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

Pedro Jeuris and Jan Niehues

Department of Data Science and Knowledge Engineering, Maastricht University
Maastricht, The Netherlands
p.jeuris@student.maastrichtuniversity.nl, jan.niehues@maastrichtuniversity.nl,

Abstract

Recently, we have seen an increasing interest in the area of speech-to-text translation. This has led to astonishing improvements in this area. In contrast, the activities in the area of speech-to-speech translation is still limited, although it is essential to overcome the language barrier. We believe that one of the limiting factors is the availability of appropriate training data. We address this issue by creating LibriS2S, to our knowledge the first publicly available speech-to-speech training corpus between German and English.

For this corpus, we used independently created audio for German and English leading to an unbiased pronunciation of the text in both languages. This allows the creation of a new text-to-speech and speech-to-speech translation model that directly learns to generate the speech signal based on the pronunciation of the source language.

Using this created corpus, we propose Text-to-Speech models based on the example of the recently proposed FastSpeech 2 model that integrates source language information. We do this by adapting the model to take information such as the pitch, energy or transcript from the source speech as additional input.

Keywords: Speech-to-Speech translation, Speech synthesis, Dataset creation

1. Introduction

As the world becomes more and more connected, both online and offline, the need for language translation systems rises to overcome the large existing language barrier (Hutchins, 2003; Nakamura, 2009). In the past few years the advances made within the Machine Translation (MT) domain have been significant. Computers have become better at translating text between two languages and can even translate speech to text in another language by using a single model (Bahdanau et al., 2016). To solve the Speech-to-Speech Translation (S2ST) problem, in which a spoken phrase needs to be translated and spoken aloud in a target language, the problem is typically broken into three steps. In the first step the source speech gets transformed into text by using a Speech-to-Text (STT) model. The second step consist of translating the text to the target language. The last step takes in the translated text as input and synthesises speech in the target language using a Text-to-Speech (TTS) model. Because of the concatenation of these three steps, and the intermediate text representation, there is a loss of information from the original speech which does not get passed on after the first step (Sperber and Paulik, 2020). This information loss mainly consists of spoken language characteristics such as the pitch and energy in the source speech which are key elements for prosody. Prosody is essential for a conversation as it contains, for example, information on the emotional state of the speaker.

In this work, we want to address this issue by proposing dedicated TTS systems for speech translation that also take the source audio into account. The idea is that when synthesising the speech of a target language, information about the pitch and energy from the source
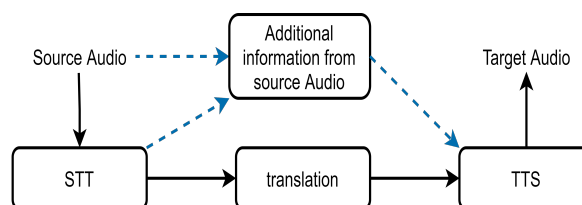

Figure 1: Conventional (black, full lines) and proposed (blue, dashed lines) additions to the pipelines.

language can be valuable to generate the right prosody. Of course, there is no one-to-one mapping. Therefore, we propose to integrate this information into the neural TTS system such that the model can automatically learn when to make use of the source language information. This is done by adding speech characteristics such as the pitch and energy from the speech in the source language into the TTS system as an additional input to the system as shown in Figure 1.

The main challenge we need to address is the creation of appropriate training data. In order to train the proposed model, we need to have aligned speech in both the source and target language. This parallel speech corpus should be created independently from each other to have high quality and unbiased audio in both languages. Therefore, we build the corpus based on the Librevox[1] project. In this case, we aligned the audio of audiobooks in German and English. The speech in both languages was created independently of each other.

The contributions of this research are first of all the

---

[1] https://librivox.org/

creation and release of LibriS2S[2] (Libri Speech-to-Speech), a speech aligned corpus between German and English with audio in both languages.

In addition, we made the scripts publicly available to encourage the creation of additional S2ST data sets. With the help of these alignments and scripts a German-English speech aligned dataset can be made. By adapting the scripts this could be extended to more languages too. Secondly, we perform an investigation on how source text and audio features can be used to improve the quality of the synthesised speech from the TTS model. We propose an extension of the state-of-the-art TTS system FastSpeech 2 (Ren et al., 2021) to integrate these features as well as evaluate the influence of the presented data.

2. Related Work

2.1. Datasets

Most datasets used in current state-of-the-art translation research such as LibrivoxDeEn (Beilharz et al., 2020), CoVoST2 (Wang et al., 2020), MuST-C (Cattoni et al., 2021) or LibriSpeech ASR(Panayotov et al., 2015) augmented with French translations (Kocabiyikoglu et al., 2018) are focused on Speech-to-Text translation. These datasets only contain the audio and transcription in the source language paired with the translation, in text, from the target language. Furthermore, datasets that do contain the audio from the target language such as Fisher and CALLHOME Spanish-English dataset (Post et al., 2013), SP2 speech corpus (Sečujski et al., 2016), the EMIME Project (Kurimo et al., 2010), VoxPopuli(Wang et al., 2021) and MaSS (Zanon Boito et al., 2020) either require a costly license, only have a limited amount of samples, exist out of unlabeled data or might have copyright concerns.

Since audio pairs of the same sentences in two languages are needed, to build the proposed TTS system and to provide the additional information during training and testing, this work will build further upon already existing speech-translation datasets. The missing speech segments from the target language will be scraped from the internet and aligned to match the dataset.

2.2. S2ST and TTS systems

Modern day S2ST systems such as provided by Jibbigo (Eck et al., 2010), Microsoft Azure (MircosoftAzure, 2021) and Skype (Lewis, 2015) make use of a concatenation of state-of-the-art models. As mentioned in the introduction this can result in a loss of information. The oldest pipeline to take additional speech information, such as prosody, into account was that of Verbmobile (Wahlster, 2013). Within the Verbmobile project, pitch, energy and other features of speech were used to annotate the generated text with additional information and to use this additional information to help with

---

the translation. Trying to achieve the same goal, in this paper the source information will be used only to enhance the quality of the generated speech at the end of the pipeline.

Do et al. (2016) try to incorporate emphasis in their TTS model with the help of an emphasis estimator. Their emphasis estimator takes the source text as input and predicts for each word an emphasis probability between 0 and 1. They then transfer this emphasis estimation to the target language and use these values as additional input of their TTS model. Compared to this work, source information will also be mapped to the target language but instead of using an intermediate predictor, the values are extracted from the source speech to be used. Also, instead of using a Hidden Semi-Markov (HSM) TTS Model, our approach uses a deep learning method.

The work of Skerry-Ryan et al. (2018) focuses on transferring prosody between two speakers within the same language. By adapting the Tacotron (Wang et al., 2017) architecture to contain an extra prosody embedding layer they are able to transfer prosodic characteristics between speakers in the same model.

The feature generation model used in this work is the recently introduced FastSpeech 2 (Ren et al., 2021). This model is of particular interest because it has separate predictors for pitch and energy as can be seen in Fig. 2. Compared to the original paper, the implementation[3] used has a few differences: 1) The pitch and energy are predicted per phoneme. 2) The pitch and energy are normalized using the z-normalization and these are being predicted. 3) Additionally a post-net similar to the one introduced in the Tacotron 2 (Shen et al., 2018) paper is used. This post-net is used to "predict a residual to add to the prediction to improve the overall reconstruction" (Shen et al., 2018).

3. Dataset creation

In order to generate the most natural speech in the target language, we were targeting a data source where the source and target language speech was generated as naturally as possible. We found the Librivox project as a promising resource, as here the audio books in both languages are generated independently of each other.

The dataset used as a starting block in this research is the LibrivoxDeEn (Beilharz et al., 2020) dataset. This dataset consists out of German speech, transcriptions and their English translation of audiobooks. These audio books are all available on Librivox and their texts are also freely available. The English translations within this dataset are based on the official translations, which means that the English counterparts of the audiobook is possibly available on Librivox too. Using the

---

[2]https://github.com/PedroDKE/LibriS2S

[3]The baseline implementation used in this work can be found in the following GitHub repository: https://github.com/TensorSpeech/TensorFlowTTS
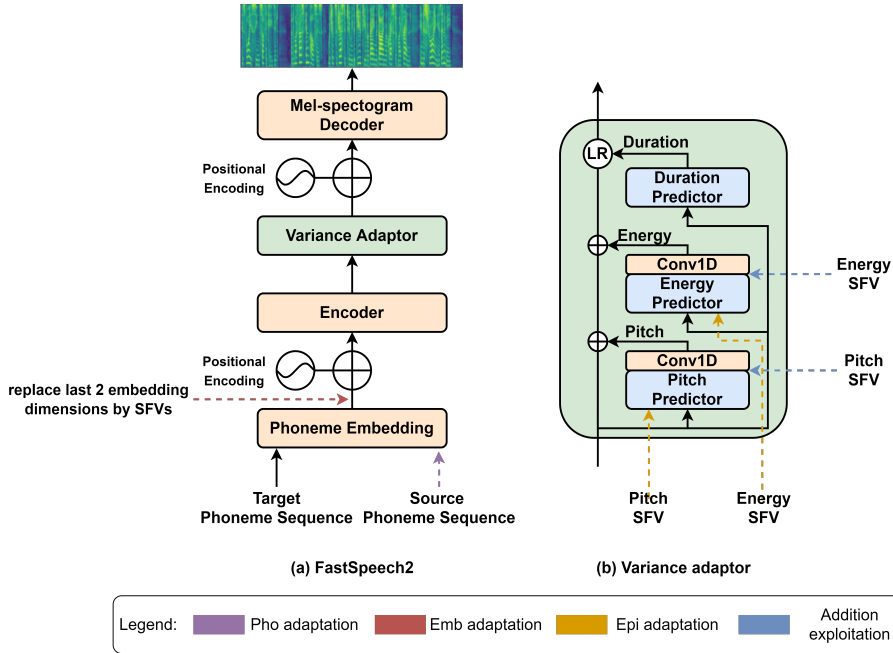
Figure 2: FastSpeech 2 architecture, figure adapted from (Ren et al., 2021). Abbreviations: LR: Length Regulator (introduced in the original FastSpeech paper (Ren et al., 2019), further information on the embedding, encoder, decoder and predictors can be found in the FastSpeech papers)

audiobooks from Librivox has multiple advantages:
1) it is possible to collect parallel data with limited involvement of human aid, 2) some books have multiple speakers which could be used to construct for multi-speaker TTS systems, 3) to build further upon the previous advantage, this method also provides multiple-to-one speaker and one-to-multiple speakers alignments which could be interesting for future research, 4) there is only a limited number of legal and privacy concerns as all data used is part of the public domain according to Librivox and 5) a wide variety of languages are available to further facilitate multi-lingual S2ST.

3.1. Data collection and allignment

The first step to enhance the LibrivoxDeEn dataset was to download the English counterparts of the audiobooks from Librivox. As some of the audiobooks do not have the same amount of chapters across languages, for example the German audio book used in this research contains 29 chapters while the English audio book has 27, these were manually split or stitched together to use the texts given within the LibrivoxDeEn dataset.
The next step was to align the newly gathered English speech to the texts. This was done using the tool aeneas (Pettarin, 2017). Given the text and audio of a book's chapter that we want to align, this tool gives us the timestamps of when a phrase was said and makes it possible to extract the spoken sentences necessary to create the required data for S2ST research.

	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

Table 1: Collected data sizes

3.2. Dataset analysis

Using this method it was possible to align 12 audio books. From those 12 books, 8 did not require the additional prepossessing to align the chapters. Those alignments are made publicly available to download together with the tools used to scrape the audio from librivox and align the audio to their text.
Additional metrics on the 12 books that were aligned can be found in Tab. 1. To count the number of speakers these were counted separately for each book and some speakers can thus have been counted more than once. The speakers that record the audio books on librivox do not always have access to a high-quality microphone. Because of this it can be challenging to create a TTS system for a single speaker. Moreover, the average length of audio per book, about 4.77 hours for English and 4.38 hours for German, is on the lower side when compared to the much used datasets LJSpeech (Ito and Johnson, 2017) for English or Thorsten (Müller, 2019) for German with both over 23 hours of speech. However, using multiple speakers or audiobooks from the dataset to train a multi-speaker

TTS model can be beneficial for the final audio quality (Chen et al., 2020) and makes it possible to make use of the full dataset. Furthermore, the scripts released with this work make it possible to align more audiobooks to the LibrivoxDeEn dataset. Combining both methods from this work and from Beilharz et al. (2020), used to create LibrivoxDeEn, it is possible to create speech aligned datasets for multiple languages given that the right audiobooks are available in both languages on librivox.

Using audio books does not give us a guarantee that sentences in both languages are expressed with the same prosody or emotion. Given the scarcity of such data, audio books can be used as a starting point as these have multiple advantages as discussed in the beginning of Sec 3.

4. Source guided speech synthesis

Given the newly created data we are now able to build TTS models that can use the audio from the source language to improve the generated speech in the target language. In this work we suggest models that use this information to help predict the energy and pitch of the target language. We do this by making three adaptations of the model that are then trained with the newly gathered data and one exploitation model where we try to improve the baseline without retraining. The different implementations are shown in Fig. 2 and further explained in this section.

4.1. Source phoneme input

The first adaptation takes a concatenation of both the source and target transcripts (in phonemes) as input. The embedding layer then maps both phoneme sequences to a vector space and then the encoder is used to mix the information from both embedded sequences. This is the most straightforward integration, where only very limited adaptation to the existing architecture is needed. However, only source text information is passed to the model and the information is not encoded specifically making the task of using this additional information very difficult. Furthermore this representation does not give any additional information on the expressed prosody in the source speech. This model will later on be referred to as 'pho' because of the additional phoneme input.

4.2. Word level pitch and energy embedding

For this adaptation a Source Feature Vector (SFV) is being calculated and used to give the model additional information on the pitch and energy from the source speech.

The first step in creating the SFVs is to calculate the average pitch and energy for each word in the source speech. Afterwards with the help of multilingual alignment, where each word of a source sentence is mapped to their corresponding word in the translation, these SFVs can be mapped to the target language. This way
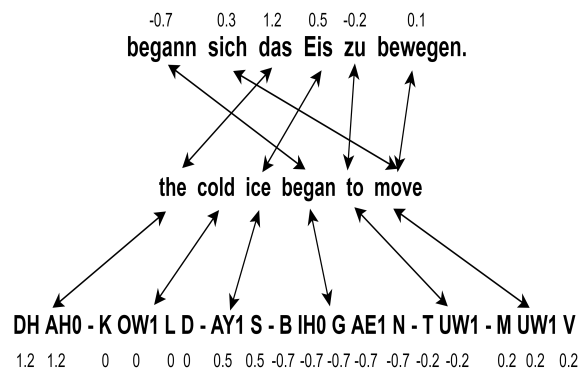

Figure 3: A (synthetic) example of an SFV to show the idea. Note how for the last word the average is being taken since it has a mapping from multiple words in the source

a vector is created where for each phoneme in the target language either contains information on the average word-level value in the source language or, if it was not alligned, a zero value. An example of an SFV can be seen in Fig. 3.

The second adaptation uses this additional information by replacing the last 2 dimensions in the embedding layer by the pitch and energy SFVs. This way the encoder takes the additional information from the SFVs as input which in turn can help the variance adaptor to predict the pitch and energy.

In the results section this model will be refered to as 'emb' because of the additional information that is incorporated in the dimensions of the embedding.

4.3. Word level pitch and energy as additional input to predictor

Instead of only using the SFVs as additional information to the embedding, the third adaptation builds further upon the previous introduced model in Sec. 4.2 and also uses the SFVs as additional input to their respective predictors. This way, the predictors also receive the values from their respective SFVs directly. This is done by stacking them to the encoder output to create a vector with sizes (2, #PHONEME). This model is refered to as 'epi' which stands for Embedding and Predictor Input.

4.4. Add pitch and energy to predictor output

Since the pitch and energy predictors try to estimate these values on a phoneme level, this can be exploited to influence the generated speech. As can be seen in Fig. 4 the energy and pitch in the generated speech can be changed by increasing the output values of their predictors before passing the values through the embedding. But it can also be seen that changing values at certain parts of the predictions can result in changes of pitch or energy in other parts of the generated speech too.
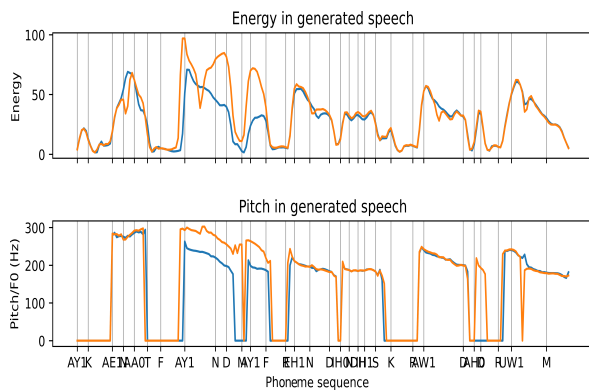
Figure 4: input sentence: "I cannot find my friend in this crowded room", "AY1 KAE1NAA0T FAY1ND MAY1 FREH1ND IH0N DHIH1S KRAW1DAH0D RUW1M" blue: using the output of the predictor, orange: increasing predicted energy and pitch at the phonemes corresponding to "find my"

In the fourth model, this is being exploited within the baseline model by summing the SFV to the output of their respective predictor. This way phonemes from words that have a below average pitch or energy in the source language get their predicted values lowered (because of having a negative z-score added to them) and the same way words with an above average value in the source language get their values increased in the prediction.

5. Experimental setup

5.1. Preprocessing

To convert the transcript to a phoneme sequence and extract the duration of each phoneme the Montreal Forced Alignment tool (McAuliffe et al., 2017) was used. The English sequences uses the ARPAbet phoneme set and the German sequences uses the prosodylab format. To get the cross-lingual alignment to create the SFVs the awesome-align tool[4] (Dou and Neubig, 2021) was used.

To train the TTS models, the speech data was resampled to 22050 kHz and the MFCC's were extracted using a hop-size of 256, a frame size of 1024 and with a basis of 80. For training, testing and validation only audio files with a length between 1s and 20s were used.

5.2. Data size

Some of the collected books either have bad microphone, need further processing to separate the speakers or do not have enough samples to train a single speaker system. Therefore in the continuation of this research the book Frankenstein by Mary Shelly[5] is used. As a

result there were a total of only 2335 audio files used for this research. 2079 of these were used for training, 129 for validation and 127 for testing. This number of samples for training is on the low side compared to datasets usually used for single speaker TTS models such as LJSpeech (Ito and Johnson, 2017) where a total of 13,100 English audio fragments are used. Nevertheless, it is still sufficient to train a TTS model but might result in more/faster overfitting and worse audio quality in the generated speech. The book used also had some misalignments on the source side originating from the LibrivoxDeEn dataset but was realigned using the same method described in Sec. 3.

5.3. Evaluation

The vocoder model used in this research is the MB-MelGAN model (Yang et al., 2020) which has been fine tuned on the samples in the training set and then used to generate the 127 samples from the test set to evaluate the speech generated. As getting a Mean Opinion Score (MOS) is not possible due to the inaccessibility to enough native speakers, the MOS is approximated by using the deep learning approach MOSNet (Lo et al., 2019). The disadvantage about this method is that it is just an approximation and might not represent the real MOS that a human evaluation would give. Fortunately, this method also has advantages such as: more samples can be evaluated, no new biases are introduced during evaluation and the same model is used for testing which results in comparable metrics. To evaluate the energy and pitch in the generated speech, the same methods as in the FastSpeech 2 (Ren et al., 2021) papers are used. To evaluate the pitch, the standard deviation, skewness and kurtosis for the pitch moments in the generated speech is calculated (Niebuhr and Skarnitzl, 2019; Andreeva et al., 2014). Ideally this would be as similar as possible to our ground truth. Additionally the average Dynamic Time Warping (DTW) distance of the pitch is also computed with the help of the dtw-python package (Giorgino, 2009). To evaluate the energy, the MAE between the generated speech and the ground truth is being calculated. Similar to Ren et al. (2021), the durations extracted by the MFA are used in the length regulator to ensure the same duration in the generated audio as the ground truth to calculate the MAE of the energy in the generated speech.

6. Results

6.1. Speech metrics of generated samples

The metrics achieved by evaluating the generated speech of all models can be seen in Tab. 2. The first metric is the predicted MOS of the generated speech. According to the predicted score the three model adaptations improve over the baseline and result in a slightly more natural audio quality. The only model not improving on this metric is the addition exploitation of the baseline model.

The pitch moments from the generated speech can be

	MOSNet prediction	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	40.779	-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

Table 2: Predicted MOS scores, pitch moments/DTW and energy MAE from the GT and generated audio

found between the third and fifth column. Ideally these are as close as possible to the ground truth from the same test set. As can be seen, all trained model adaptations have an improvement on the baseline when looking at the standard deviation (σ) and skewness (γ) but perform worse on the resulting kurtosis (κ). The addition exploitation however achieves the best results on the skewness and kurtosis of the pitch but performs worse when looking at the standard deviation. Besides the pitch moments, the DTW distance to the GT is also given in the sixth column. We can see that again for the three adaptations this distance has decreased when compared to the baseline model indicating a better modeling of the pitch. For the addition model this distance has increased, indicating a worse modeling compared to the baseline. This is probably due to the effect shown in Fig. 4 where unwanted changes in the pitch can happen or because of the fact that prosody is not a one-to-one mapping across two languages.

For the MAE between the energy within the GT and generated speech we can see that only the emb adaptation improves over the baseline and the other two adaptations perform similar to each other. The addition model however performs the worst again, possibly due to the same reasons as to why the DTW distance increased for this model.

6.2. Ablation study on the influence of SFV

To have a closer look at the influence of the SFV on the pitch and energy in the generated speech an ablation study is done where the SFVs are set to zero before passing them to the emb and epi model during inference. Afterwards the same metrics as the previous three sections are calculated. The values achieved with the SFVs extracted from the source speech are also given as GT in the table, these values are the same as from Tab. 2.

Tab. 3 shows the achieved metrics when the SFVs values are put to zero during inference. Due to the z-normalization a zero value means that this part has an average pitch or energy. For the epi model, setting the SFVs values to zero only results in the improvement on one metric. This indicates that the information from the SFV is used in the epi model and improves the final performance of the model. For the emb model it is not as clear and we only see improvements in half the metrics. One reason could be that the SFVs are more directly integrated in the epi approach and therefore it

	MOS	σ	γ	κ	DTW	MAE
emb zero sfv	**3.193**	40.187	-1.039	3.114	**20.094**	**9.985**
emb GT sfv	3.163	38.113	-1.000	3.104	20.329	10.002
epi zero sfv	3.153	38.988	-1.074	**2.730**	20.123	10.137
epi GT sfv	3.159	38.704	-1.039	2.979	19.948	10.103

Table 3: pitch and energy metrics from the generated speech with SFV values being zero. Improvements over non-zero SFV are shown in bold.

is easier for the model to use this information to generate the MFCC.

7. Discussion

Given the results presented in Tab. 2 and Tab. 3, it can be said that while changing the output of the predictors might be a promising way to control the pitch and energy (as seen in Fig. 4) in the generated speech, combining them with the SFVs fails to improve over the baseline on most metrics and has the overall worst performance out of the methods presented. The emb model, where the SFVs are used as additional input to only the embedding space, does improve over the baseline and performs best out of the proposed methods on half of the metrics, but this method gives conflicting results in the ablation study and is therefore a less interesting method. Using the SFVs also as input to the predictors gives the second best results on all metrics and the ablation study shows that using this method can be beneficial for the epi model. While the pho model does not receive any additional information on the energy and pitch from the source speech, only information of what was said, this still seems to improve the generated speech. On most metrics it improves over the baseline and seems to perform similarly to the other two adaptations, emb and epi models, in terms of predicted MOS, DTW distance and the energy MAE.

8. Conclusion

In this paper, LibriS2S, a dataset consisting of sentence aligned German and English audio pairs and their transcriptions, is introduced and released together with the methods used to make this dataset. The presented method has multiple advantages and is shown to be able to be used for a single-speaker TTS system. This

dataset is also used to incorporate additional source information during training and inference of a Fast-Speech 2 model in a S2ST environment. The trained models seem to mainly improve on having a better pitch characteristic as the target and have a higher predicted MOS score by using additional information from the source audio. On the energy level of the generated speech the adaptations do not seem to benefit much from the additional source information. For some of the adaptations conflicting results are achieved in the ablation study and thus might need further investigation. However, this is not the case for all models presented in this work.

The presented data opens up several interesting research directions. While the baseline experiments focused on a single speaker scenario, the provided data offers data from multiple speakers, which can be beneficial for the quality of generated speech(Latorre et al., 2018). The use of the full data will also enable the creation of a full speech translation pipeline and allow to investigate the challenges in modeling the full pipeline. Furthermore, the provided toolkit enables the creation of similar data for other language pairs. This will enable the analysis on the similarity of speech characteristics between different speakers and improve the transfer between languages.

9. Bibliographical References

Andreeva, B., Demenko, G., Möbius, B., Zimmerer, F., Jügler, J., and Oleskowicz-Popiel, M. (2014). Differences of pitch profiles in germanic and slavic languages.

Bahdanau, D., Cho, K., and Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate.

Beilharz, B., Sun, X., Karimova, S., and Riezler, S. (2020). Librivoxdeen: A corpus for german-to-english speech translation and speech recognition. *Proceedings of the Language Resources and Evaluation Conference*.

Cattoni, R., Di Gangi, M. A., Bentivogli, L., Negri, M., and Turchi, M. (2021). Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Chen, M., Tan, X., Ren, Y., Xu, J., Sun, H., Zhao, S., Qin, T., and Liu, T.-Y. (2020). Multispeech: Multi-speaker text to speech with transformer.

Do, T., Toda, T., Neubig, G., Sakti, S., and Nakamura, S. (2016). Preserving word-level emphasis in speech-to-speech translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP.

Dou, Z.-Y. and Neubig, G. (2021). Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Eck, M., Lane, I., Zhang, Y., and Waibel, A. (2010). Jibbigo: Speech-to-speech translation on mobile de-vices. In *2010 IEEE Spoken Language Technology Workshop*, pages 165–166.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in r: The dtw package. *Journal of Statistical Software, Articles*, 31(7):1–24.

Hutchins, W. J. (2003). *The development and use of machine translation systems and computer-based translation tools*. Bahri.

Ito, K. and Johnson, L. (2017). The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Kocabiyikoglu, A. C., Besacier, L., and Kraif, O. (2018). Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation.

Kurimo, M., Byrne, W., Dines, J., Garner, P. N., Gibson, M., Guan, Y., Hirsimäki, T., Karhila, R., King, S., Liang, H., Oura, K., Saheer, L., Shannon, M., Shiota, S., and Tian, J. (2010). Personalising speech-to-speech translation in the EMIME project. In *Proceedings of the ACL 2010 System Demonstrations*, pages 48–53, Uppsala, Sweden. Association for Computational Linguistics.

Latorre, J., Lachowicz, J., Lorenzo-Trueba, J., Merritt, T., Drugman, T., Ronanki, S., and Viacheslav, K. (2018). Effect of data reduction on sequence-to-sequence neural TTS. *CoRR*, abs/1811.06315.

Lewis, W. D. (2015). Skype translator: Breaking down language and hearing barriers. *Translating and the Computer (TC37)*, 10:125–149.

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. *CoRR*, abs/1904.08352.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *INTERSPEECH*, pages 498–502.

MircosoftAzure. (2021). Speech translation. https://azure.microsoft.com/en-us/services/cognitive-services/speech-translation/#features.

Müller, T. (2019). Thorsten: Open german voice dataset. https://github.com/thorstenMueller/deep-learning-german-tts#dataset-thorsten-neutral.

Nakamura, S. (2009). Overcoming the language barrier with speech translation technology. Technical report, Citeseer.

Niebuhr, O. and Skarnitzl, R. (2019). Measuring a speaker's acoustic correlates of pitch -but which? a contrastive analysis based on perceived speaker charisma.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5206–5210.

Pettarin, A. (2017). aeneas. `https://www.readbeyond.it/aeneas/`.

Post, M., Kumar, G., Lopez, A., Karakos, D., Callison-Burch, C., and Khudanpur, S. (2013). Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany.

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.

Sečujski, M., Gerazov, B., Csapó, T., Delić, V., Garner, P., Gjoreski, A., Guennec, D., Ivanovski, Z., Melov, A., Németh, G., Stojkovikj, A., and Szaszák, G. (2016). Design of a speech corpus for research on cross-lingual prosody transfer. pages 199–206.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.

Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R. J., Clark, R., and Saurous, R. A. (2018). Towards end-to-end prosody transfer for expressive speech synthesis with tacotron.

Sperber, M. and Paulik, M. (2020). Speech translation and the end-to-end promise: Taking stock of where we are.

Wahlster, W. (2013). *Verbmobil: Foundations of Speech-to-Speech Translation*. Artificial Intelligence. Springer Berlin Heidelberg.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis.

Wang, C., Wu, A., and Pino, J. (2020). Covost 2 and massively multilingual speech-to-text translation.

Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. (2021). Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Yang, G., Yang, S., Liu, K., Fang, P., Chen, W., and Xie, L. (2020). Multi-band melgan: Faster waveform generation for high-quality text-to-speech. *CoRR*, abs/2005.05106.

Zanon Boito, M., Havard, W., Garnerin, M., Le Ferrand, É., and Besacier, L. (2020). MaSS: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6486–6493, Marseille, France. European Language Resources Association.