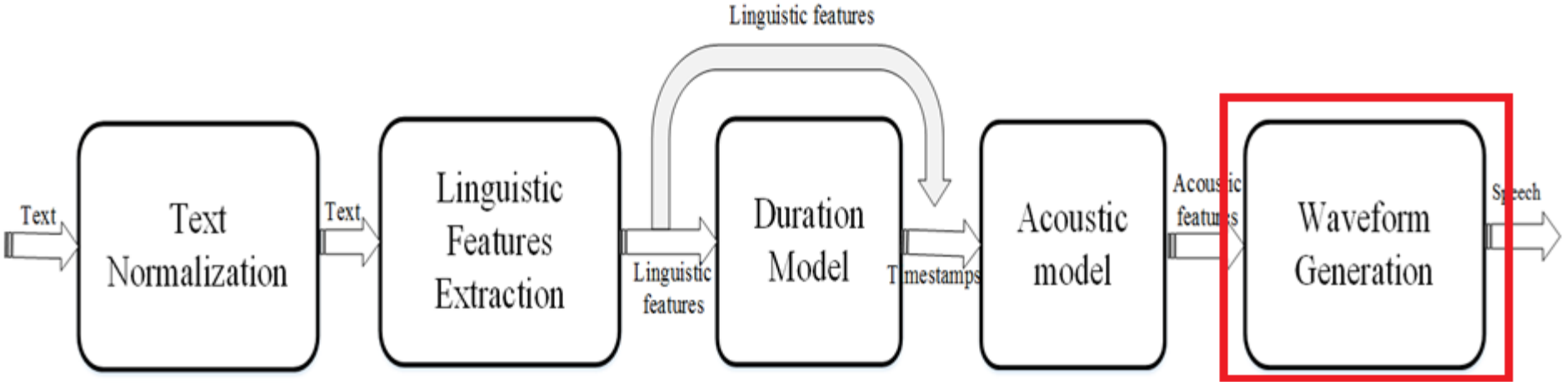
Mô hình thời gian - Duration model



Mô hình âm học - Acoustic Model

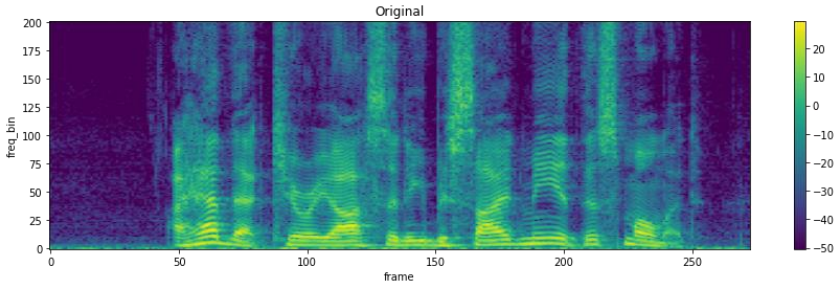


Vocoder

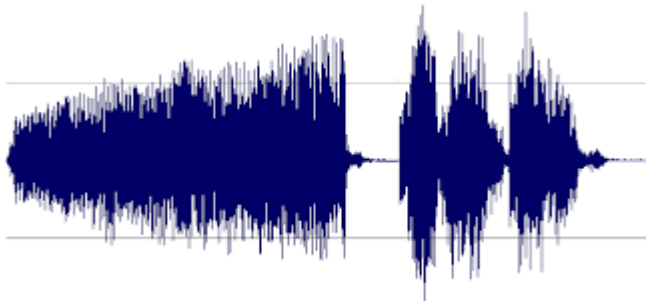

Linguistic features
Text
Text Normalization
Text
Linguistic Features Extraction
Linguistic features
Duration Model
Timestamps
Acoustic model
Acoustic features
Waveform Generation
Speech

Light Vocoder

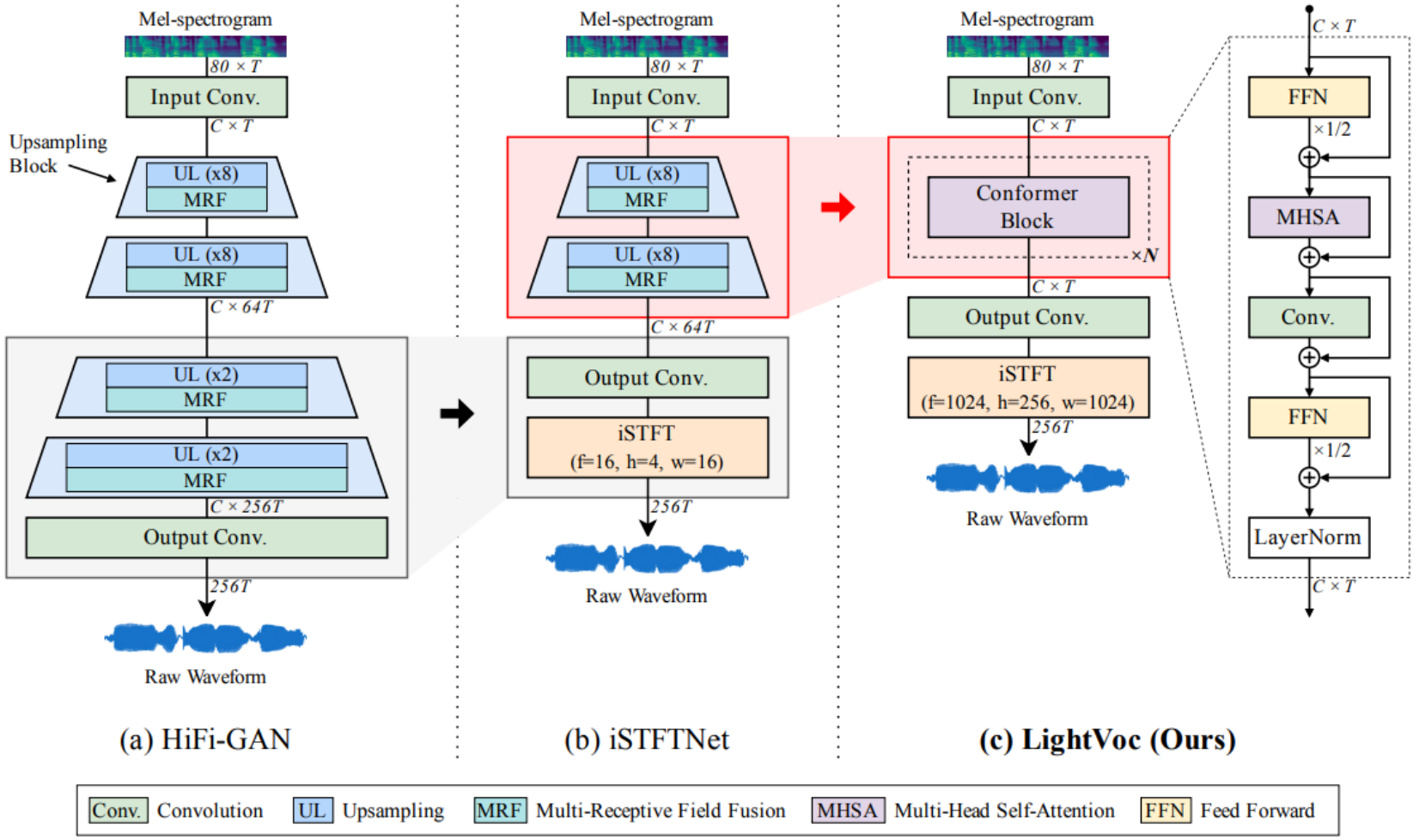
Dinh Son Dang, Tung Lam Nguyen, Bao Thang Ta, Tien Thanh Nguyen, Thi Ngoc Anh Nguyen, Dang Linh Le, Nhat Minh Le, Van Hai Do, “LightVoc: An Upsampling-Free GAN Vocoder Based On Conformer And Inverse Short-time Fourier Transform”, in Proceedings of INTERSPEECH 2023



Vocoder



LightVoc: Generator



LightVoc: Discriminator

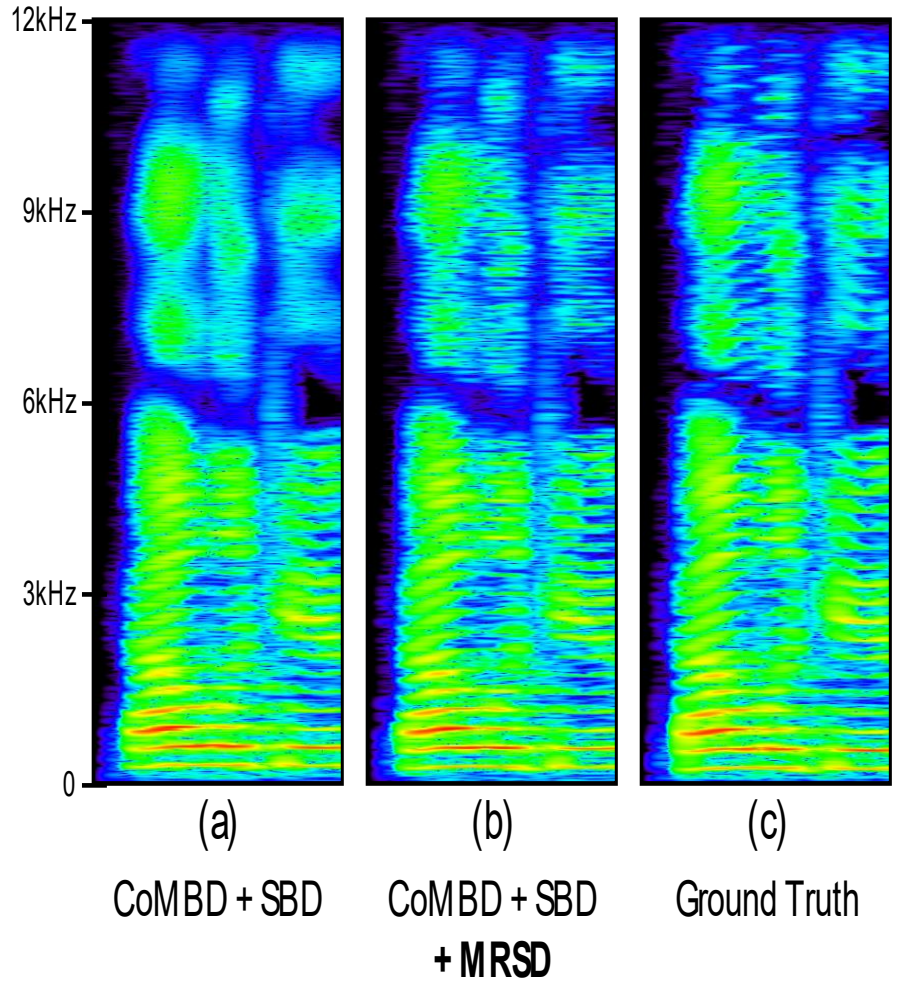
Our Discriminator:
(1) a collaborative multiband discriminator (CoMBD)
(2) a sub-band discriminator (SBD)
(3) a multi-resolution spectrogram discriminator (MRSD)

- *A combination of (1) and (2) was proposed in Avocodo* [Bak'22]

⇒ *over-smoothing problem at high-frequency band*

- *(3) was proposed in UnivNet [Jang'21] to solve over-smoothing problem*

We first propose a combination of (1,2) and (3) to generate a high-resolution waveform over the full band



Experiments

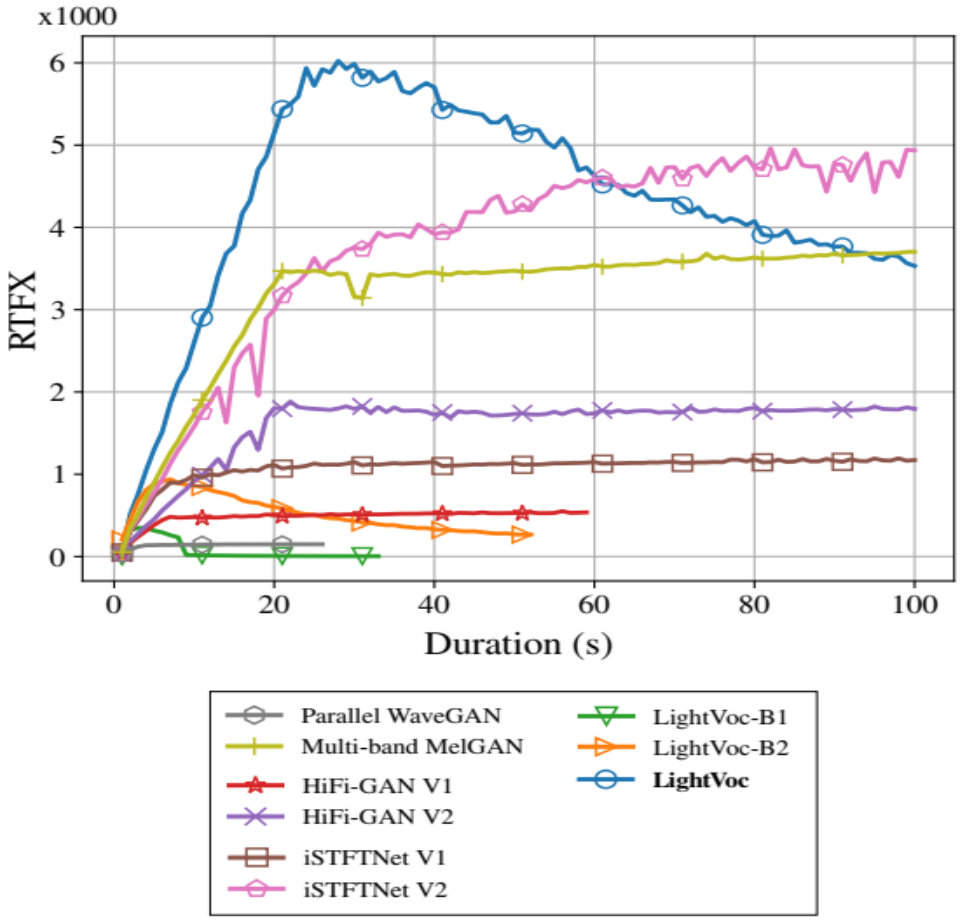


Figure 4: *Impact of audio length on GPU-based synthesizing speed. Parallel WaveGAN, HiFi-GAN V1, and LightVoc-B1/2 are unable to complete the test due to out-of-memory.*

- ✓ **52,5 times faster than HiFi-GAN V1 on CPU**
- ✓ **4 times faster than iSTFTNet V2**
- ✓ **Audio quality is competitive to HiFI-GAN V1**
- ✓ **Much smaller model size than HiFI-GAN V1 and iSTFTNet V1**

Xây dựng một hệ thống TTS trong thực tế cần làm gì?

Các vấn đề đối với TTS

Mô hình

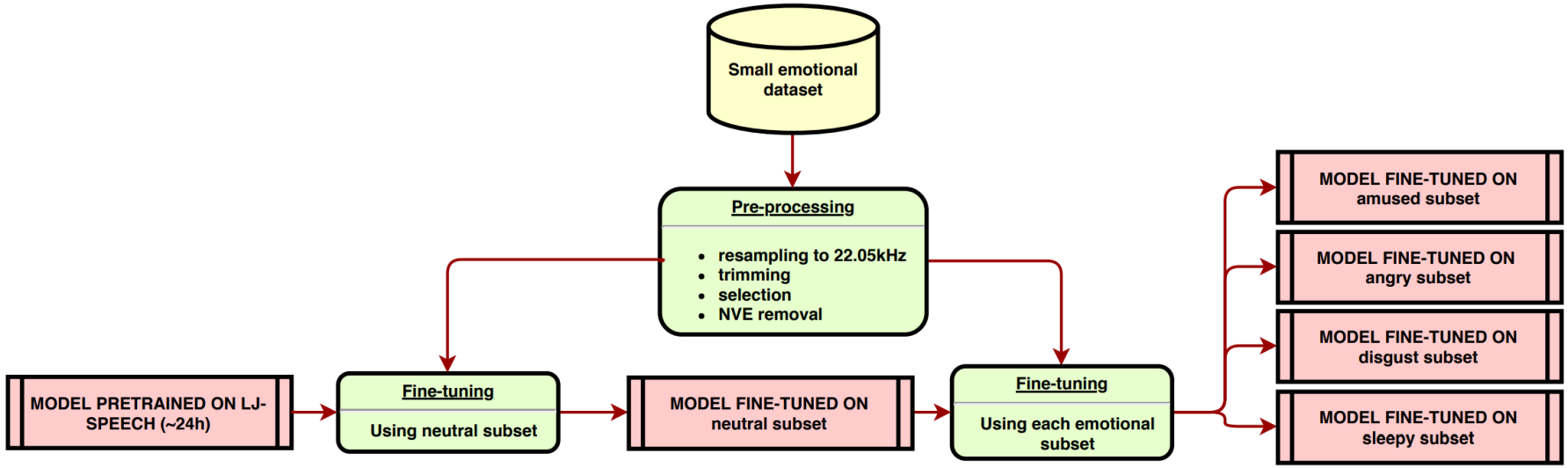
- Chuẩn bị dữ liệu huấn luyện
  - *Select 'good' speaker.*
  - *Clean audio data.*
  - *Correct transcription.*
- Lựa chọn đặc trưng, kiến trúc mô hình, kiến trúc End2End.
- Emotional TTS
- Multi-speaker TTS
- Speaker adaptation
- Voice Conversion

Các vấn đề đối với TTS

Triển khai

- Chuẩn hóa text
- Từ nước ngoài, từ viết tắt
  - AI là “ai”, hay “ây ai”
  - “ĐT: => điện thoại, đội tuyển ???
- Nâng cao hiệu năng tính toán
- Streaming

Một số vấn đề nâng cao của TTS
Emotional TTS



Tits, Noé, Kevin El Haddad, and Thierry Dutoit. "Exploring transfer learning for low resource emotional tts." *Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys).*

Một số vấn đề nâng cao của TTS

Multi-speaker TTS

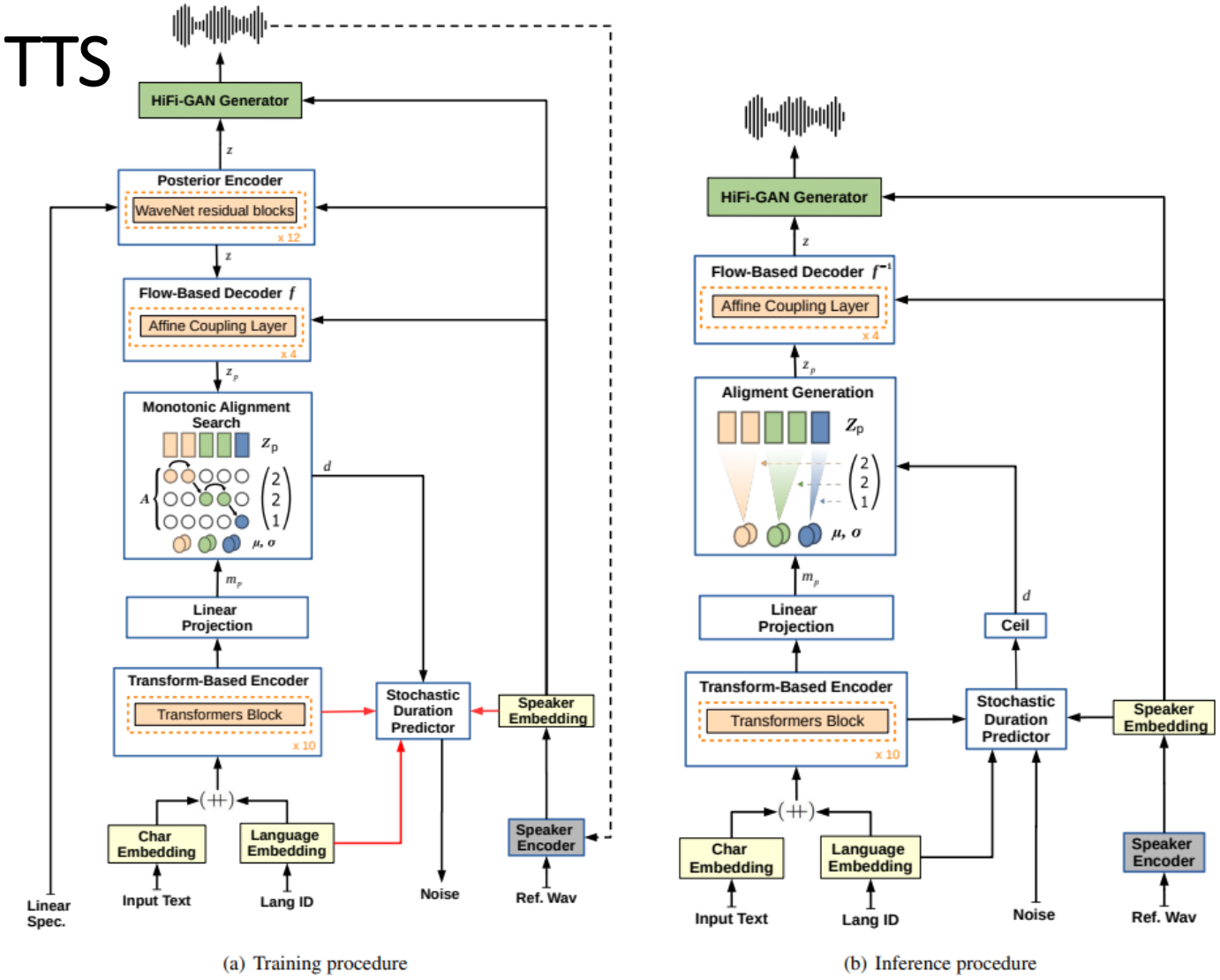
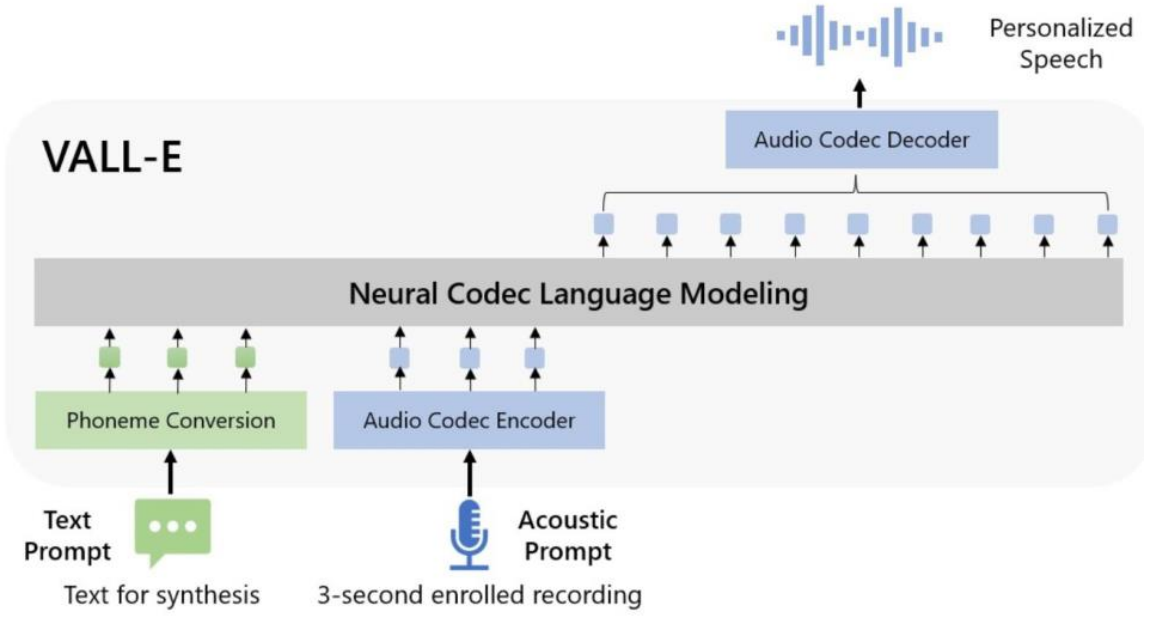


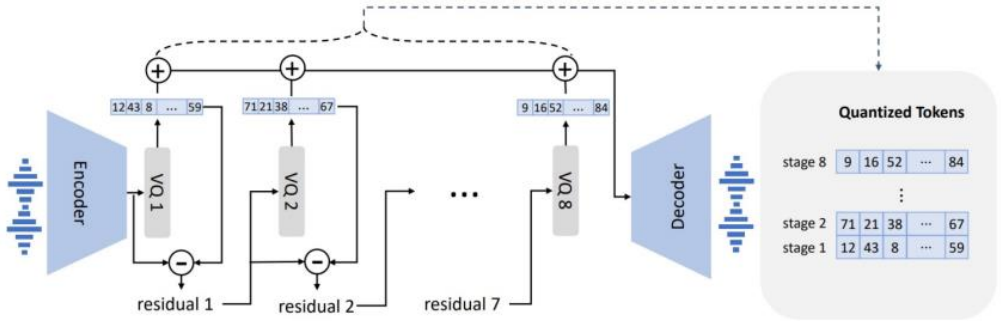
Figure 1: *YourTTS diagram depicting (a) training procedure and (b) inference procedure.*

Casanova, Edresson, et al. "YourTTS: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone." *International Conference on Machine Learning*. 2022.

Một số vấn đề nâng cao của TTS

Multi-speaker TTS





Wang, Chengyi, et al. "Neural codec language models are zero-shot text to speech synthesizers." arXiv preprint arXiv:2301.02111 (2023).

Speech to Text

(Automatic Speech Recognition – ASR)

Ứng dụng tiếng nói tiếng Việt

Virtual Assistant

Smart Home

Voice Messaging

Viettel Voice Note
Viettel Group Application Music & Audio
Everyone

Interview
Meeting Note

BIZ
Reputa
Viettel Group Application Business

Social Listening

Supervise Call Centers
Telesales

Call-Bot

Audio NewsPaper

Một ví dụ về ứng dụng của ASR

- Giám sát cuộc gọi chăm sóc khách hàng

Hiện trạng tổng đài CSKH

- CSKH là bài toán cực kì quan trọng của doanh nghiệp làm dịch vụ.
- Số lượng khách hàng càng tăng, doanh nghiệp gặp khó khăn trong việc giám sát các cuộc gọi chăm sóc khách hàng để đảm bảo chất lượng dịch vụ.
- Nếu không giải quyết sớm có thể dẫn tới các sự cố truyền thông.

Hiện trạng tổng đài CSKH

Mỗi ngày có hàng trăm nghìn cuộc gọi CSKH cần theo dõi và giám sát.

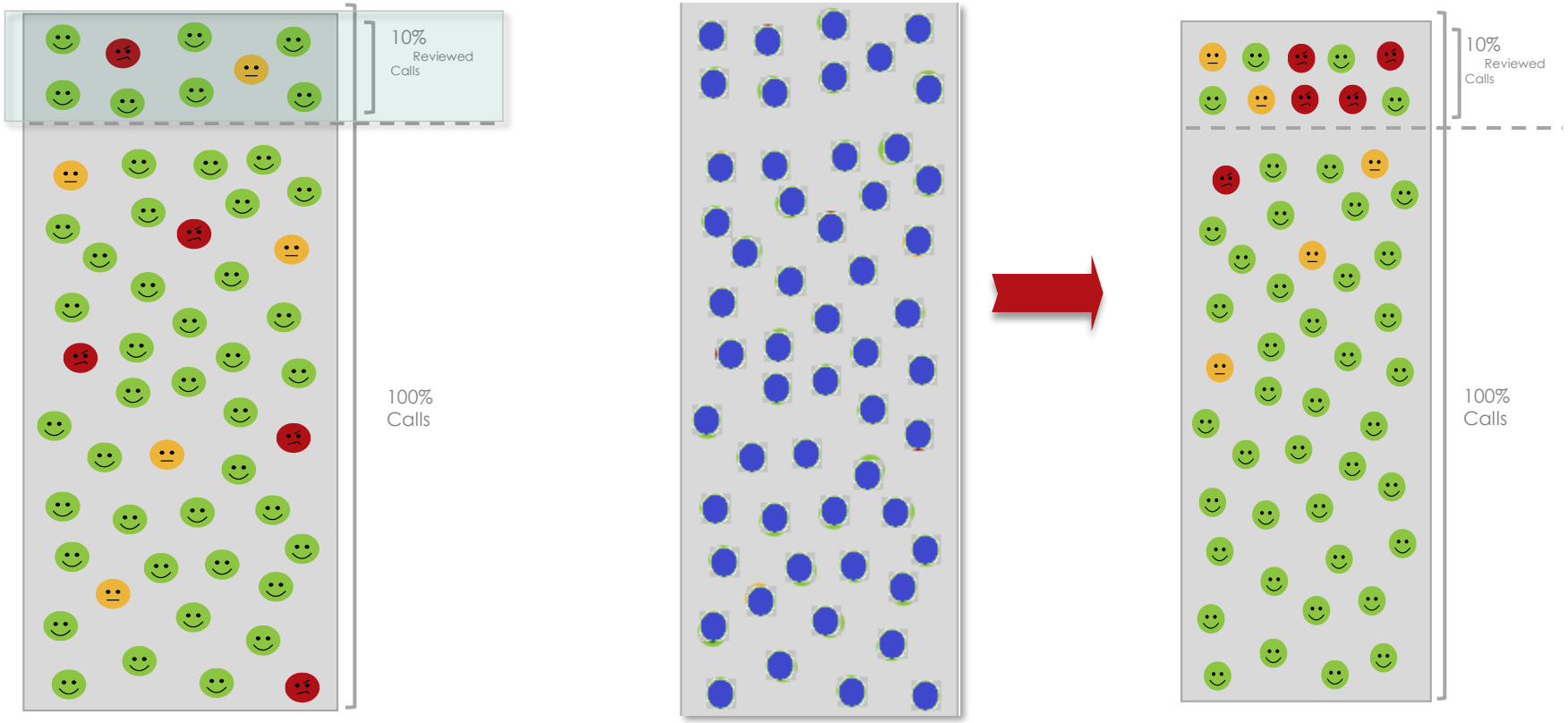
Đánh giá & Phân loại cuộc gọi chưa chính xác

- Thực hiện thủ công bằng cách nghe ngẫu nhiên với lượng mẫu nhỏ.
- Không đủ nhân sự và thời gian để giám sát toàn bộ cuộc gọi CSKH.

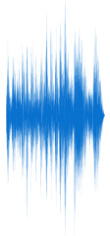
Tìm kiếm các dịch vụ bất thường gặp nhiều khó khăn

- Lựa chọn các cuộc gọi khả nghi hầu hết dựa trên cảm tính.
- Thiếu các tiêu chuẩn đánh giá khoa học. (*dựa vào các tiêu chí thô sơ như độ dài cuộc gọi, thậm chí thâm niên nhân viên CSKH*).
- Việc tìm kiếm các cuộc gọi theo nội dung hầu như không thể thực hiện được.

Hiện trạng tổng đài CSKH



Giải pháp đề xuất



Cuộc gọi CSKH

NHẬN DẠNG TIẾNG NÓI

KH: Em ơi sao điện thoại anh lại bị khóa một chiều à?
ĐTV: Dạ, do anh chưa thanh toán cước tháng này ạ.
KH: Chả có thông báo gì cả, làm thế thì không thể chấp nhận được, tôi sẽ kiện.

KHAI PHÁ DỮ LIỆU

- *Khóa một chiều*
- *Không thể chấp nhận được*
- *Kiện*
- *...*

Phân loại, chấm điểm

GIAO DIỆN WEB

- ✓ Giám sát, thống kê, phân loại
- ✓ Tìm kiếm cuộc gọi khả nghi
- ✓ Đánh giá điện thoại viên
- ✓ Xác định xu hướng khách hàng
- ✓ Tìm kiếm theo nội dung

Bài Báo:

[1] Nguyen. Q.B. et al. Development of a Vietnamese Speech Recognition System for Viettel Call Center in proceedings of Oriental COCOSDA, pp. 104-108, 2017.

[2] Nguyen. Q. B et al. Development of a Vietnamese Large Vocabulary Continuous Speech Recognition System under Noisy Conditions, in proceedings of the 9th International Symposium on Information and Communication Technology, pp. 222-226, 2018.

[3] Do. V. H et al. "Agent/Client Speech Identification for Mixed Channel Conversation in Customer Service Call Centers", accepted by the International Conference on Asian Language Processing (IALP), 2020.

So sánh với Google - độ chính xác

Do được tối ưu hóa cho bài toán cuộc gọi CSKH viễn thông nên:

- Độ chính xác về nhận dạng từ đạt 85%, từ khóa đạt 91%.
- Độ chính xác về nhận dạng ngữ điệu đạt trên 80%.

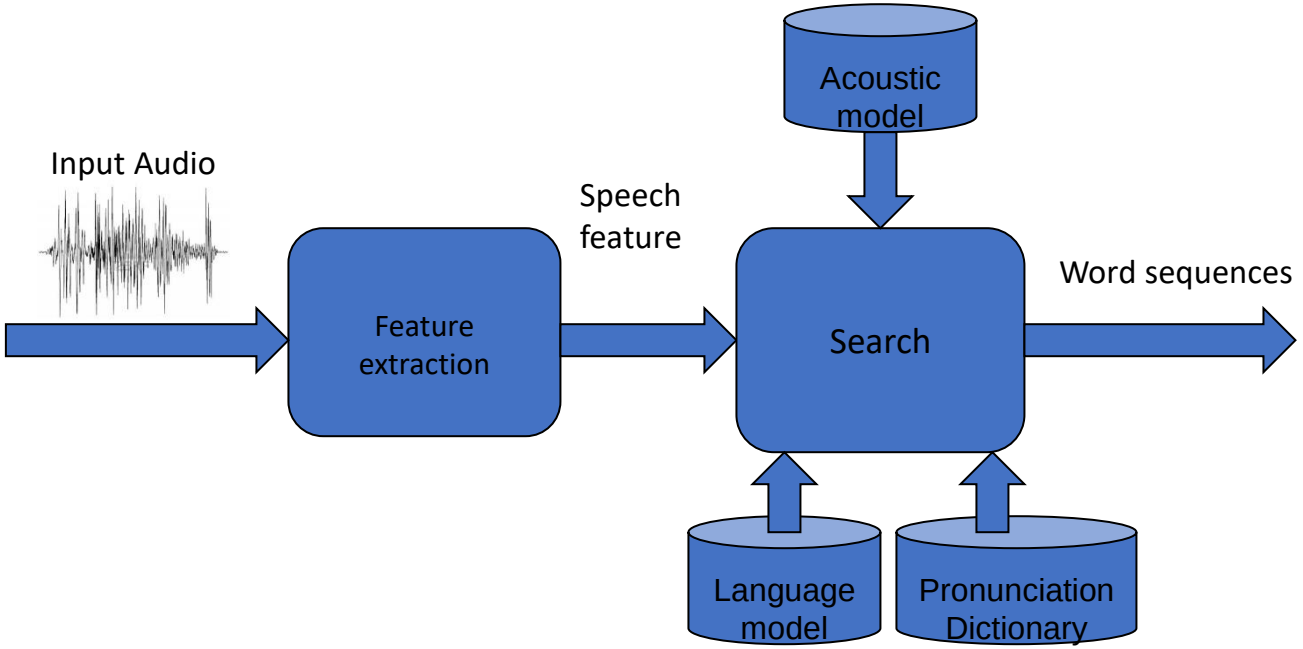
	Công nghệ đề xuất	Google API
Nhận dạng từ chung	**85%**	63%
Nhận dạng từ khóa	**91%**	82%

VIETTEL	mà Youtube mà nói chậm thì chắc chắn là có vấn đề phải không anh
GOOGLE	*phải kiến thiết mà làm gì chắc chắn là có vấn đề gì không anh*

VIETTEL	bình thường có hai cái đèn ADSL và đèn Internet nếu mà nó sáng ở trạng thái mổ sáng nháy ý
GOOGLE	*bình thường có hai cái đấy tí teo và để Internet nếu mà nó sáng ở trạng thái mà sáng nhái*

VIETTEL	này nó gắn anh có cái anten một hai đâu
GOOGLE	tại con rắn ăn con cái ăn kem một có hay đâu

Hệ thống ASR điển hình



$$\mathbf{W}^* = \arg\max_{W} \Pr(\mathbf{W} \mid \mathbf{O})$$

$$= \arg\max_{W} \Pr(\mathbf{O} \mid \mathbf{W}) \Pr(\mathbf{W})$$

Language Model

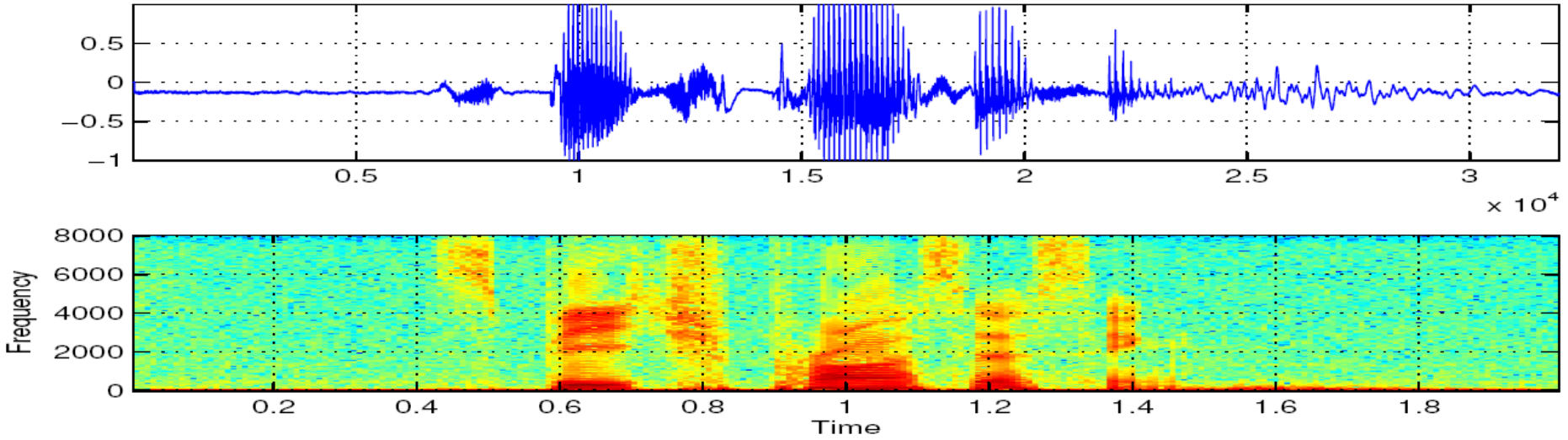
Acoustic Model

O - chuỗi đặc trưng quan sát đầu vào

W* - chuỗi text nhận dạng ở đầu ra

Trích chọn đặc trưng - Feature extraction

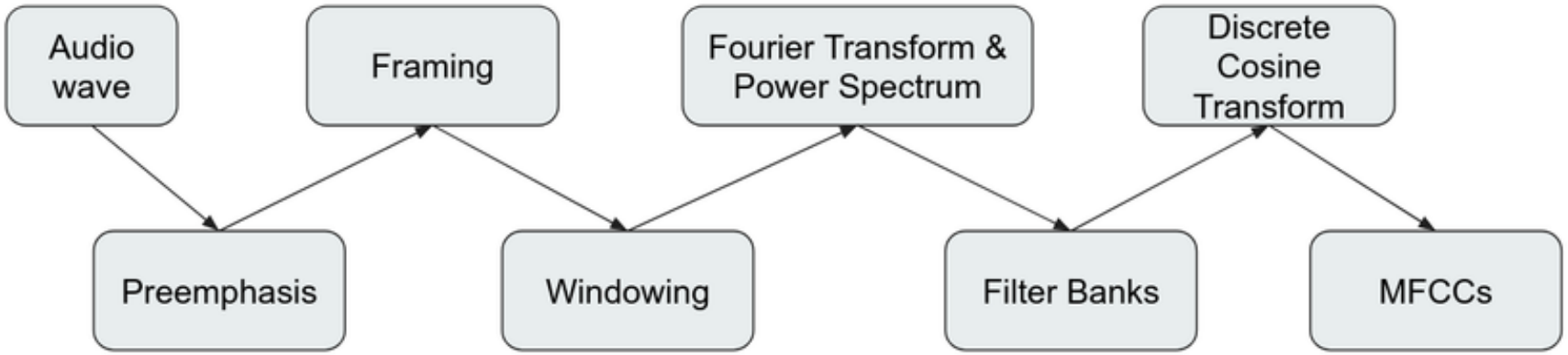
- Module này có nhiệm vụ trích xuất ra những thông tin có ích và loạ

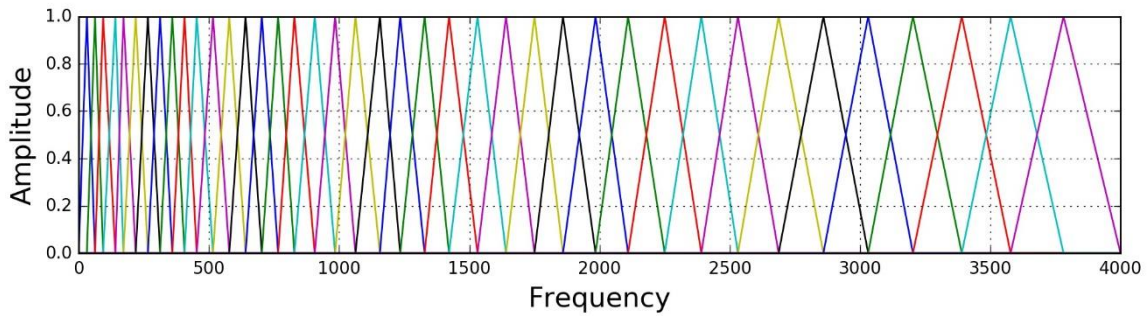


Spectrogram (Short Time Fourier Transform)

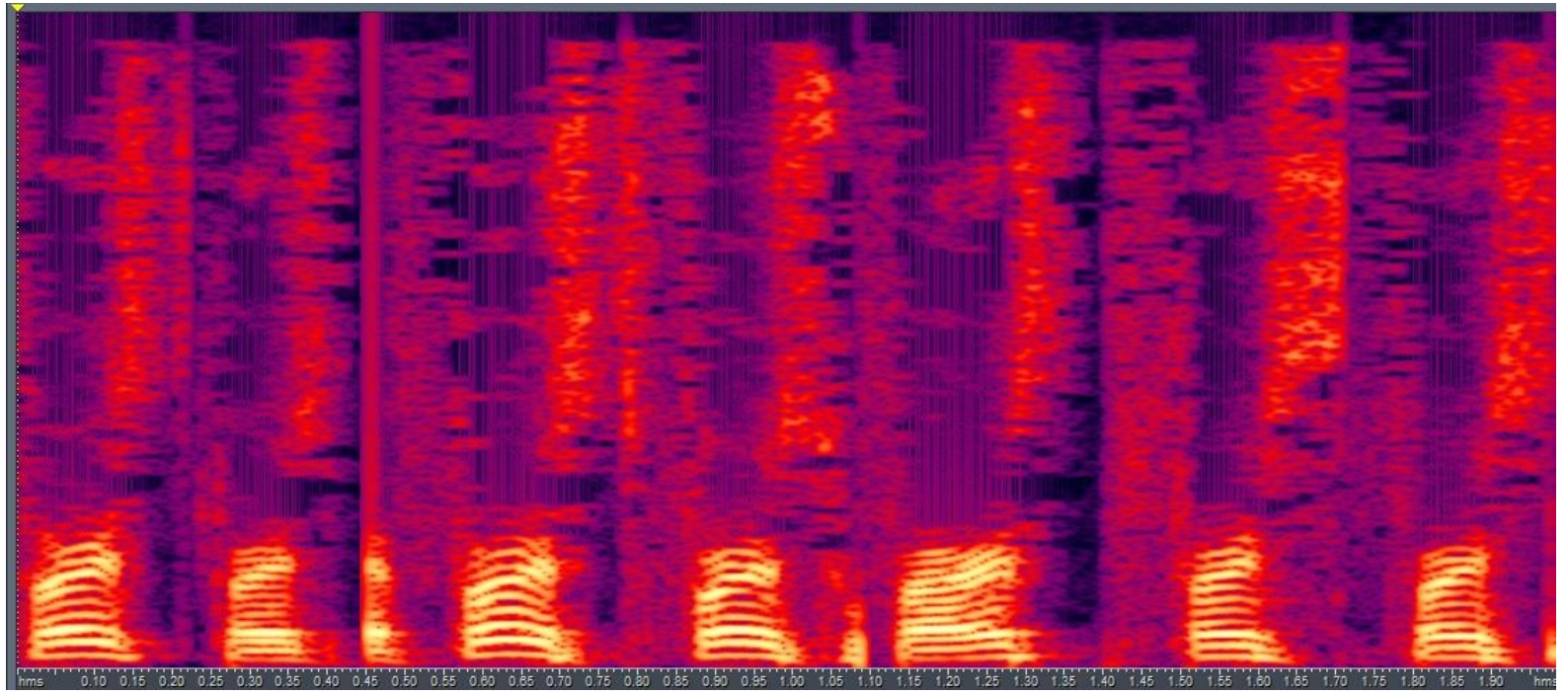
10ms -> 1 vector
1 giờ -> 360k vector
10.000 giờ -> 3,6 tỷ vector

Trích chọn đặc trưng - MFCC



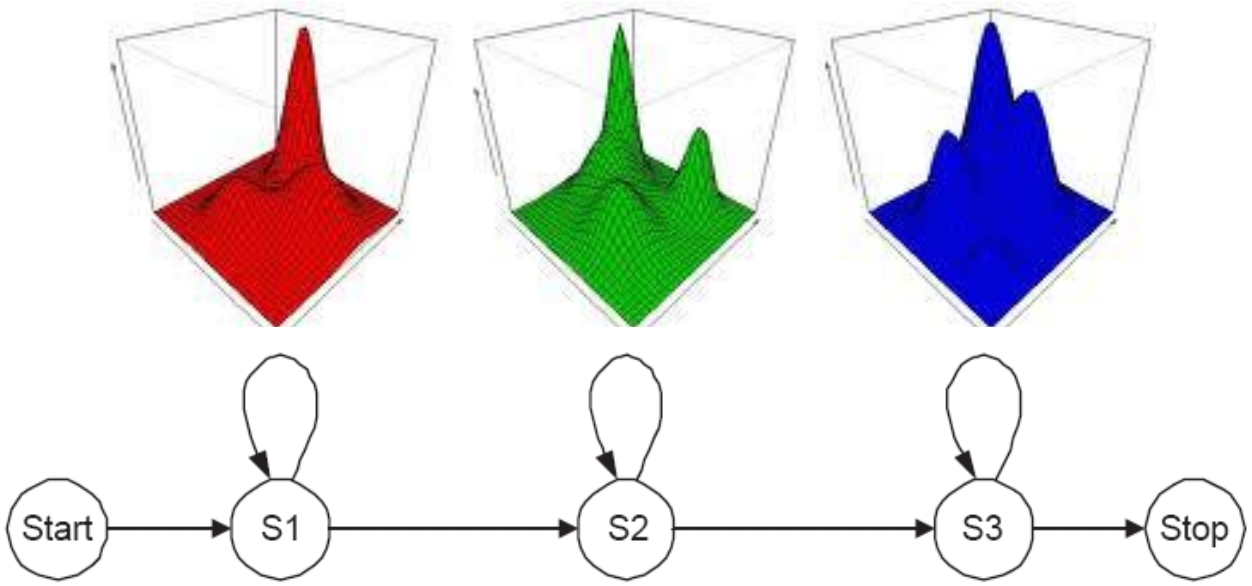


Tại sao ASR lại khó?



Spectrogram

Mô hình âm học



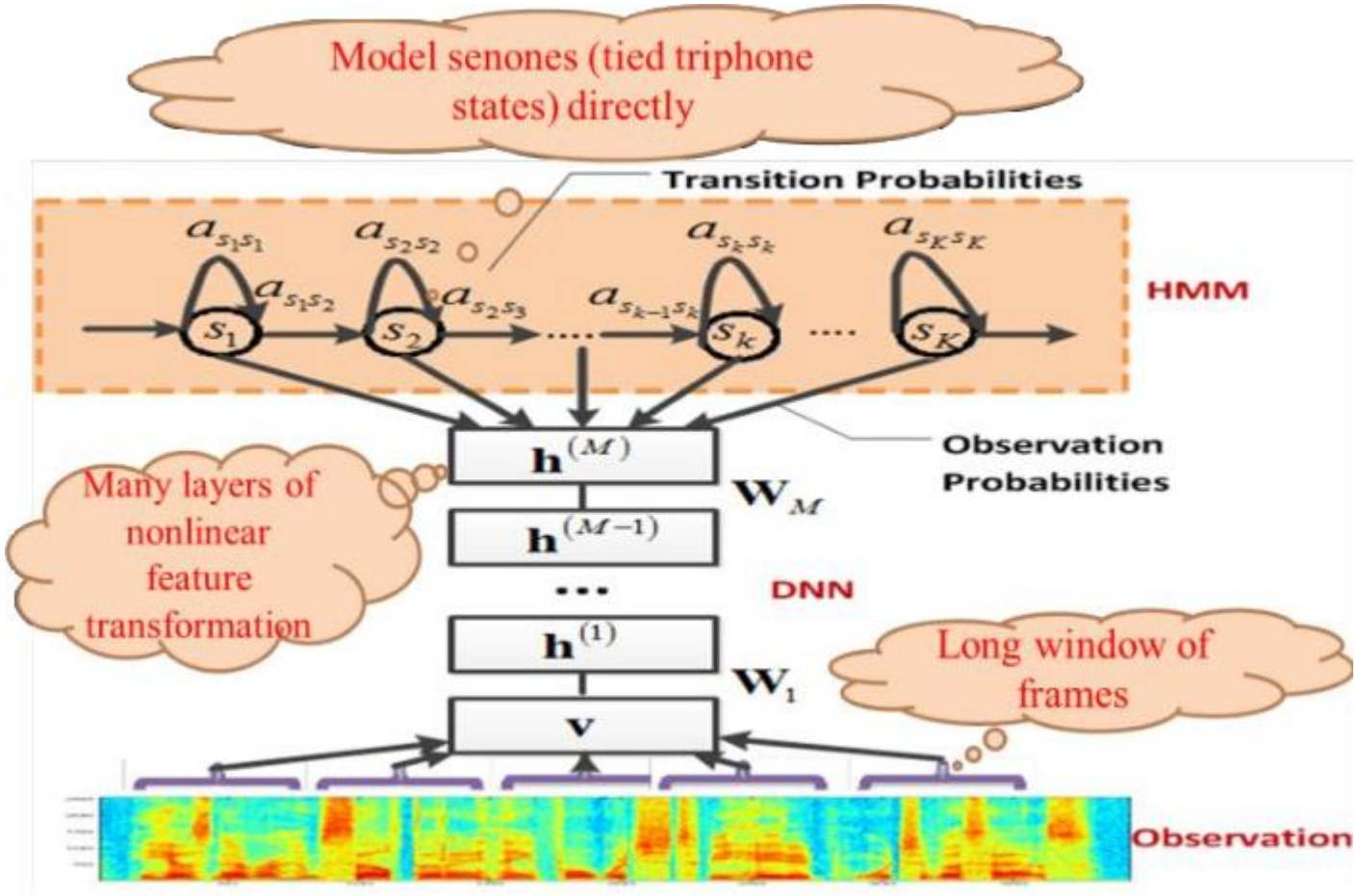
Hidden Markov Model
(HMM)

Mô hình thống kê
(GMM)

Học sâu
(DNN)

Nó hoạt động giống như tai người

Mô hình âm học sử dụng Deep Neural Networks



Mô hình âm học

- Xây dựng mô hình ổn định trong môi trường nhiễu, nhiều tạp âm.
- Mô hình hóa được giọng vùng miền khác nhau.
- Mô hình hóa được các kênh truyền khác nhau.
- Mô hình hóa được cách nói khác nhau
- Mô hình hóa được người nói khác

Mô hình ngôn ngữ

"He likes ice cold drinks" v.s "He likes eyes cold drinks"

⇒Mô hình âm học là không đủ

⇒Mô hình ngôn ngữ mô tả não người

- $P(w_n | w_1;w_2; \dots ;w_{n-1})$ => n-gram language model

- 4-gram LM (n=4)
  - Tập đoàn công => next word ?
  - Tập đoàn công nghiệp = 0.7
  - Tập đoàn công nghệ = 0.28
  - Tập đoàn công <others> = 0.02

Mô hình ngôn ngữ

- Thay đổi theo domain
- Thay đổi theo thời gian
- Thay đổi theo người nói
- Thay đổi theo vùng miền

Đánh giá trong nhận dạng tiếng nói

- Tỷ lệ sai số từ (Word Error Rate)
  - https://en.wikipedia.org/wiki/Word_error_rate
  - So sánh chuỗi chuẩn (reference) và chuỗi nhận dạng (hypothesis)

- WER=(S+I+D)/N
- Trong đó:
  - S là số từ bị thay thế.
  - D là số từ bị mất.
  - I là số từ bị chèn thêm.
  - N là số lượng từ trong chuỗi từ tham chiếu.

Ref: xin chào các bạn sinh viên tài năng

Hyp: xin cháo bạn sinh viên tài năng cực siêu nhé

WER=?

WER Benchmark

English Tasks	WER%
LibriSpeech audiobooks 960hour clean	1.4
LibriSpeech audiobooks 960hour other	2.6
Switchboard telephone conversations between strangers	5.8
CALLHOME telephone conversations between family	11.0
Sociolinguistic interviews, CORAAL (AAL)	27.0
CHiMe5 dinner parties with body-worn microphones	47.9
CHiMe5 dinner parties with distant microphones	81.3
Chinese (Mandarin) Tasks	**CER%**
AISHELL-1 Mandarin read speech corpus	6.7
HKUST Mandarin Chinese telephone conversations	23.5

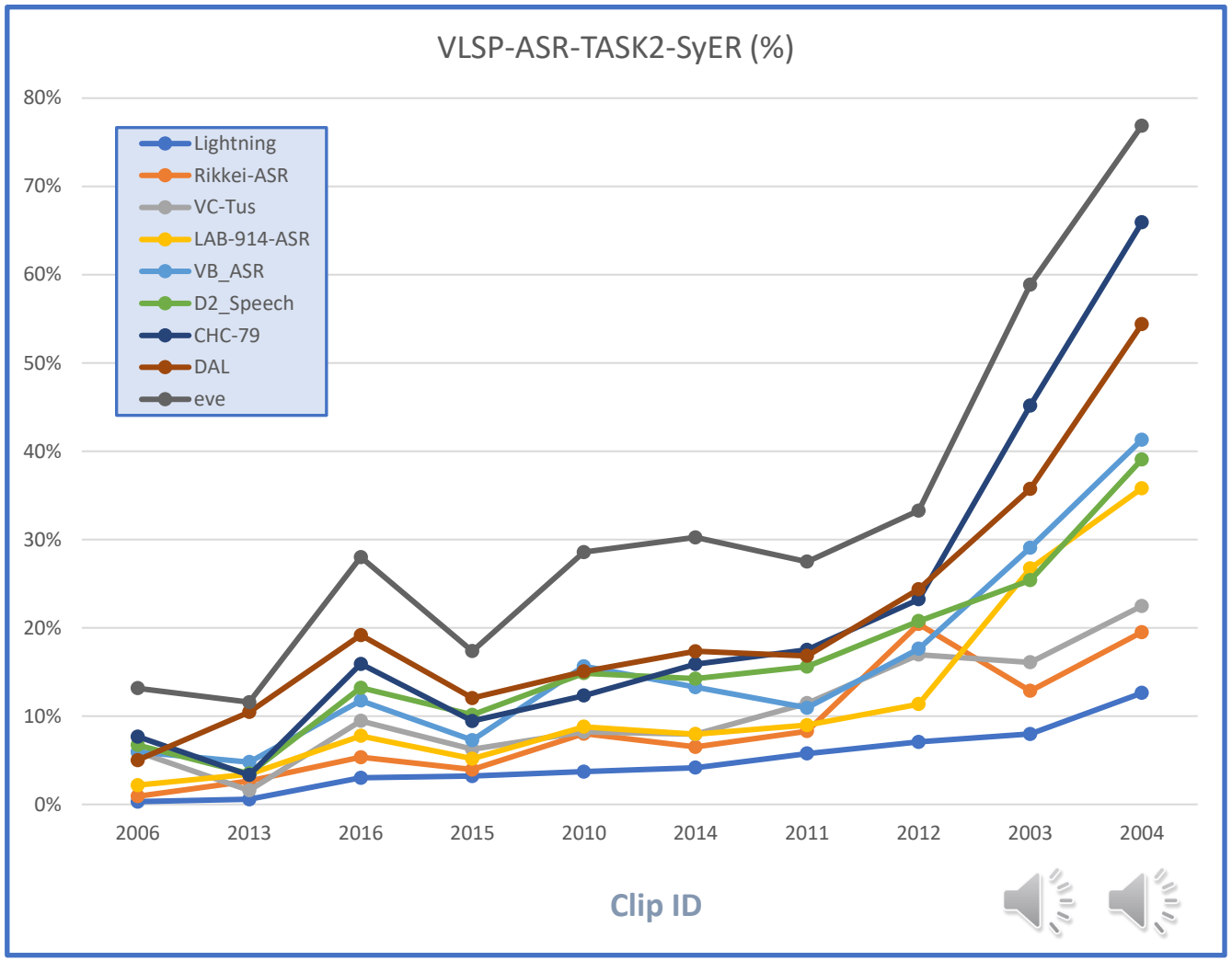
Figure 26.1 Rough Word Error Rates (WER = % of words misrecognized) reported around 2020 for ASR on various American English recognition tasks, and character error rates (CER) for two Chinese recognition tasks.

- Nguồn:
  - Daniel Jurafsky, James H. Martin, ***Speech and Language Processing***, 3rd Edition

VLSP2021 – Vietnamese ASR Challenge

Association for Vietnamese Language and Speech Processing

Rank	Task2	
	Team	SyER
1	Lightning	4.17%
2	Rikkei-ASR	6.72%
3	VC-Tus	8.83%
4	LAB-914-ASR	9.88%
5	VB_ASR	13.19%
6	D2_Speech	14.09%
7	CHC-79	18.05%
8	DAL	18.99%
9	eve	28.60%



https://vlsp.org.vn/

Thảo luận

1. Dữ liệu huấn luyện cho ASR như thế nào?
2. Cách làm dữ liệu như thế nào?
3. Tăng cường dữ liệu
4. Lựa chọn dữ liệu gán nhãn

Dữ liệu huấn luyện cho ASR như thế nào?

- Audio ⇔ text

Audio	Text
audio1	hôm nay trời đẹp lắm
audio2	chào mừng các bạn đã đến với trường hè soict
audio3	ngày mười hai tháng chín là thứ ba
...	

Cách làm dữ liệu như thế nào?

- Cho text trước rồi đọc lên
- Lấy audio có sẵn rồi gán text

Tăng cường dữ liệu (Data Augmentation)

- Thay đổi tốc độ tiếng nói
  - *Ko, Tom, et al. "Audio augmentation for speech recognition." Sixteenth Annual Conference of the International Speech Communication Association. 2015.*

    $x(t) = x(\alpha t)$

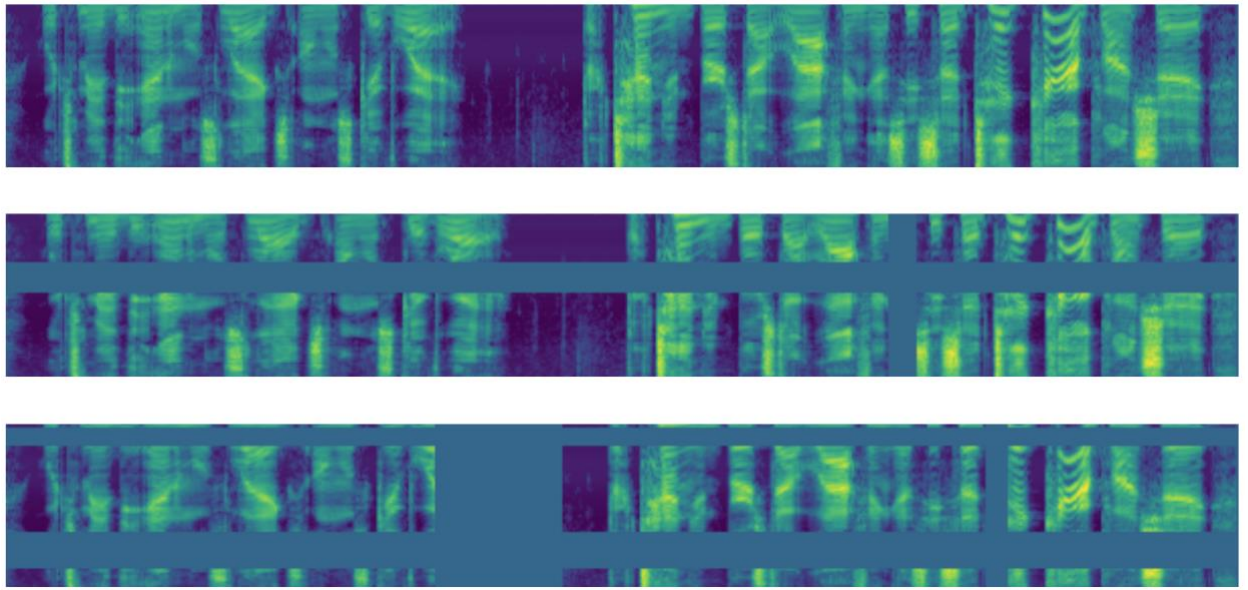
- Bổ sung thêm nhiễu và vọng

$$x_r[t] = x[t] * h_s[t] + \sum_i n_i[t] * h_i[t] + d[t]$$

  - *Ko, Tom, et al. "A study on data augmentation of reverberant speech for robust speech recognition." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.*

Tăng cường dữ liệu (Data Augmentation)

- Dùng mặt nạ che một số vùng trong phổ tín hiệu
  - Park, Daniel S., et al. "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition." Proc. Interspeech 2019 (2019): 2613-2617.



Thảo luận: Lựa chọn dữ liệu gán nhãn

Các bài toán khác trong xử lý tiếng nói

1. Text-to-Speech (Tổng hợp tiếng nói)
2. Speech-to-Text (Nhận dạng tiếng nói)
3. Speech Emotion Recognition (Nhận dạng ngữ điệu)
4. Speech Accent Recogntion (Nhận dạng phương ngữ)
5. Speech Quality Assessment (Đánh giá chất lượng tiếng nói)

Speech Emotion Recognition

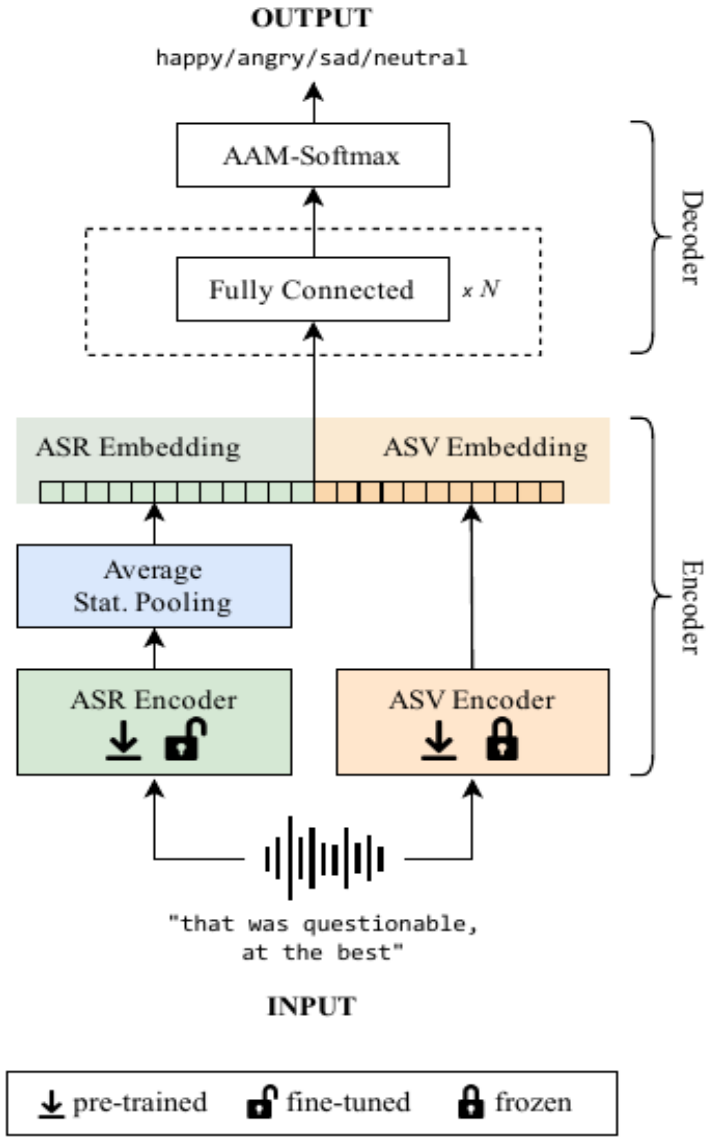


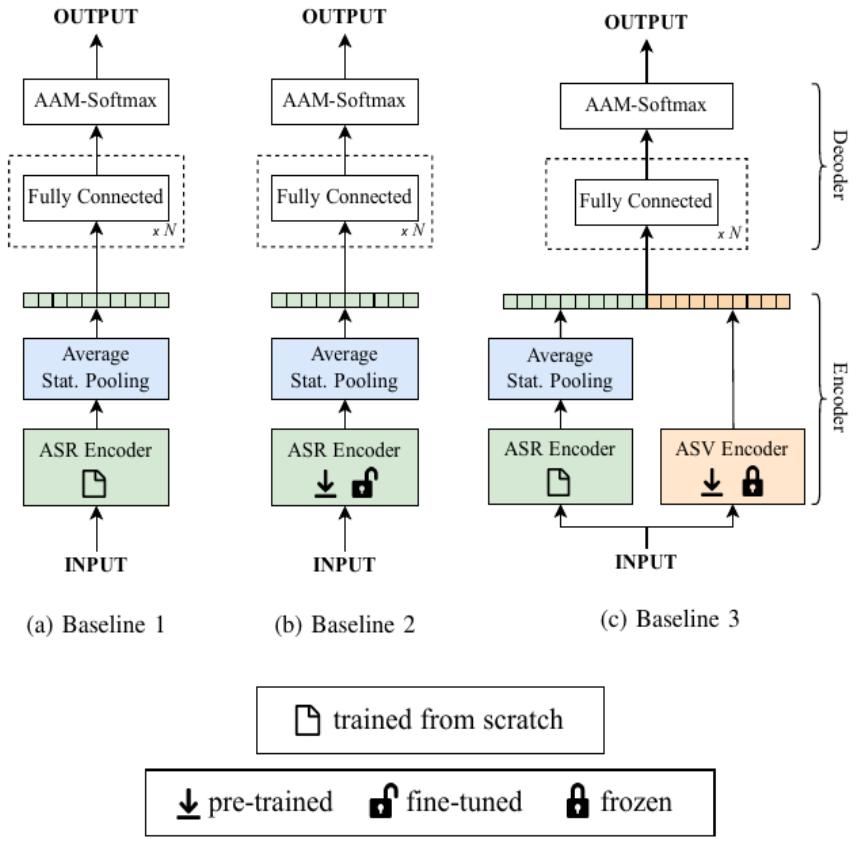
Fig. 3: An illustration of the proposed SER model. Abbreviation: *stat.* - *statistical*.

Bao Thang Ta, Tung Lam Nguyen, Dinh Son Dang, Nhat Minh Le, Van Hai Do, Improving Speech Emotion Recognition via Fine-tuning ASR with Speaker Information, in APSIPA 2022

Speech Emotion Recognition

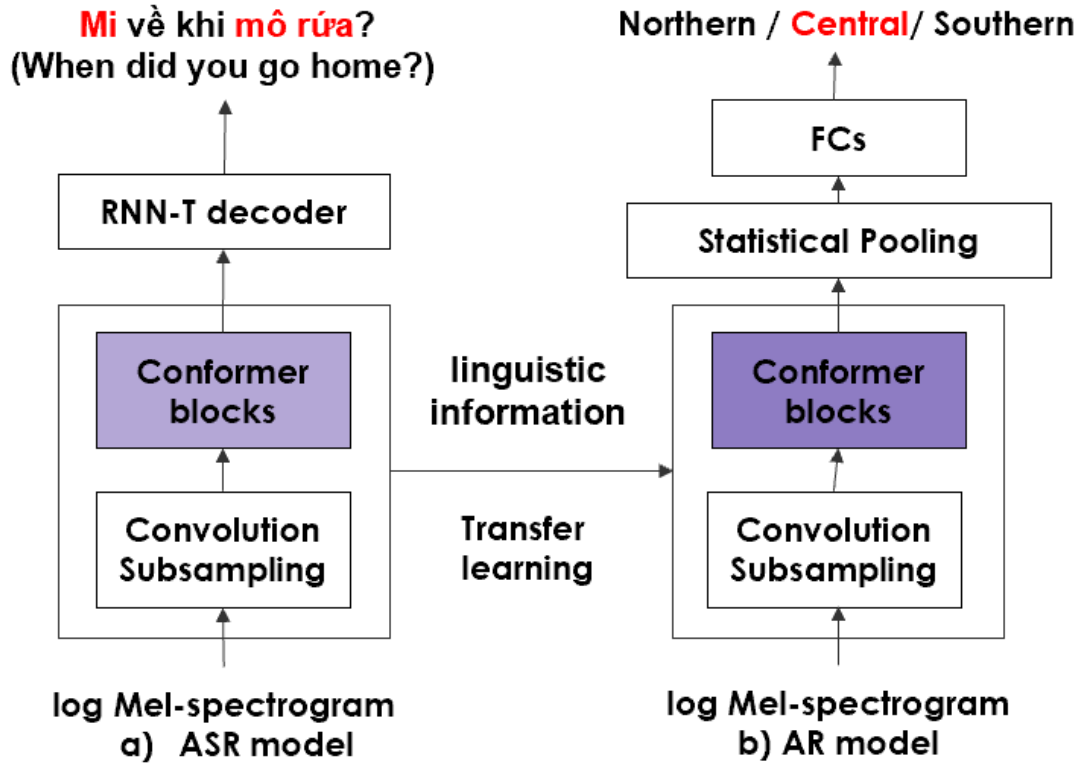
TABLE III
PERFORMANCE COMPARISON OF OUR MODELS WITH PRIOR WORKS IN UNWEIGHTED ACCURACY (UA) (%) ON IEMOCAP

Model	# Params	UA (%)
Deep Belief Network [5]	-	60.0
SpecAug + ResNet34 [12]	65M+	64.1
Wav2vec2-PT [20]	95M+	67.2
Wav2vec2-Vanilla Fine-tuning [21]	95M+	69.9
TDNN-iVector [41]	9M	71.7
Baseline 1 (Ours)	16M	71.8
Baseline 3 (Ours)	32M	72.0
MGF-fine-tune [19]	-	73.1
Wav2vec2-TAPT [21]	95M+	73.5
Baseline 2 (Ours)	16M	73.7
SYSCOMB: BLSTMATT + CSA [42]	-	74.0
HuBERT-Large-TAPT [43]	316M+	74.2
Wav2vec2-PTAPT [21]	95M+	74.3
CAP12 (Conformer based) [37]	600M+	75.0
Proposed Model (Ours)	32M	**75.3**



Speech Accent Recognition

Bao Thang Ta, Tung Lam Nguyen, Dinh Son Dang, Nhat Minh Le, Van Hai Do, Improving Vietnamese Accent Recognition via ASR Transfer Learning, in O-COCOSDA 2022



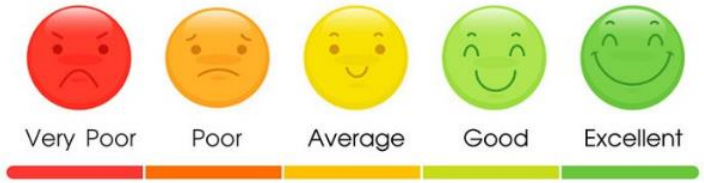
Speech Accent Recognition

Table 2: Compare with state-of-the-art Vietnamese AR model.

Model	Accuracy
CNN [12]	50.1
ResNet50[12]	60.2
LSTM (transcript only)	60.9
Conformer	62.0
Conformer (Vietnamese ASR pretraining)	**88.3**

Speech Quality Assessment

SQA models

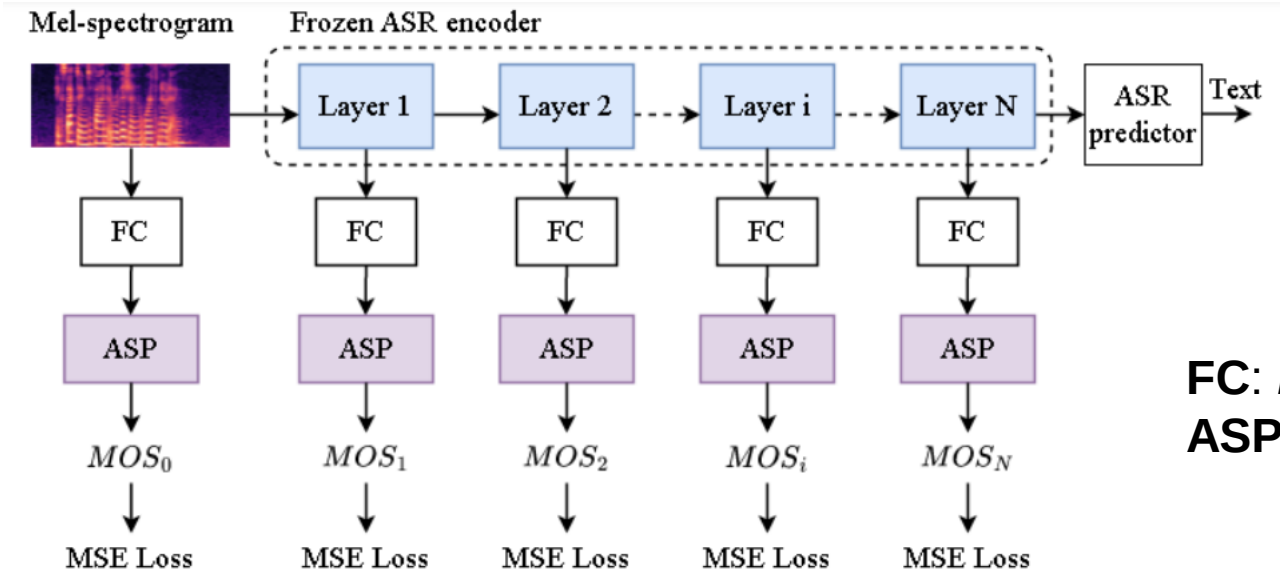


Bao Thang Ta, Minh Tu Le, Nhat Minh Le, Van Hai Do, "Probing Speech Quality Information in ASR Systems", in INTERSPEECH 2023

Probing Speech Quality Information in ASR

The probing experiments are conducted as follows:

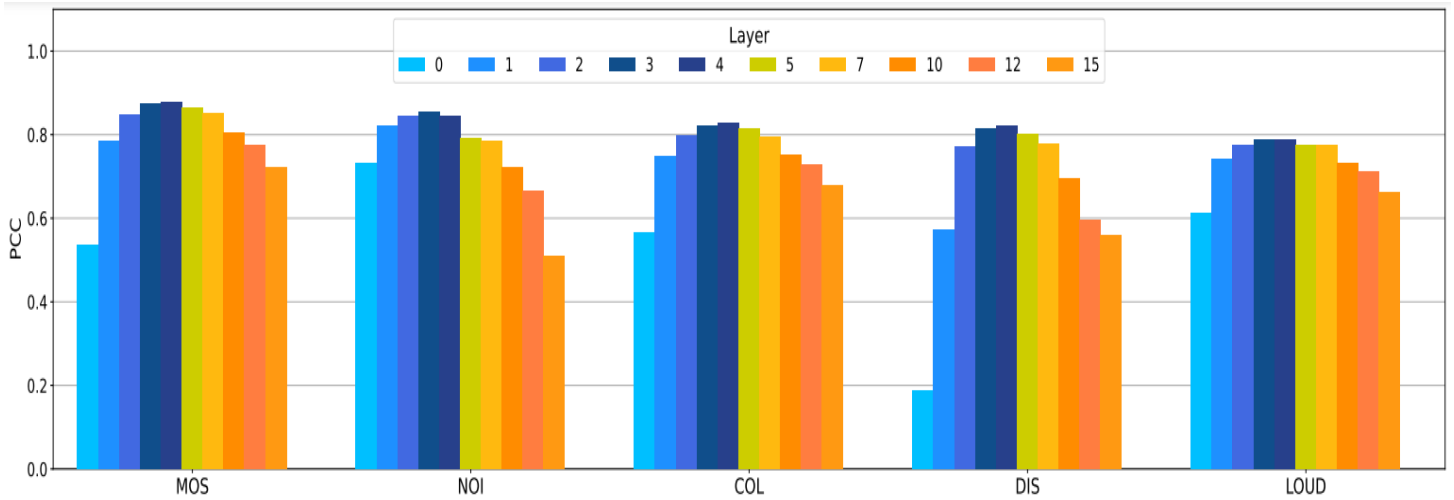
1. Build an end2end ASR based on Conformer, a recent SoTA for ASR tasks
2. Freeze all its encoder layers and utilize them as a feature extractor for SQA tasks.
3. Embedding gained from each layer goes through FC and ASP to get the predicted quality value



FC: *Fully Connected Layer*
ASP: *Attentive Statistical Pooling*

Probing Speech Quality Information in ASR

Probing experiments are conducted on four datasets sourced from the NISQA corpus [Mittag ’21]



PCC: Averaged Pearson Correlation Coefficient between predicted and subjective speech quality.
(Higher is Better)

MOS - *Mean Opinion Score*, **NOI** - *Noisiness*, **COL** - *Coloration*, **DIS** - *Discontinuity*, and **LOUD** - *Loudness*

✓ Information from ASR layers is much richer than using Mel-spectrogram (layer 0)

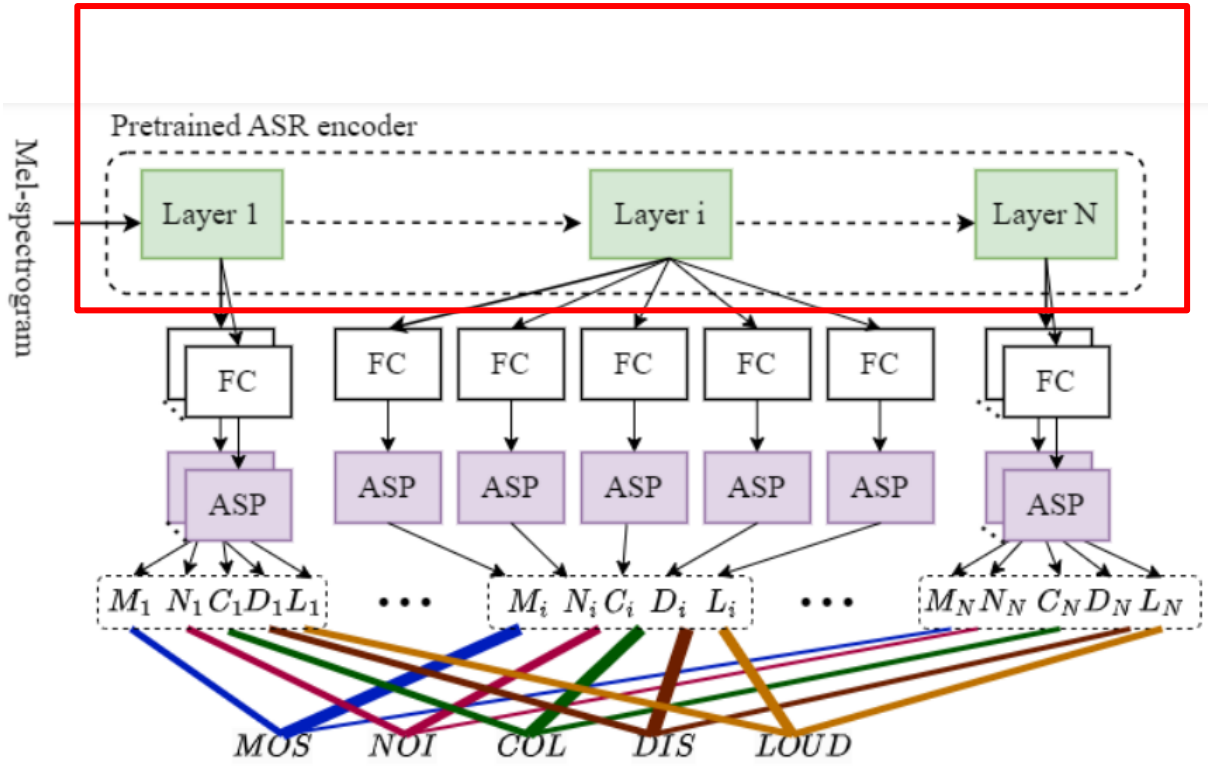
✓ Last layers fading information more than the earlier layers

ASR can be alternative approach for transfer learning for SQA tasks

How to handle different information between layers?

The proposed SQA

- Pretraining SQA encoder in ASR tasks
- Finetune directly instead of freezing to reduce negative from ASR tasks
- Combine information from different layers via Attention
- Predict multiple quality indicators simultaneously
- Employ an uncertainty loss [Kendall'18] to weigh multiple loss functions



Experimental Results

Averaged PCC and RMSE over all speech quality indicators of compared models.

Model	NISQA_VAL_LIVE		NISQA_VAL_SIM		NISQA_TEST_FOR		NISQA_TEST_LIVETALK		NISQA_TEST_P501	
	$\overline{\text{PCC}}$ ↑	$\overline{\text{RMSE}}$ ↓	$\overline{\text{PCC}}$ ↑	$\overline{\text{RMSE}}$ ↓	$\overline{\text{PCC}}$ ↑	$\overline{\text{RMSE}}$ ↓	$\overline{\text{PCC}}$ ↑	$\overline{\text{RMSE}}$ ↓	$\overline{\text{PCC}}$ ↑	$\overline{\text{RMSE}}$ ↓
Resnet34	0.582	0.549	0.754	0.632	0.650	0.639	0.664	0.692	0.731	0.679
ECAPA-TDNN	0.574	0.570	0.753	0.640	0.705	0.602	0.654	0.671	0.742	0.723
NISQA2	0.662	0.525	0.846	0.550	0.865	0.451	0.684	0.829	0.881	0.542
wav2vec2	0.690	0.507	0.849	0.553	0.863	0.447	0.778	0.650	0.834	0.666
NISQA59	0.676	0.500	0.838	0.545	0.866	0.528	0.702	0.725	0.877	**0.409**
Conformer (from scratch)	0.644	0.496	0.824	0.542	0.809	0.517	0.693	0.733	0.823	0.627
Conformer (pretrained)	0.693	0.480	0.860	**0.492**	**0.889**	0.386	0.769	0.641	0.874	0.525
Conformer (pretrained) + Attn	**0.695**	**0.478**	**0.869**	0.496	0.888	**0.385**	**0.795**	**0.609**	**0.884**	0.543

NISQA59 is trained on 59 datasets, while other methods are trained only on two datasets .

- ✓ **SoTA results in most cases**
- ✓ **Using less training data**
- ✓ **Reasonable model size**

Giới thiệu về VTCC

viettel
cyberspace

Tài nguyên

- Data Resources
  - > 10k hours of labeled training data
  - Almost unlimited unlabeled training data
- Computing Resources
  - DGX A100, H100
  - Many other powerful GPUs such as V100
- Human Resources
  - World-class experts.
  - Senior engineers.

https://viettelgroup.ai

Lợi ích

- Competitive salary
- Join large-scaled projects
- Work with experienced experts, engineers.
- Support to write scientific papers, patents.
- Support to attend international conferences.
- Support to study Master, PhD.

Gương mặt gen Z tiêu biểu

Kết luận

Thank you!