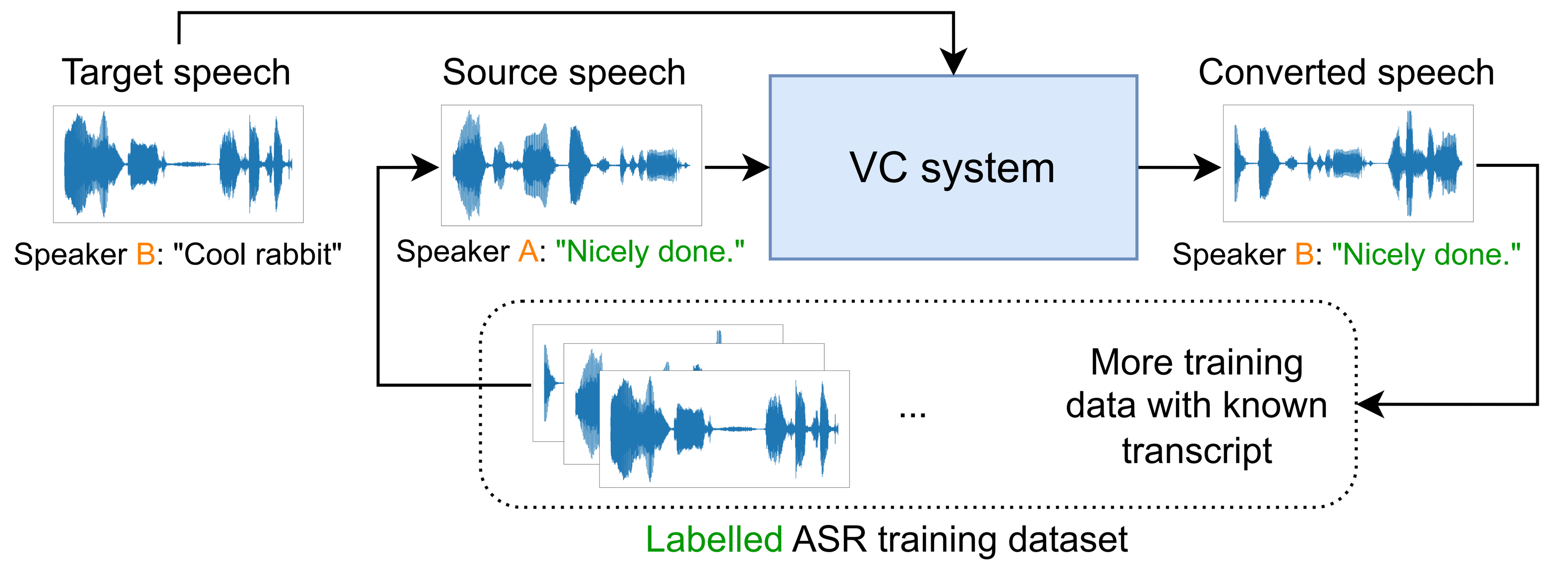
A Temporal Extension of Latent Dirichlet Allocation for Unsupervised Acoustic Unit Discovery

Werner van der Merwe, Herman Kamper, Johan du Preez
MediaLab, E&E Engineering, Stellenbosch University, South Africa

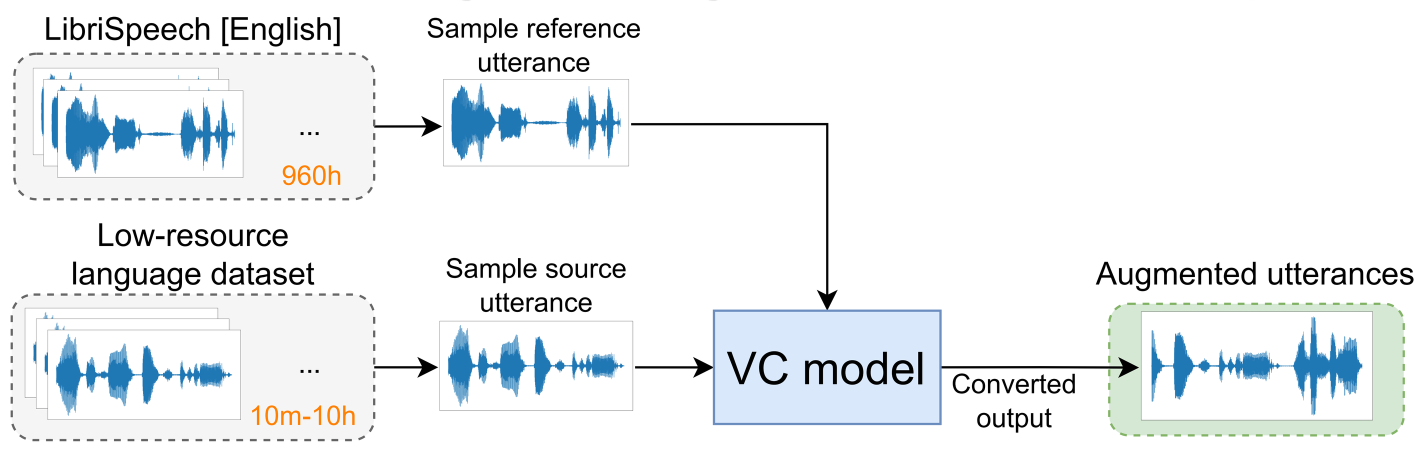
Background

- Voice conversion (VC) can be used to augment training data of automatic speech recognition (ASR) systems:



- Attempts using VC for English ASR augmentation have met limited success.
- ASR still struggles in very low-resource settings with < 1 h of labeled speech.
- **Research question:** can we design a VC system which can be used cross-lingually to improve ASR performance in very low-resource settings?
- To be practically useful for ASR augmentation, it needs to:
  1. work on unseen languages and speakers (any-to-any VC model)
  2. run reasonably fast so that augmenting an entire dataset is feasible.
  3. retain high quality in low-resource settings

Cross-lingual augmentation setup

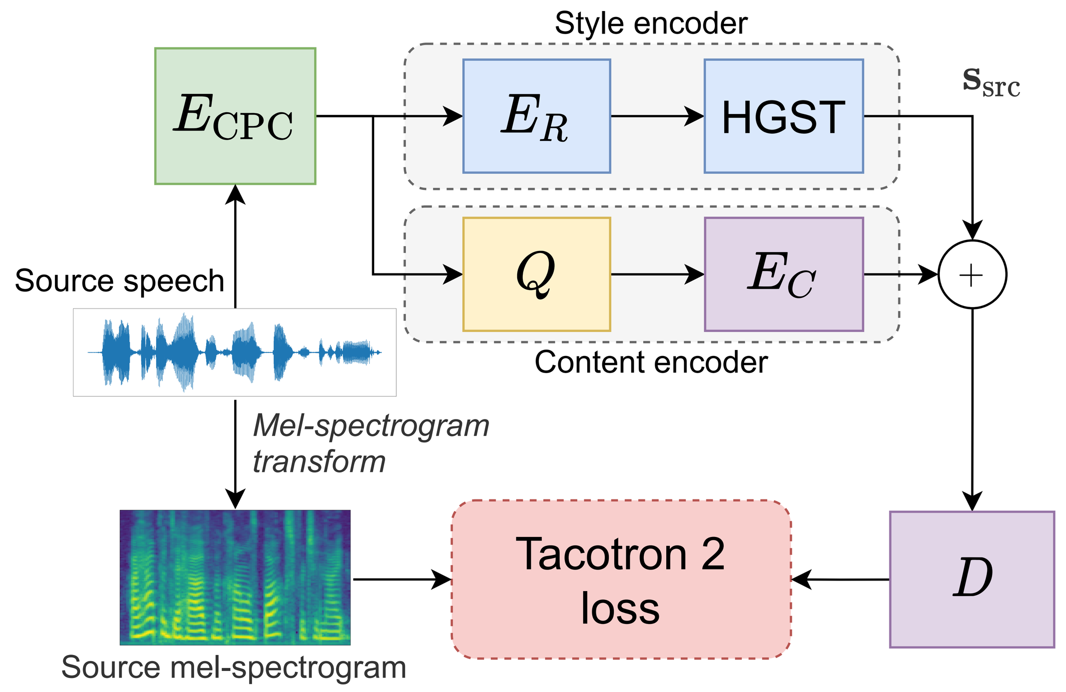


Combined augmented and original low-resource data now contains greater speaker diversity $\implies$ improve ASR generalization.

Experimental setup

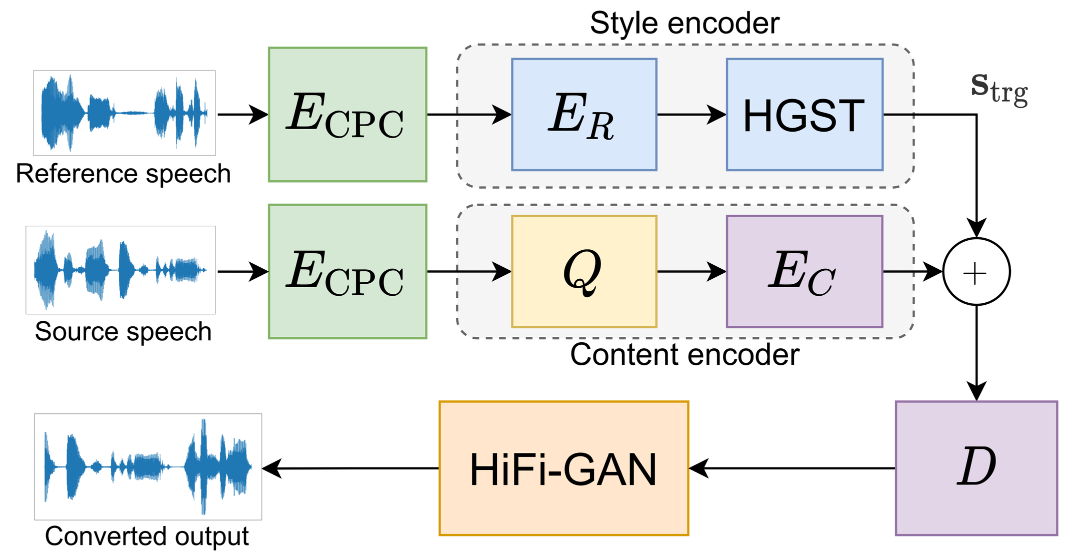
- **VC training:** use fixed pretrained `CPC-Big` encoder and train rest on non-parallel LibriSpeech 960 h training set.
- **ASR setup:** use pretrained `XLSR-53` wav2vec 2.0 model, fine-tune with CTC on labelled low-resource settings with varying amounts of augmented data. No LM used in very low-resource settings – unlikely to be available.
- **Low-resource settings**: simulated low-resource English baseline, final evaluations with 10 min of Afrikaans, Setswana, isiXhosa, Sepedi speech.

Training setup: a new VC approach



- Style encoder isolates speaker identity in style vector $\mathbf{s}_{\text{src}}$; content encoder captures linguistic content as CPC feature sequence.
- Self-supervised objective: L_2 loss between source and predicted mel-spectrogram from Tacotron 2 style decoder D.

Inference setup



Feed different utterances into style and content encoder branches $\implies$ output obtains speaker from style encoder, linguistic content from content encoder.

Ablations and conversion quality

Table: Re-synthesis results in terms of ASR performance on original and converted data. Lower W/CER (%) is more intelligible.

	English		Afrikaans		Sepedi	
VC Model	WER	CER	WER	CER	WER	CER
Original data	5.7	1.9	6.3	4.3	2.1	0.9
Full model	20.6	9.6	32.5	11.0	20.4	9.5
Without HGST	21.3	9.9	34.0	11.9	21.1	9.9
Without Q	**7.1**	**2.6**	**17.0**	**4.6**	**3.7**	**1.6**

Table: Speaker similarity error rates (%) – Lower values mean converted is closer to reference than source (better VC).

VC Model	English	Afrikaans	Sepedi
Full model	8.7	22.0	58.3
Sans HGST	**2.3**	**6.9**	**34.4**
Sans Q	99.7	99.8	99.9

Without HGST: better VC; without Q: more intelligible.
Summary: decent VC and intelligibility when using both HGST and Q.

Validating data augmentation

Table: WERs (%) on LibriSpeech test data for ASR models trained with increasing amounts of VC- and SpecAug-augmented data, with and without 4-gram LM decoding.

		No LM			LM decoded		
Augmentation	Amount	10 min	1 h	10 h	10 min	1 h	10 h
None	0%	47.7	**30.4**	13.4	17.4	**10.6**	**7.5**
VC	100%	43.8	32.7	13.5	**17.2**	11.4	7.6
VC	500%	43.5	34.4	14.4	17.9	11.9	8.1
SpecAug	100%	44.3	31.8	**13.1**	18.8	11.2	7.6
SpecAug	500%	43.1	34.4	13.3	17.7	12.1	7.7
VC→SpecAug	100%	42.5	31.3	13.2	18.5	11.2	7.6
VC→SpecAug	500%	**42.4**	35.0	14.2	18.4	12.5	8.1

Insight: augmentation only helps when we have very poor speaker diversity $\implies$ only helps in very low-resource settings with $\sim$ 10 min of labelled data.

Very low-resource settings

Table: ASR results (%) on test data of four low-resource languages when trained on 10 min of real audio data and different amounts of additional VC- and combined VC-SpecAug augmented data. Sepedi* uses a non-default training procedure.

Language	Augmentation	Amount	WER	CER
Afrikaans	None	0%	52.3	15.9
	VC	100%	**48.9**	**15.0**
	VC→SpecAug	100%	53.5	16.5
Setswana	None	0%	68.9	26.1
	VC	100%	**65.9**	**25.1**
	VC→SpecAug	100%	69.3	26.8
isiXhosa	None	0%	63.2	15.5
	VC	100%	**56.5**	**13.8**
	VC→SpecAug	100%	69.3	26.8
Sepedi*	None	0%	92.6	50.7
	VC	100%	**52.8**	**19.9**
	VC→SpecAug	100%	97.8	69.1

* different training configuration to allow training convergence.

Conclusions

- VC can be used for data augmentation to improve ASR, but primarily in very low-resource settings with limited speaker diversity.
- Cross-lingual VC to drastically different languages unseen during training works well enough to improve ASR performance.
- **Future**: while our cross-lingual VC augmentation is complementary with SpecAugment, how does it compare/combine with other forms of augmentation? And how well do different VC systems work for cross-lingual VC to unseen languages?

Poster presented at Interspeech 2022. [Incheon Korea, September 2022]

Audio demo and samples: `https://rf5.github.io/interspeech2022/`