srce
Sveučilište u Zagrebu
Sveučilišni računski centar

Korištenje različitih reprezentacija govora i ansambala modela za pretvorbu govora u tekst

Marin Jezidžić, Matej Mihelčić

Sadržaj

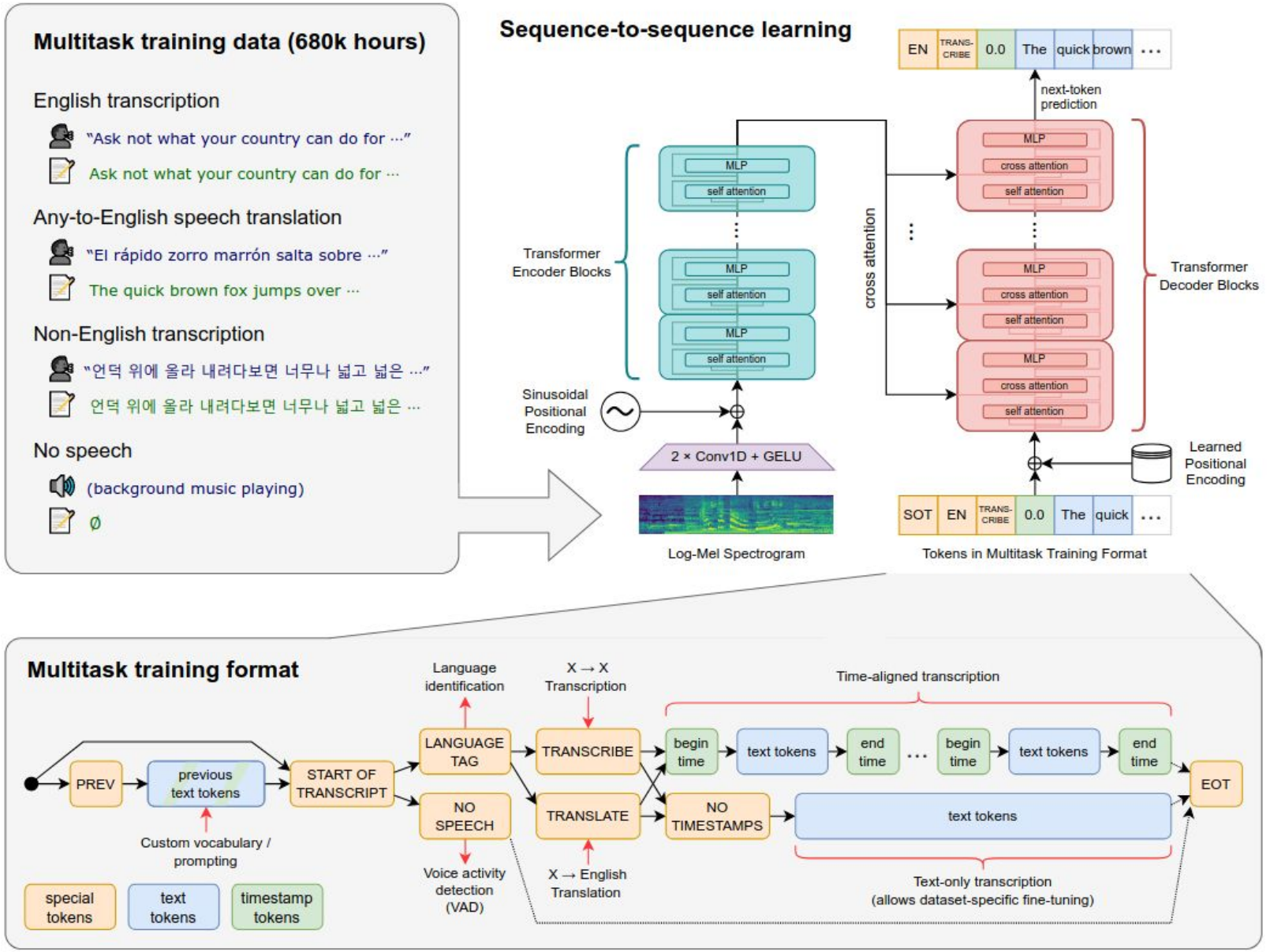
- Uvod
- Dvodimenzionalne reprezentacije zvučnog signala
- Metode kombiniranja različitih reprezentacija zvučnog signala
- Predloženi pristup

Sadržaj

- **Uvod**
- Dvodimenzionalne reprezentacije zvučnog signala
- Metode kombiniranja različitih reprezentacija zvučnog signala
- Predloženi pristup

Uvod

- Sveobuhvatni problem nazivamo *automatsko prevođenje govora*, a osnovni potproblemi su:
  - transkripcija govora
  - prepoznavanje ključne riječi
  - identifikacija govornika
- Za modeliranje imamo unikatne izazove u odnosu na NLP probleme



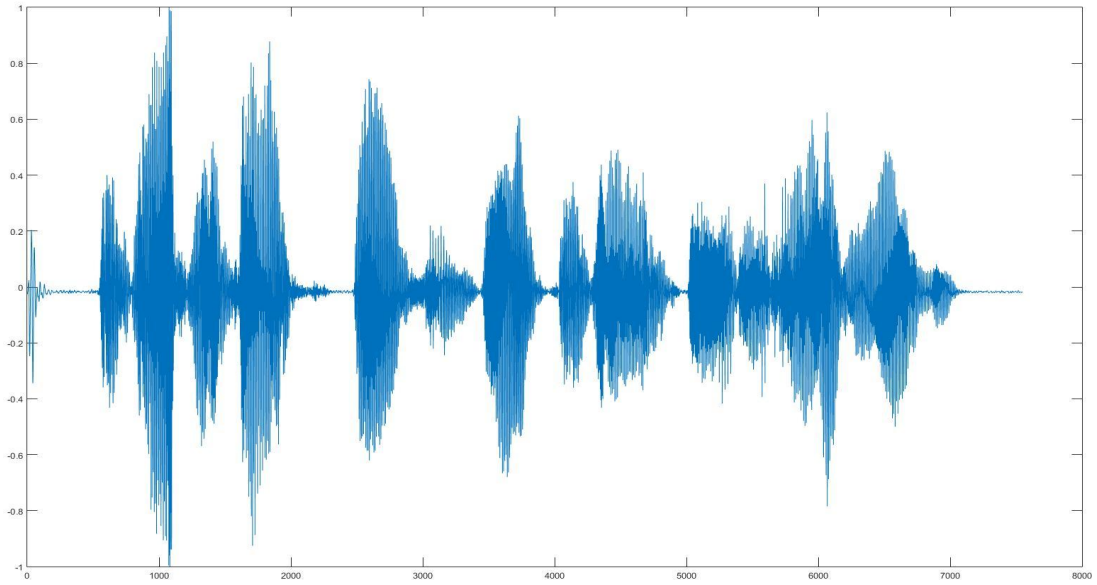
[1] Radford et al. Robust Speech Recognition via Large-Scale Weak Supervision. CoRR abs/2212.04356

Sadržaj

- Uvod
- **Dvodimenzionalne reprezentacije zvučnog signala**
- Metode kombiniranja različitih reprezentacija zvučnog signala
- Predloženi pristup

Dvodimenzionalne reprezentacije zvučnog signala

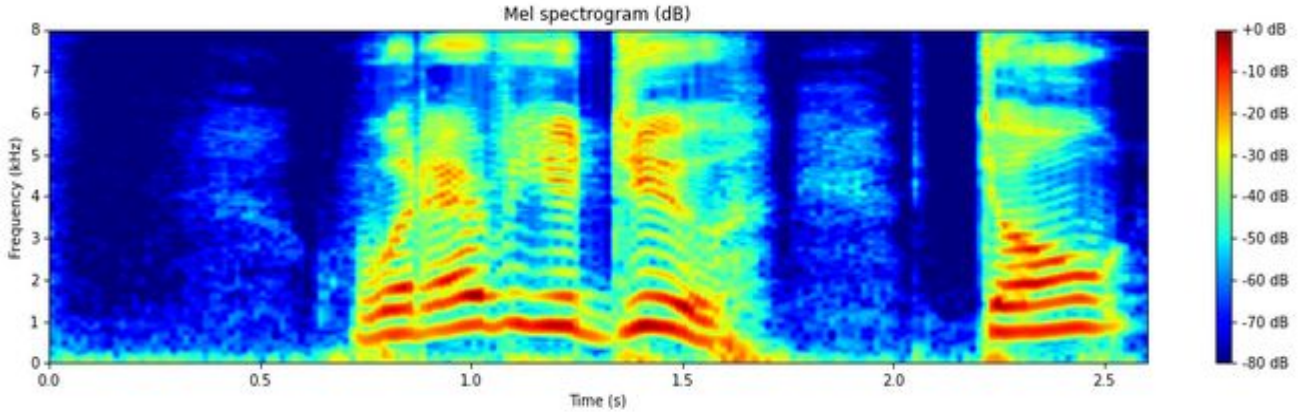
- Kako najbolje reprezentirati zvučni signal otvoreno je pitanje



?

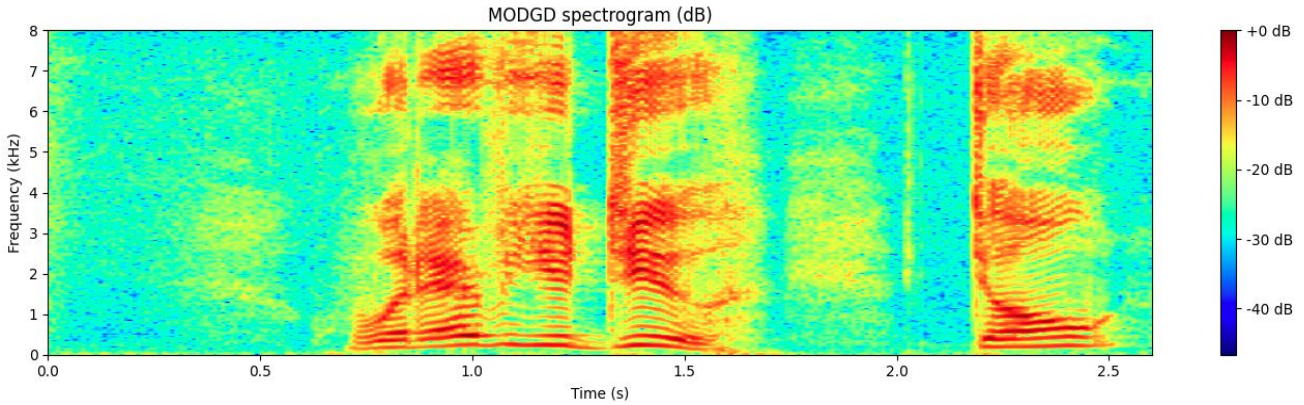
Mel Spektrogram i MFCC

- Najpoznatije reprezentacije zvuka
- Dobar balans frekvencijske i vremenske rezolucije
- Brza ekstrakcija; fiksne duljine prozora!
- MFCC reprezentacija se dobije iz Mel Spektrograma apliciranjem DCT transformacije



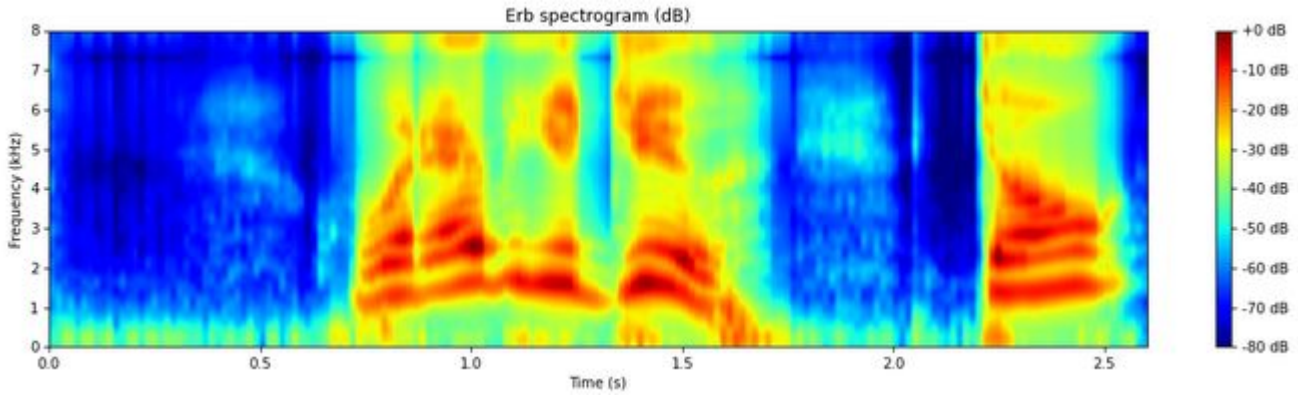
Modificirani Group Delay Spektrogram

- Dobije se iz Group Delay-a spektralnim omekšavanjem i uvođenjem dva parametra za kontroliranje dinamičkog dosega
- Za razliku od Mel Spektrograma uključuje i faznu informaciju



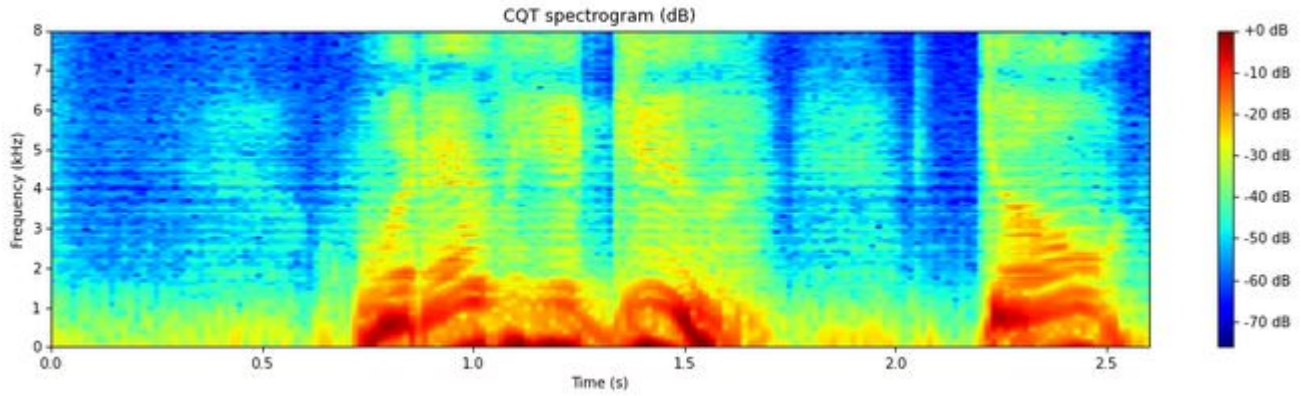
Gammatone Spektrogram

- Temelji se na gammatone filterima koji oponašaju rad pužnice (cochlea) u ljudskom uhu
- Svaki filter ima impulsni odziv sličan gama funkciji pomnožen s tonom (sinusoidom)
- Dobra otpornost na šum



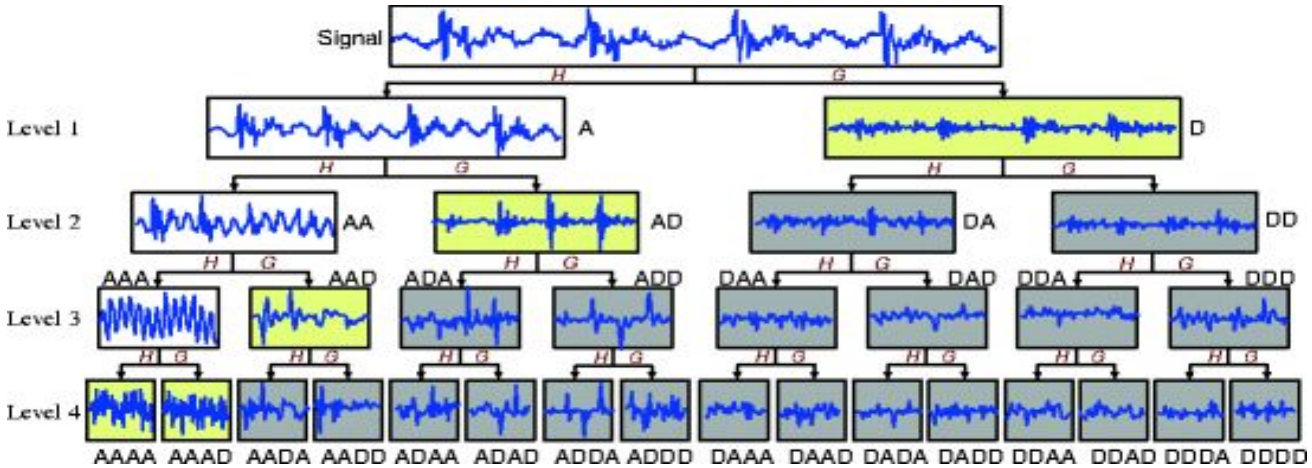
Constant Q Transformirani Spektrogram

- Duljina prozora analize mijenja se s frekvencijom: duži prozori za niže frekvencije, kraći za više
- Vrsta vremensko-frekvencijske analize gdje je omjer frekvencije i frekvencijske rezolucije (Q) konstantan kroz cijeli spektar
- Brza ekstrakcija, no neke optimizacije kao s DFT nisu moguće radi prozora različite veličine



Transformacija Paketom Valića

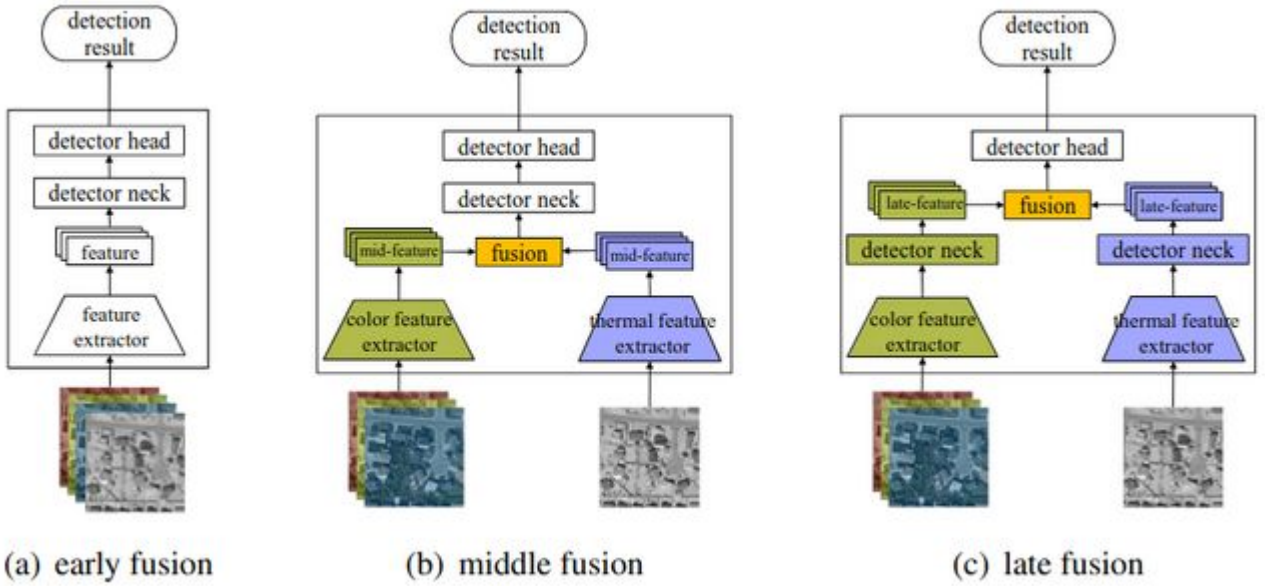
- Proširenje diskretne transformacije valićima (DWT) koje omogućuje finiju analizu signala na svim frekvencijskim razinama
- WPT dijeli i niskofrekventne i visokofrekventne dijelove signala
- Stvara potpuno binarno stablo gdje svaki čvor predstavlja različiti frekvencijski pojas signala



Sadržaj

- Uvod
- Dvodimenzionalne reprezentacije zvučnog signala
- **Metode kombiniranja različitih reprezentacija zvučnog signala**
- Predloženi pristup

Metode kombiniranja različitih reprezentacija zvučnog signala



[2] Fang et al.: Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. Pattern Recognit. 130: 108786

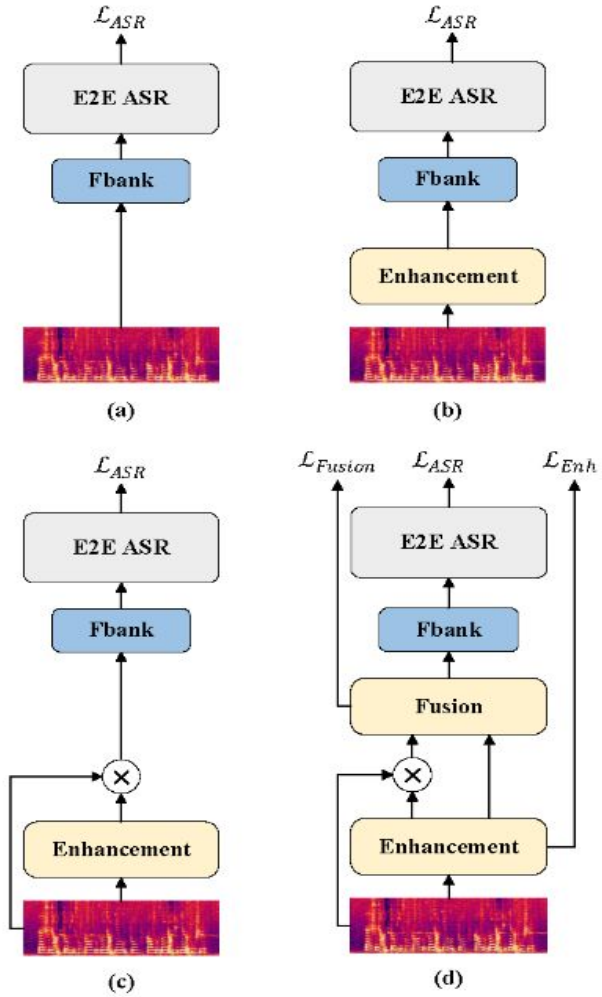
- Optimiziramo:

WER (Word error rate) $\longrightarrow$ $\mathrm{WER} = \frac{S_w + D_w + I_w}{N_w}$

CER (Character error rate) $\longrightarrow$ $\mathrm{CER} = \frac{S_c + D_c + I_c}{N_c}$

U problemu prevođenja govora

Primjer: Rana fuzija



- Rana fuzija [3]
- Troslojni BLSTM kao modul za poboljšanje govora [4]
- Transformer enkoder-dekoder kao bazni model [5]

[3] Shi et al.: Spectrograms Fusion-based End-to-end Robust Automatic Speech Recognition. APSIPA ASC 2021: 438-442
[4] Schuster et al.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 45(11): 2673-2681
[5] Vaswani et al.: Attention Is All You Need. CoRR abs/1706.03762

$$\mathcal{L}_{Joint} = \beta\mathcal{L}_{ASR} + (1-\beta)\mathcal{L}_{SE} + \gamma\mathcal{L}_{SF}$$

$$\mathcal{L}_{ASR} = -\ln P(S^*|\hat{X})$$

$$\mathcal{L}_{SE} = \alpha\mathcal{L}_{Mapping} + (1-\alpha)\mathcal{L}_{Masking}$$

$$\mathcal{L}_{SF} = \frac{1}{TF}\sum_i ||\widetilde{MDM}_i - MDM_i||_F^2$$

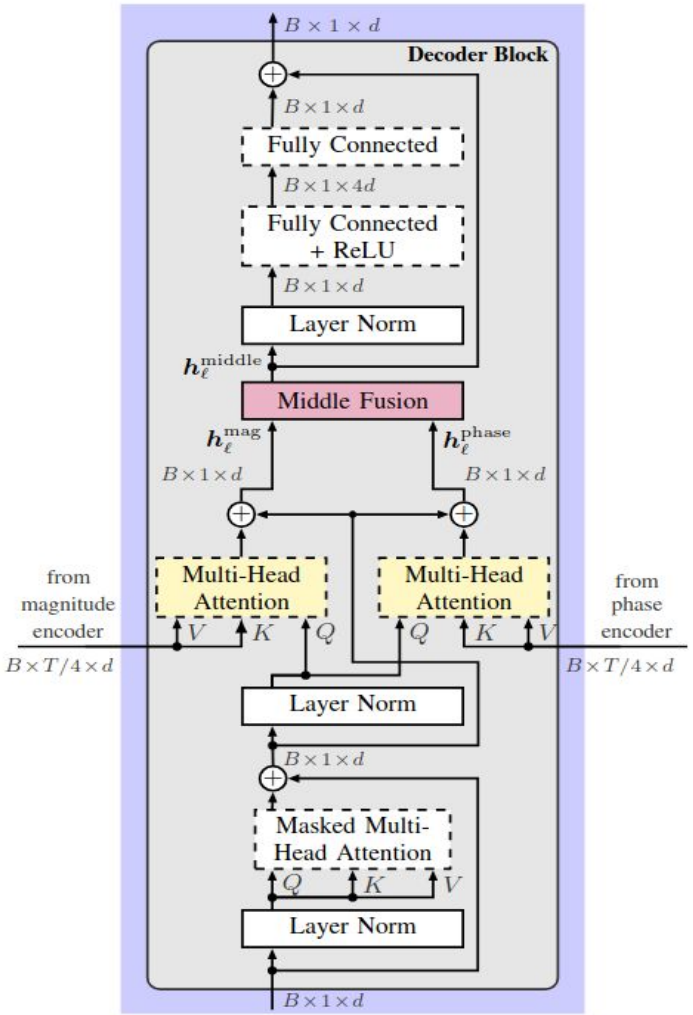
$$\hat{X} = M\ddot{D}M \odot \ddot{X} + \widetilde{MDM} \odot \widetilde{X}$$

$$\mathcal{L}_{Mapping} = \frac{1}{TF}\sum |||\widetilde{X}| - |X|||_F^2$$

$$\mathcal{L}_{Masking} = \frac{1}{TF}\sum ||\widetilde{M} \odot |Y| - |X|||_F^2 = \frac{1}{TF}\sum |||\ddot{X}| - |X|||_F^2$$

- T - time
- F - frequency
- $\widetilde{M}$ - estimated mask from speech enhancement
- $|Y|$ - noisy input magnitude spectrogram
- $|\widetilde{X}|$ - mapping based speech magnitude spectrogram
- $|\ddot{X}|$ - masking based speech magnitude spectrogram
- $|X|$ - target clean speech magnitude spectrogram
- S^* - ground truth of the whole sequence of output labels
- MDM - minimal distance mask

Primjer: Srednja fuzija



- Srednja fuzija [6]
- Predložili zajednički trening više tipova značajki uz inferenciju zasebnih

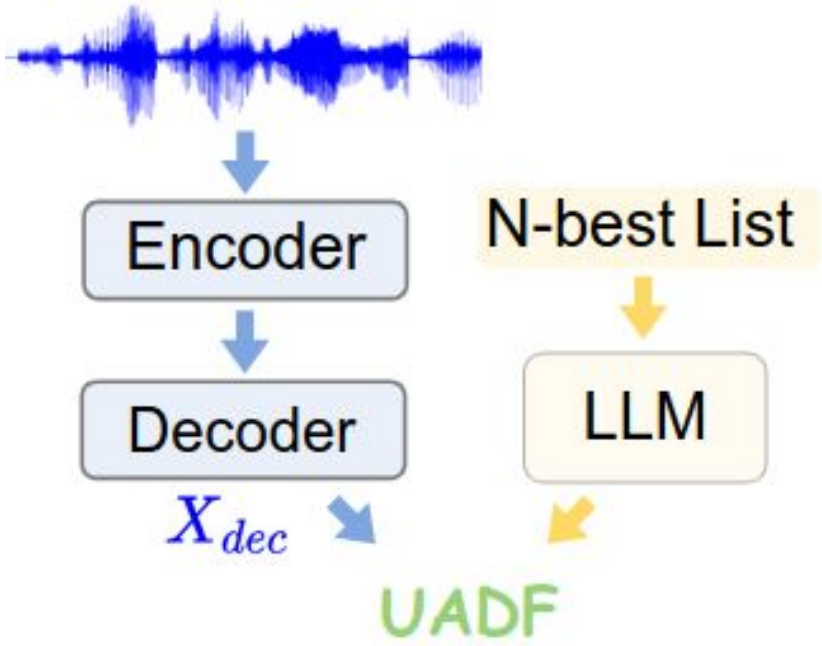
$$\mathbf{h}_\ell^{\text{middle}} = \alpha \boldsymbol{h}_\ell^{\text{mag}} + (1-\alpha)\boldsymbol{h}_\ell^{\text{phase}}$$

[6] Lohrenz et al.: Multi-Encoder Learning and Stream Fusion for Transformer-Based End-to-End Automatic Speech Recognition. CoRR abs/2104.00120

Approach	Inference complexity # of parameters	Inference complexity relative runtime	Without language model dev93 WER	Without language model dev93 CER	Without language model eval92 WER	Without language model eval92 CER	With language model dev93 WER	With language model dev93 CER	With language model eval92 WER	With language model eval92 CER
Baseline-mag	16.8M	1.0	14.62	5.28	11.66	4.03	6.51	3.64	4.43	2.37
Baseline-phase	16.8M	1.0	15.75	5.79	12.90	4.33	7.32	4.23	5.48	3.17
Fusion-Mid-WS	28.8M	1.37	16.82	5.75	13.43	4.11	6.40	3.55	4.38	2.46
Fusion-t-Mid-WS	27.2M	1.33	15.89	5.58	12.32	4.07	6.23	3.57	4.09	2.10
MEL-t-mag	16.8M	1.0	15.22	5.36	11.73	3.95	6.29	3.43	4.31	2.50
MEL-t-phase	16.8M	1.0	16.01	5.85	12.09	4.25	7.00	4.04	4.68	2.74

[6] Lohrenz et al.: Multi-Encoder Learning and Stream Fusion for Transformer-Based End-to-End Automatic Speech Recognition. CoRR abs/2104.00120

Primjer: Kasna fuzija



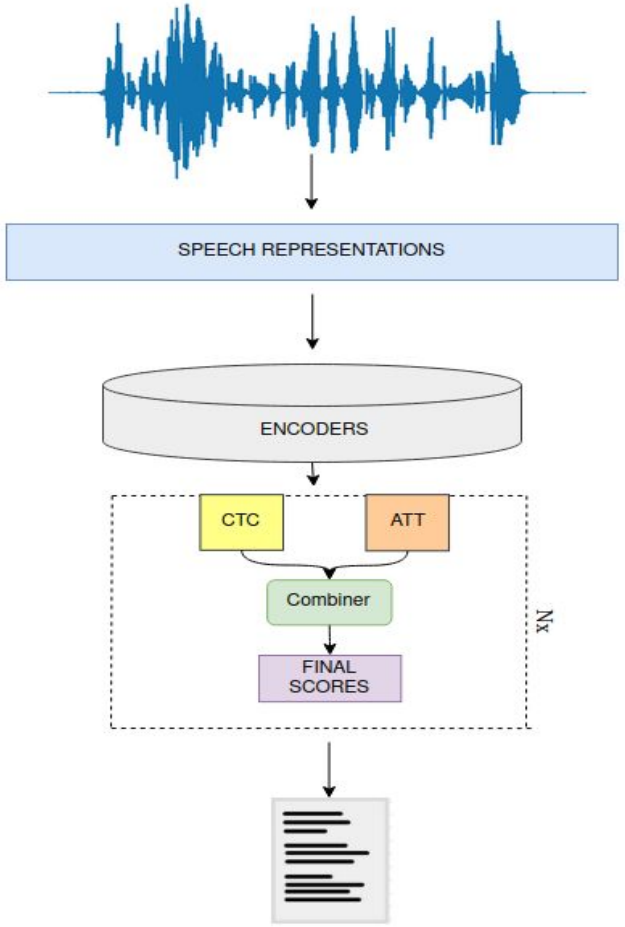
- Kasna fuzija [7]
- Integrirali veliki jezični model u procesu dekodiranja (beam search) - (eng. Uncertainty Aware Dynamic Fusion)

[7] Chen et al.: It's Never Too Late: Fusing Acoustic Information into Large Language Models for Automatic Speech Recognition. ICLR 2024

Sadržaj

- Uvod
- Dvodimenzionalne reprezentacije zvučnog signala
- Metode kombiniranja različitih reprezentacija zvučnog signala
- **Predloženi pristup**

Predloženi pristup: Kasna fuzija



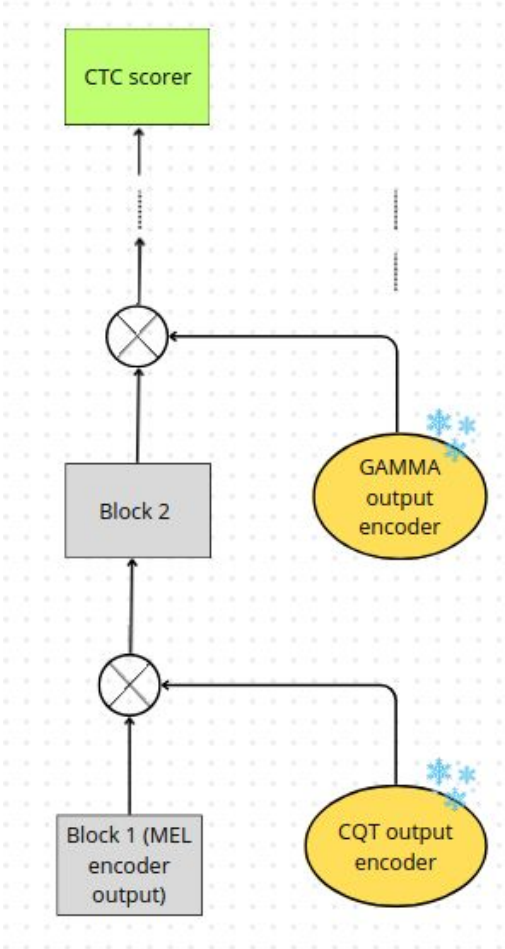
- Metoda **ne** zahtjeva dotreniravanje za sinkronizaciju izlaza modela
- Kombiniramo predikcije hibridnog modela pozornosti i CTC na razini tokena [8]
- Rad prihvaćen na PAKDD 2025. - specijalnu sesiju

$$\text{Score} = \lambda \log p^{ctc}(\hat{y}_n|\hat{y}_{1:n-1}, h) + (1-\lambda) \log p^{att}(\hat{y}_n|\hat{y}_{1:n-1}, h)$$

$$\text{Combined score} = \sum_{i=1}^{\text{num\_repr}} \sum_{j=1}^{2} \alpha_{ij} \text{Score}_{ij}$$

[8] Niko et al.: Streaming End-to-End Speech Recognition with Joint CTC-Attention Based Models. ASRU 2019: 936-943

Predloženi pristup: Srednja fuzija

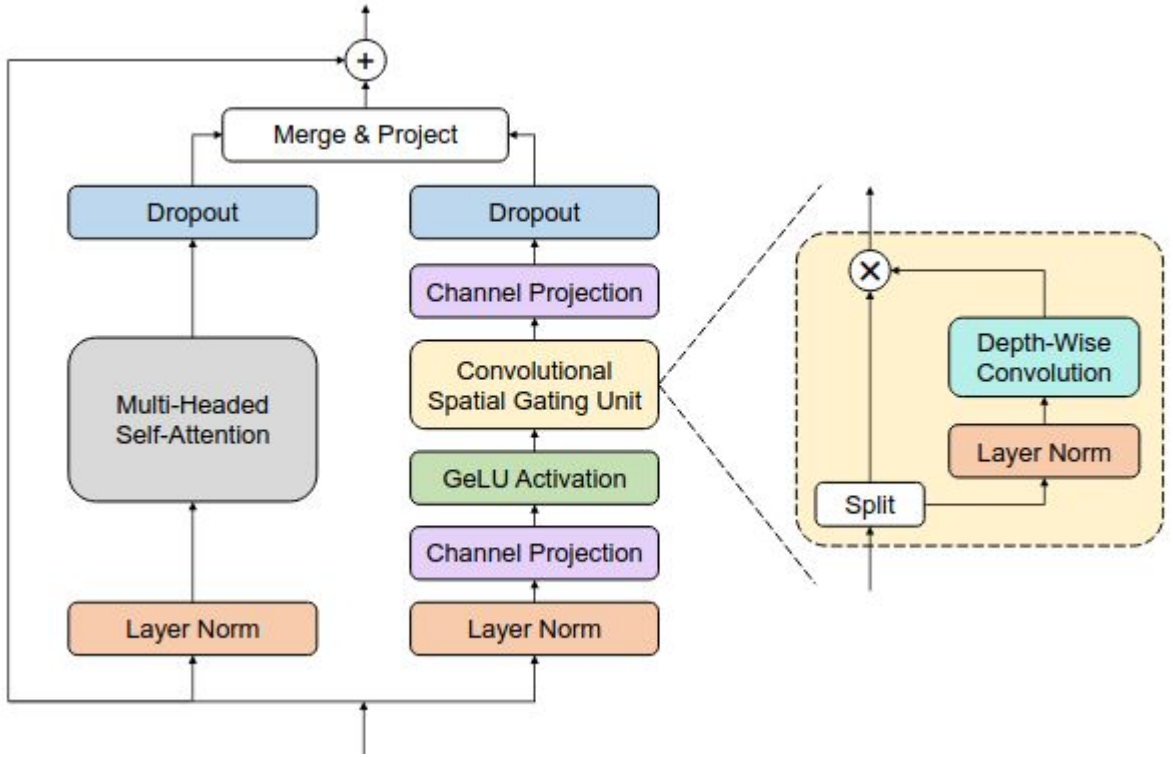


- Koristimo “gotove” enkodere
- Iterativno uvodimo reprezentacije kroz fuzijski modul
- Ne mijenjamo težine enkodera tokom treniranja
- Postižemo bolje rezultate od standardne konkatenacije

Detalji eksperimenata

- E-branchformer [9] kao bazni model
- Modeli trenirani između 30 i 80 epoha po otprilike tjedan dana na 4 A100 gpu
- Spektralna augmentacija [10]
- Optimalnu kombinaciju modela odabrali na temelju ablacijske studije pohlepnim pristupom

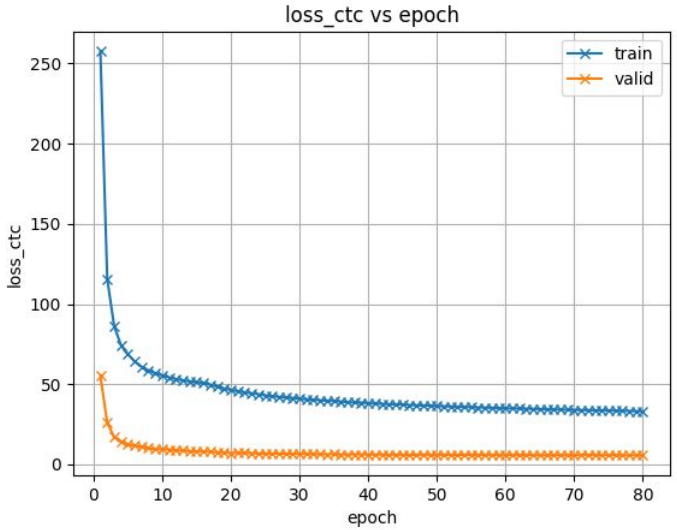
[9] Kim et al.: Branchformer with Enhanced merging for speech recognition. CoRR abs/2210.00077
[10] Tsunoo et al.: Data Augmentation Methods for End-to-end Speech Recognition on Distant-Talk Scenarios. CoRR abs/2106.03419 (2021)

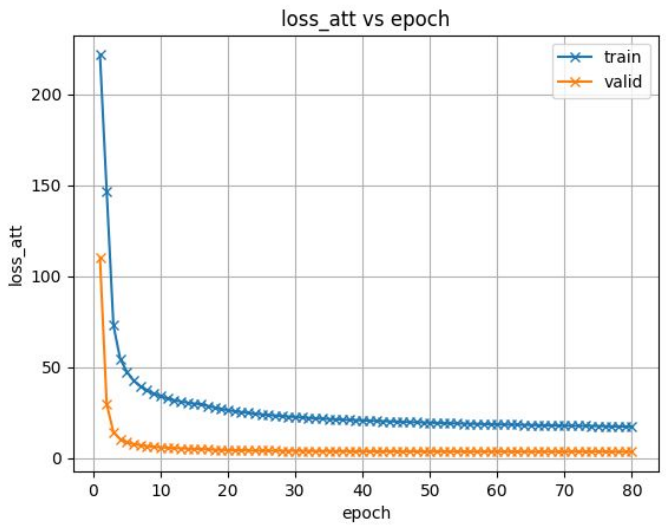


[11] Peng et al.: Branchformer: Parallel MLP-Attention Architectures to Capture Local and Global Context for Speech Recognition and Understanding. CoRR abs/2207.02971

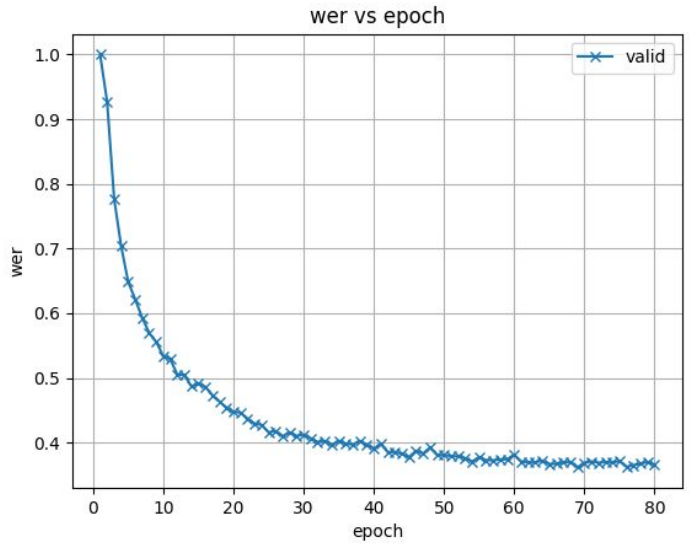
- Skupovi podataka:

Skup podataka	Veličina	Jezik	Broj govornika	Brzina uzorkovanja
LibriSpeech	1,000 sati	Engleski	~2,500	16 kHz
GigaSpeech	10,000 sati	Engleski	~40,000	16 kHz
AISHELL-1	170 sati	Mandarinski	400	16 kHz
TED-LIUM v2	207 sati	Engleski	~1,200	16 kHz



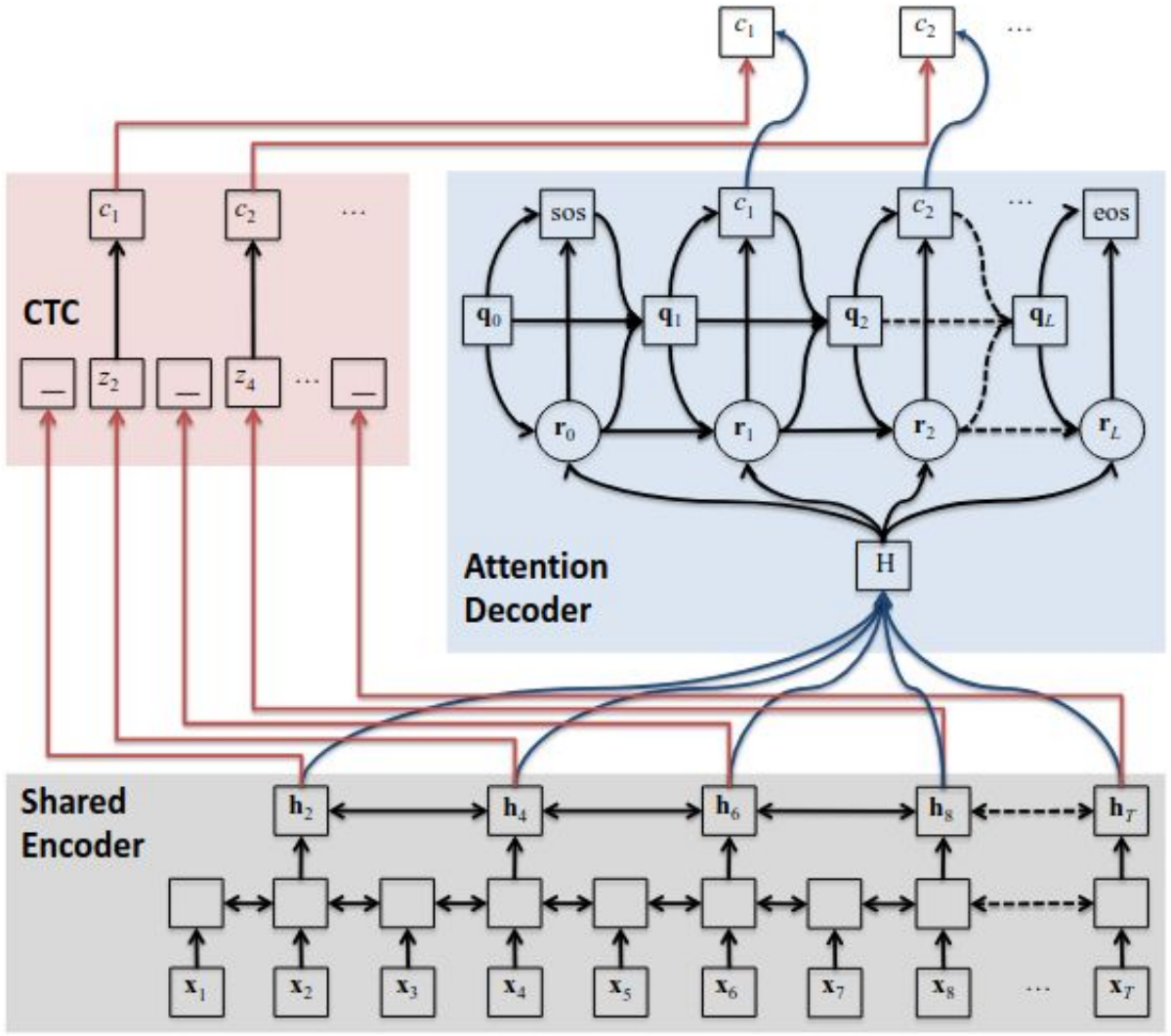


??



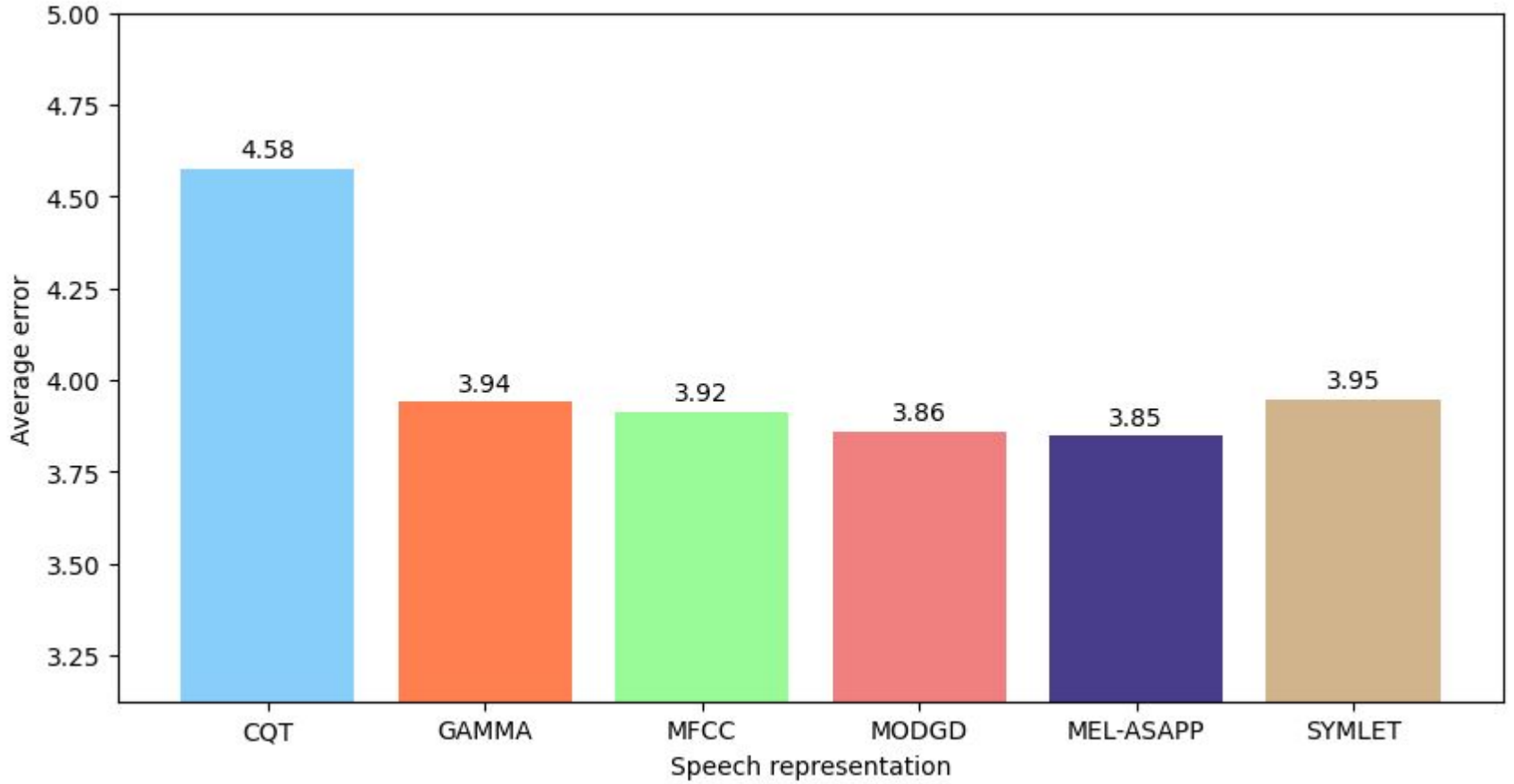
- Nestabilan trening modela [12]

[12] Peng et al.: A Comparative Study on E-Branchformer vs Conformer in Speech Recognition, Translation, and Understanding Tasks. CoRR abs/2305.11073


CTC
Attention Decoder
Shared Encoder
sos
eos
H

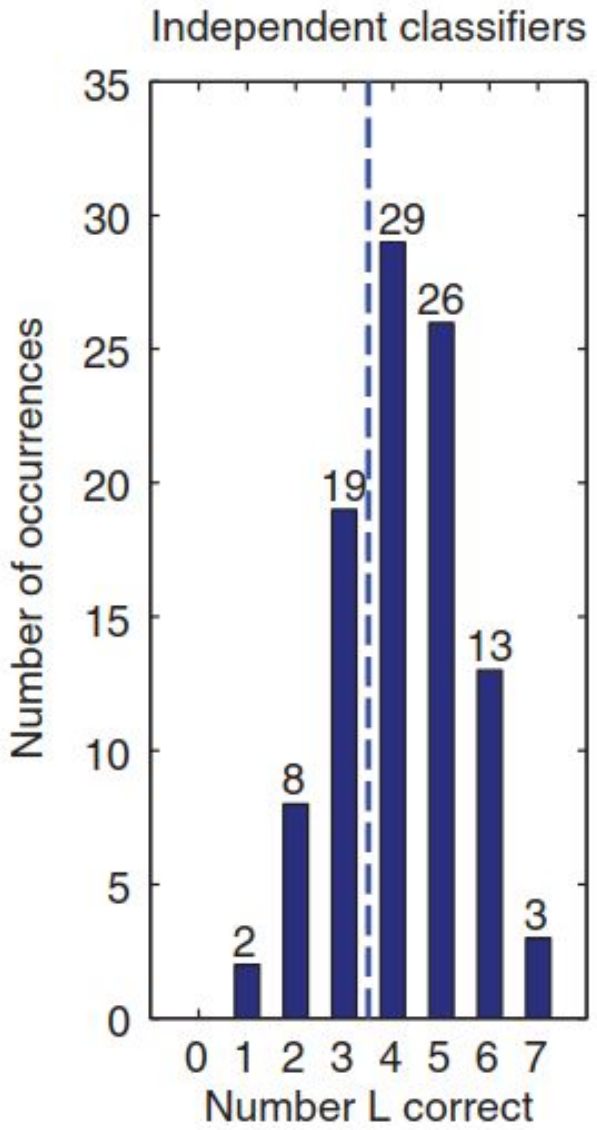
Rezultati

- Ansambl pokazuje relativno poboljšanje od oko **8-14%** u usporedbi s najboljim pojedinačnim modelima
- Značajnija poboljšanja vidimo u šumovitim uvjetima
- Dodavanje jezičnog modela u ansambl, slično kao u [6] dodatno snižava grešku


5.00
4.75
4.50
4.25
4.00
3.75
3.50
3.25
Average error
4.58
3.94
3.92
3.86
3.85
3.95
CQT
GAMMA
MFCC
MODGD
MEL-ASAPP
SYMLET
Speech representation

Raznolikost ansambla

- Koristimo Težinski histogram za određivanje raznolikosti
- Pokazujemo da predloženi ansambl ostvaruje raznolikost dovoljnu za postizanje rezultata industrijskih sistema treniranih na mnogo više podataka



[13] Ludmila I. Kuncheva: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, 2004.

Zaključak i nastavak

- Meta-klasifikatori logičan idući korak
- Zahvala SRCE organizaciji na računalnim resursima

Hvala na pažnji!

Pitanja?

srce
Sveučilište u Zagrebu
Sveučilišni računski centar

Srce politikom otvorenog pristupa široj javnosti osigurava dostupnost i korištenje svih rezultata rada Srca, a prvenstveno obrazovnih i stručnih informacija i sadržaja nastalih djelovanjem i radom Srca.

www.srce.unizg.hr

creativecommons.org/licenses/by/4.0/deed

www.srce.unizg.hr/otvoreni-pristup

srce
otvoreni pristup