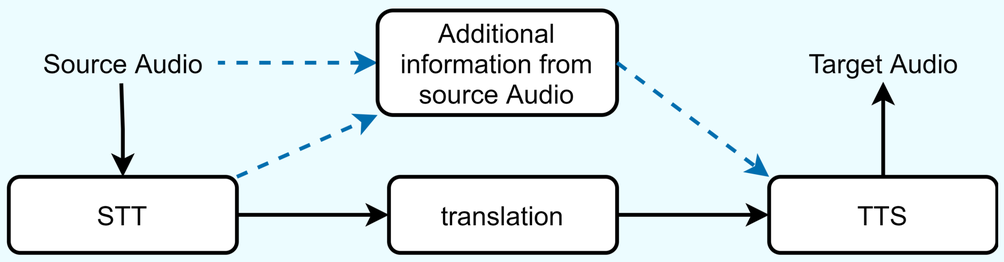
Maastricht University

LibriS2S: A German-English Speech-to-Speech Translation Corpus

Pedro Jeuris, Jan Niehues
Department of Data Science and Knowledge Engineering, Maastricht University

Motivation and Background Speech-to-Speech Translation (S2ST)

- Translate speech from one language to speech in another language.
- Current approaches (black/full arrows) use concatenation of Speech-to-Text (STT), text-to-text translation and Text-to-Speech (TTS) models.
- Can result in a loss of information on speech characteristics such as the pitch and energy (Sperber and Paulik, 2020).
- Proposed (blue/dashed arrows) pipeline to pass information from the source speech to the TTS system to synthesize the target speech:



- Requires parallel speech in both languages.
- Most datasets do not meet the requirement, require a costly license or have a limited amount of samples/no labeled data.
- This paper introduces **LibriS2S** for German and English, consisting of parallel speech and transcriptions.

Dataset Creation

Requirements: parallel human speech in both languages with their transcripts.

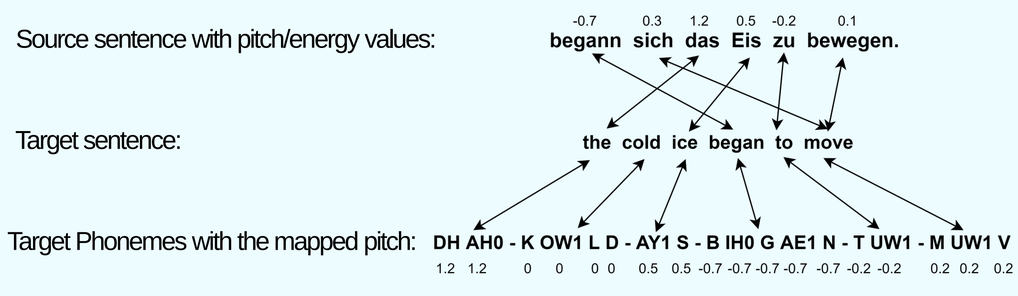
Steps:

- Start from existing dataset, LibrivoxDeEn (Beilharz et al., 2020), which contains the German speech, transcripts and English translation from audiobooks.
- Scrape the English audiobooks from librivox.org.
- Align the scraped data using speech alignment tools.
- Combine the scraped data and librivoxDeEn dataset.

	German	English
# Audio files	25 635	25 635
# Unique tokens	10 367	9 322
# Words	49 129	62 961
# Speakers	42	29
Duration (hh:mm:ss)	52:30:57	57:20:10

Source Feature Vectors (SFV)

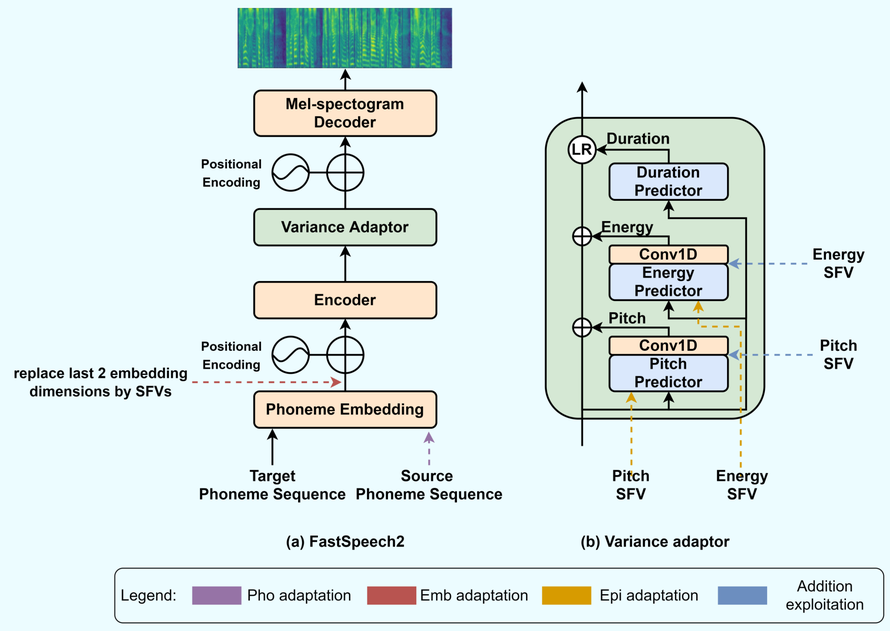
- Way to represent the pitch and energy (influences the volume and prosody of speech) for each phoneme.
- Used as additional input to give information on the speech characteristics in the source speech.



- Pitch and energy mapped from word in source language to corresponding word/phonemes in target language if possible.

Model Adaptations of FastSpeech 2 TTS model

Selected FastSpeech 2 for its ability to control pitch and energy to a certain extent by the dedicated predictors.



Adaptation of image from FastSpeech 2 (Ren et al. 2019,2021).

Multiple adaptations of the FastSpeech 2 architecture have been tested:

- "pho": only uses the phoneme sequence from the source as additional input.
- "emb": only uses the SFV's to replace the last 2 embedding dimensions.
- "epi": SFV's used in the embedding and as input to the pitch and energy predictor.
- "addition": SFV's added to the output of the energy/pitch predictor.

Models are trained on 2079 audio files from a single speaker and as vocoder an MB-MelGAN (Yang et al., 2020) model was finetuned on the same data.

Results

- MOSNet (Lo et al., 2019) used to approximate the Mean Opinion Score (MOS).
- Pitch moments and Dynamic Time Warping (DTW) to compare pitch to the Ground Truth (GT)
- Mean Absolute Error (MAE) for energy

	MOSNet	Pitch σ	Pitch γ	Pitch κ	Pitch DTW	Energy MAE
GT	3.673	31.867	0.788	1.769	\	\
baseline	3.133	41.163	-1.138	2.627	21.423	10.039
pho	3.161	40.778	-1.063	2.931	**19.876**	10.110
emb	**3.163**	**38.113**	-1.000	3.104	20.329	**10.002**
epi	3.159	38.704	-1.039	2.979	19.948	10.103
addition	3.071	42.174	**-0.807**	**2.390**	23.065	11.042

Conclusion

- New S2ST translation dataset released for the English, German language pair.
- The tools used are released together with it and can be utilized to extend it to other languages.
- Introduced dataset was used to train adaptations of FastSpeech 2 that also take information from the source speech as input.
- Results show that the adapted models improve upon the baseline model but need further investigation.

References

Beilharz, B., Sun, X., Karimova, S., and Riezler, S. (2020). Librivoxdeen: A corpus for german-to-english speech translation and speech recognition
Lo, C., Fu, S., Huang, W., Wang, X., Yamagishi, J., Tsao, Y., and Wang, H. (2019). Mosnet: Deep learning based objective assessment for voice conversion.
Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2019). Fastspeech: Fast, robust and controllable text to speech
Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2021). Fastspeech 2: Fast and high-quality end-to-end text to speech
Sperber, M. and Paulik, M. (2020). Speech translation and the end-to-end promise: Taking stock of where we are
Yang, G., Yang, S., Liu, K., Fang, P., Chen, W., and Xie, L. (2020). Multi-band melgan: Faster wave-form generation for high-quality text-to-speech.

Contact: p.jeuris@student.maastrichtuniversity.nl
https://www.linkedin.com/in/pedroantonioj/

Dataset & Code: https://github.com/PedroDKE/LibriS2S

Paper: https://arxiv.org/abs/2204.10593