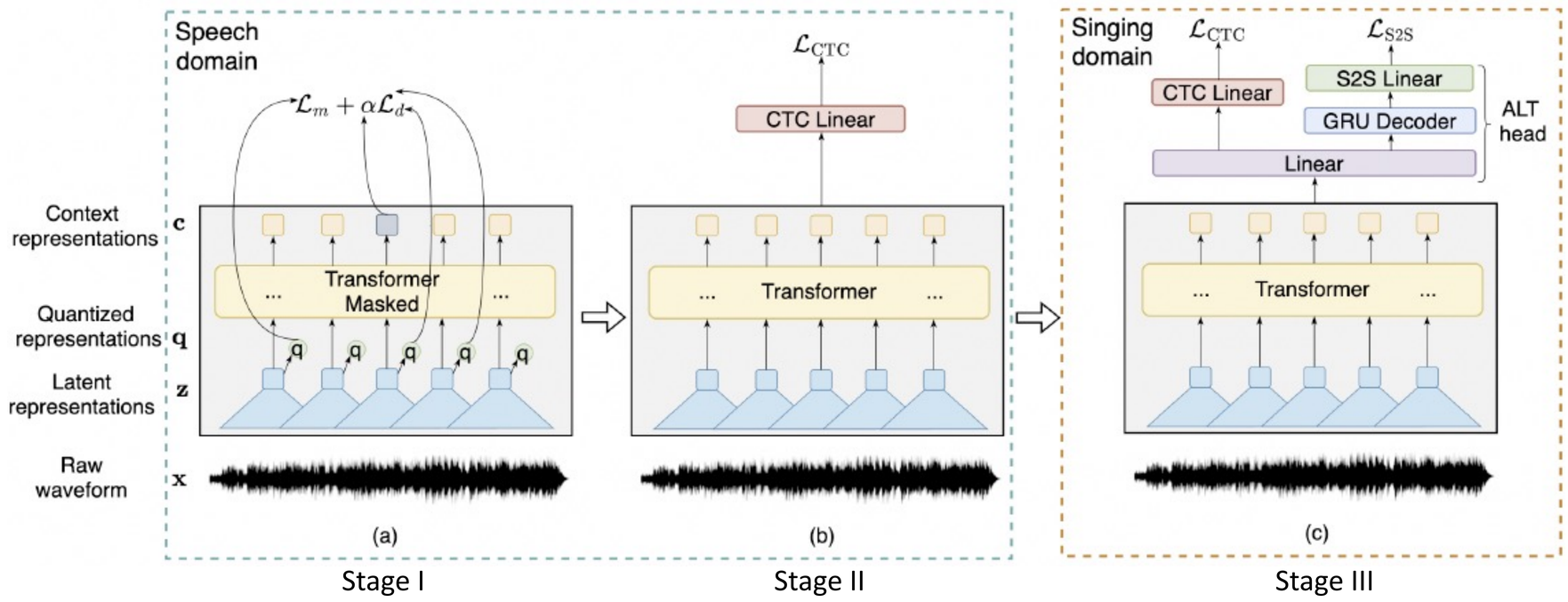
Paper ID #94

Transfer Learning of wav2vec 2.0 for Automatic Lyric Transcription

Longshen Ou*, Xiangming Gu*, Ye Wang
(*Both authors contributed equally to this research)



• Introduction:

➢ The task of Automatic lyric transcription (ALT):

Singing voice

Lyric Transcriptor

I'm sitting here in a boring room …

Lyric texts

➢ ALT is difficult, because

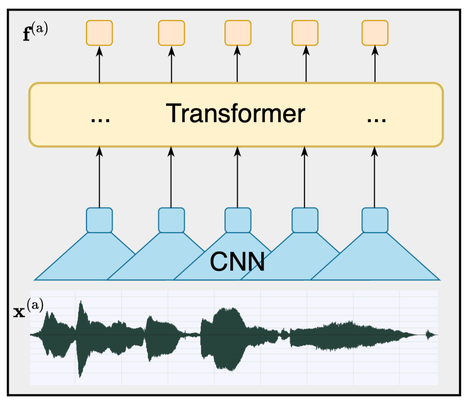
- Degraded intelligibility of sung words
- Perturbations, e.g., musical accompaniments, other ambient noise, etc.
- Small datasets

➢ We propose to design a transfer-learning-based system to solve this problem

- To utilize the similarity between speech and singing
- To reduce the amount of labeled data needed

• Our System:

➢ wav2vec 2.0 as the encoder



Structure: CNN + Transformer

Pretraining objectives:

1. **Masked input**, to learn contextualized representation
2. **Contrastive loss**, to encourage contextualized feature to only be similar to local feature of the same frame

➢ Training procedures:

1. Stage I: Pretraining on speech
2. Stage II: Finetuning on ASR task, on speech domain
3. Stage III: Finetuning on ALT task, on singing domain

➢ Extend to hybrid CTC/attention

- Training objective: weighted CTC and sequence-to-sequence loss

$$\mathcal{L}_w = \lambda_a \mathcal{L}_{\text{CTC}} + (1-\lambda_a)\mathcal{L}_{\text{S2S}}$$

- Inference: weighted likelihood of CTC, decoder, and language model

$$\begin{aligned} \boldsymbol{w}' = \arg\max_{\boldsymbol{w}} \lambda_b \log \sum_{\boldsymbol{\pi}\in\mathcal{B}^{-1}(\boldsymbol{w})} \prod_{t=1}^{T} p(\boldsymbol{\pi}_t|\boldsymbol{f}_t) \\ + (1-\lambda_b)\log \prod_{s=1}^{S} p(\boldsymbol{w}_s|\boldsymbol{w}_{<s}, \boldsymbol{f}_{1:T}) \\ + \lambda_c \log p_{LM}(\boldsymbol{w}) \end{aligned}$$

• Experiments:

➢ We achieve state-of-the-art performance on various benchmark datasets (metric: word error rate, WER, lower is better)

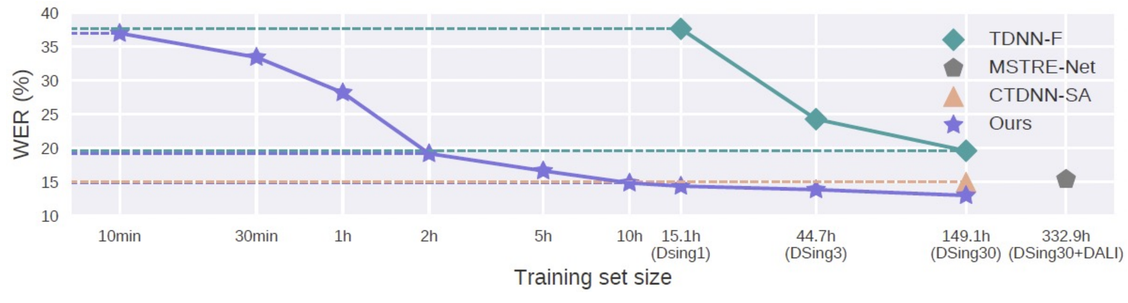
Method	DSing$^{\text{dev}}$	DSing$^{\text{test}}$	DALI$^{\text{test}}$	Jamendo	Hansen	Mauch
TDNN-F [5]	23.33	19.60	67.12	76.37	77.59	76.98
CTDNN-SA [6]	17.70	14.96	76.72	66.96	78.53	78.50
Genre-informed AM [3]	-	56.90	-	50.64	39.00	40.43
MSTRE-Net [7]	-	15.38	42.11	34.94	36.78	37.33
DE2 - segmented [4]	-	-	-	44.52	49.92	-
Ours	**12.34**	**12.99**	**30.85**	**33.13**	**18.71**	**28.48**

➢ Ablation studies show the necessity of all three training stages, and the effectiveness of hybrid CTC/attention

Method	DSing$^{\text{dev}}$	DSing$^{\text{test}}$
Ours	**12.34**	**12.99**
- Finetuning	12.64 (+ 0.30)	14.58 (+ 2.59)
- Pretraining	35.61 (+24.27)	39.13 (+26.14)

Method	DSing$^{\text{dev}}$	DSing$^{\text{test}}$
CTC	19.86	20.99
+ S2S	15.63 (-4.23)	16.95 (-4.04)
+ LM	**12.34** (-7.52)	**12.99** (-8.00)

➢ Our system demonstrates effectiveness in low-resource experiment setups:



• Take-aways

➢ We presented an ALT solution that takes advantage of similarity between speech and singing, by transfer learning of wav2vec 2.0 on singing data

➢ Our method surpass previous ones on various ALT datasets, and still demonstrate competitive results with very limited labeled data.

➢ Check out our project website github.com/guxm2021/ALT_SpeechBrain, or scan this QR code: