User De-Identification over Speech/Dialogue exchange

Abhishek Das (addas) [1] **Sweta Priyadarshi (swetap)** [1] **Subhadra Gopalakrishnan (subhadrg)** [1]
Wren Dsilva (wdsilva) [1]

1. Abstract

The project focuses on studying different approaches that could eventually help in the enhancement of the privacy of the user while using voice assistant systems or AI based dialogue exchange units. The goal of the project is to take an input of speech segment (i.e. audio file) and transform the voice in a way that makes it highly intelligible i.e. easy for a listener to understand each individual word, while ensuring that a person or machine listening to the transformed audio cannot determine who the original speaker was. The major task that we are trying to accomplish in the project is speaker de-identification to hide the identity of the speaker while transmitting the context of the speech and maintaining privacy by hiding or redacting sensitive information. Our problem statement consists of three major subsections- Automatic speech recognition model, Text to Speech generator model and hiding of sensitive information from the transcripts generated from the ASR module. In our work, we have designed and implemented an ASR+TTS module for voice conversion and compared to an end to end model for multi-speaker inputs.

2. Introduction

Recent research have shown interest in the user de-identification model with the increase in the development of applications utilizing conversational AI and use of speech as command to operate multiple electronic devices including smart home, vehicle navigation, smart vehicles, smart phones etc.

Voice conversion (VC) refers to digital cloning of a person's voice; it can be used to modify audio waveform so that it appear as if spoken by someone else (target) than the original speaker (source). VC is useful in many applications, such as customizing audio book and avatar voices, dubbing, movie industry, teleconferencing, singing voice modification, voice restoration after surgery, and cloning of voices of historical persons. Since VC technology involves identity conversion, it can also be used to protect the privacy of the individual in social media and sensitive interviews, for instance. For the same reason, VC also enables spoofing (fooling) voice biometric systems and has therefore potential security implications.(Zhao et al., 2020)

As the demand and applications for smart devices increases, the utilization of devices being operated on user command would increase. Since, speech has been the most interactive form of communication between humans, any applications based on speech command get easily adapted in the market. With the increase in such applications, there is an increase of threat to the data privacy. These speech-operated devices utilize trigger command to activate but there isn't an established way to stop the device from listening content we didn't intend it to listen. Most of these data are eventually collected to make the speech based models robust to the diversity of accent, dialect, and voice pitch.

It is crucial to address the user data privacy at the early stage of production to ensure the ethical usage of data. Our problem statement is motivated by this research idea of user de-identification and we aimed not only to remove the features inherited in the voice of the individual but also the content that could be deemed as sensitive information.

We begin the report with background details and the related works to propose our method. Following it, we have described the proposed approach, explaining the overall pipeline and the architectures which we adopted to build our user de-identification system. The next section following that is the experimental analysis and results after which we end the report by addressing some of the suggested future work and concluding our research project.

3. Related works

Our inspiration for this study was from the Voice Conversion Challenge 2020 which dealt with the task of voice conversion in same language (English) as well as cross-lingual conversion (English to German, Mandarin and Finnish). (Huang et al., 2020) proposed a sequence-to-sequence (seq2seq) baseline system for this task. They used a cascade of ASR and TTS model using the ESPnet framework (Watanabe et al., 2018) which is a well-developed open-source end-to-end (E2E) speech processing toolkit. The system consists of three main modules: a speaker-independent ASR model, a separate speaker dependent TTS model for each

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University.

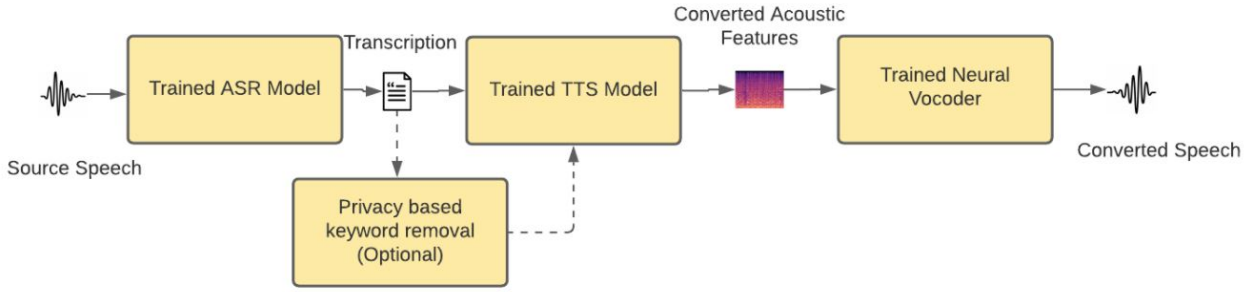

Figure 1. ASR-TTS Pipeline with Text Redaction

target speaker, and a neural vocoder that synthesizes the final speech waveform. The ASR module is a transformer based model trained on the LibriSpeech dataset and the TTS module is an x-vector + transformer based model trained on the LibriTTS dataset (Zen et al., 2019), finetuned on the target audio clips of the actual voice conversion dataset. The model is very similar in structure to the Transformer based Text to Speech model (Transformer-TTS), as proposed in (Shen et al., 2018), which is a combination of the Transformer Architecture and the Tacotron Text to Speech system. For implementation of voice conversion, the input text is replaced with the source voice as mel-spectrogram data and the output is the predicted mel-spectrogram features of the target speaker. By using the Transformer-TTS pretrained model and training for this new problem, the authors are able to achieve good performance. (Tobing et al., 2020) presents a cyclic variational autoencoder-based voice conversion system with parallel WaveGAN-based vocoder. CycleVAE is a nonparallel VAE-based voice conversion that utilizes converted acoustic features to consider cyclically reconstructed spectra during optimization. On the other hand, PWG is a non-autoregressive neural vocoder that is based on a generative adversarial network for a high-quality and fast waveform generator.

4. Proposed Approach

We build a baseline end-to-end trainable multi-speaker input to a single target source voice conversion system based on the transformer TTS architecture and compare the results of this model with another such model that performs multi-speaker to target voice speaker conversion, by first utilizing an automatic speech recognizer based on DeepSpeech 2 (Amodei et al., 2016) Architecture to convert input audio to text, and then applying a text-to-speech converter based on Tacotron 2 (Shen et al., 2018) architecture which converts the output text obtained from the speech recognizer to audio voiced by the target speaker. In between the automatic speech recognizer and the text-to-speech engine lies the NLP based sensitive data removal module which removes sensitive content like names, phone numbers, credit card information, and street address from the transcript output from the speech recognizer. This sanitized transcript is fed as input to the text-to-speech model.

4.1. Baseline Model

The baseline model (Huang et al., 2020) is an end to end trainable voice conversion system based on the Transformer TTS architecture. Instead of word/character embeddings, the input is now source mel-spectrograms. The ESPNet implementation only supports single source - single target mapping, we extend this to multispeaker by retraining with the multispeaker data from CMU Arctic dataset

4.2. ASR Model

The ASR model is based on the DeepSpeech2 (Amodei et al., 2016) architecture which has up to 11 layers including many bidirectional recurrent layers and convolutional layers. These models have nearly 8 times the amount of computation per data example as the models in Deep Speech 1 making fast optimization and computation critical. In order to optimize these models successfully, Batch Normalization is used for RNNs and a novel optimization curriculum called SortaGrad. They also exploit long strides between RNN inputs to reduce computation per example by a factor of 3.

4.3. TTS Model

The TTS model is based on the Tacotron 2 (Shen et al., 2018) architecture with LSTM based encoder and decoder systems connected by locally sensitive attention blocks. The model consists of two components (1) a recurrent sequence-to-sequence feature prediction network with attention which predicts a sequence of mel spectrogram frames from an input character sequence, and (2) a modified version of

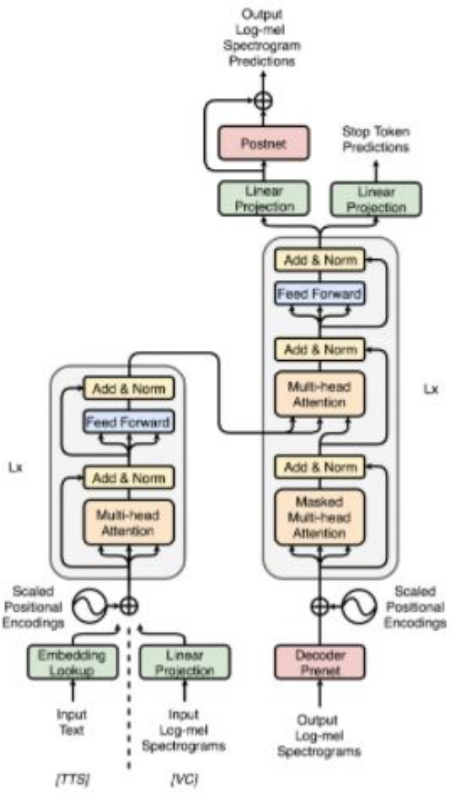

Figure 2. Baseline Model

WaveNet which generates time-domain waveform samples conditioned on the predicted mel spectrogram frames.

4.4. Sensitive Text Redaction

We use the text reduction model as proposed in (Yang & Liang, 2018), which redacts sensitive data such as credit card number, name, phone number, URL, street address, and email addresses. It allows custom rules to be added as well. The model is built by using an NLP based approach which classifies the each token in the text as either sensitive or non-sensitive by passing the text through four modules:

- Preprocess: This phase includes stop word removal, lemma recovery and part-of-speech tagging
- Analysis : Use context surrounding the word to classifying to one of the predefined topic classes
- analysis : Establish relationships between noun phrases and verb phrases from the syntax tree
- analysis : Use semantic analysis techniques like sentiment analysis to generate the probability of the word being sensitive or not.

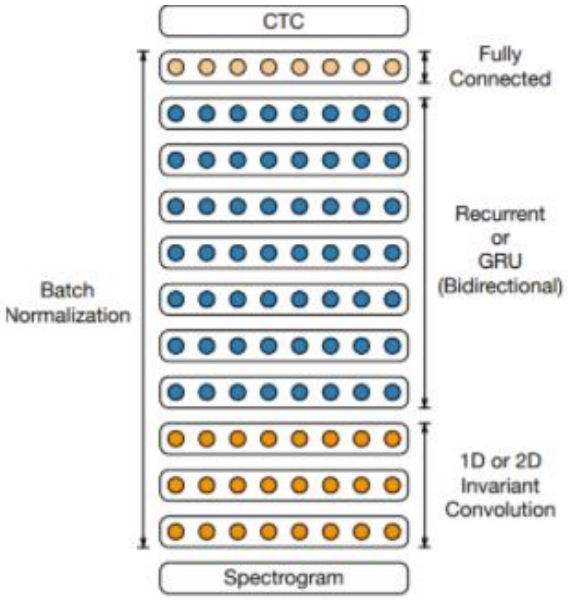

Figure 3. ASR Model

5. Experimentation

5.1. Dataset

There are multiple datasets available publicly for different speech related tasks and we explored WSJ (Wall Street Journal), CMU arctic and VCC2018 dataset. Most of the ASR tasks have utilized WSJ dataset for speech recognition models due to the data size and ease of use. Although WSJ is widely used and popular among multiple ASR Tasks, for our project we used the CMU Arctic dataset and the LibriSpeech dataset.

The brief overview of each dataset is provided below:

- CMU arctic (Kominek & Black, 2004) databases consist of around 1150 utterances carefully selected from out-of-copyright texts from Project Gutenberg. The databases include US English male and female speakers (both experienced voice talent) as well as other accented speakers. For training our multispeaker voice conversion model we used the slt (female) speaker as our target and clb (female), rms (male) and bdl (male) as our source speakers. The training set consisted of 932 single speaker target utterances paired with a random source speaker's corresponding utterance. The evaluation set was 100 utterances large and was used to calculate the final WER metric.
- LibriSpeech dataset: LibriSpeech (Zen et al., 2019) is a corpus of approximately 1000 hours of 16kHz read English speech. The training portion of the corpus is split into three subsets of approximately 100, 360 and

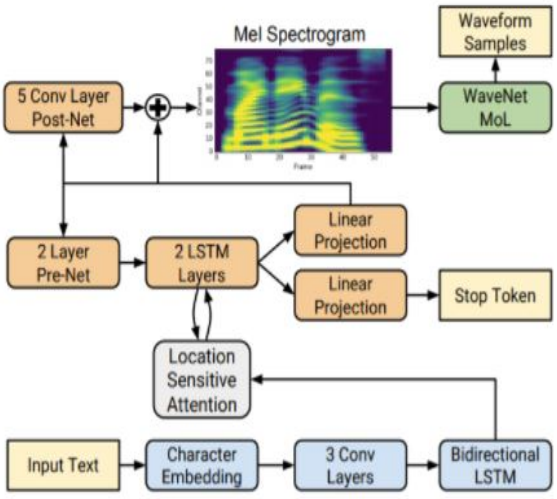

Figure 4. TTS Model

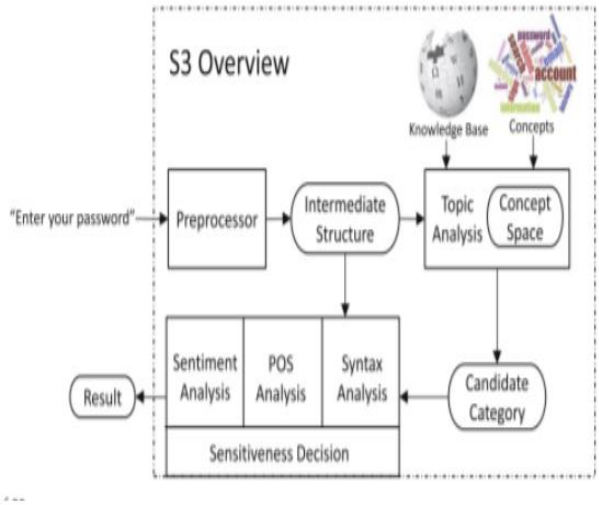

Figure 5. Sensitive Text Redaction Model

500 hours of audio respectively. The 100 hour train split consists of 126 male speakers and 125 female speakers. The 360 hour train split consists of 482 male speakers and 439 female speakers, and the 500 hour split consists of 602 male speakers and 564 female speakers.

5.2. End-to-End Voice Conversion Baseline using ESPNet

The VC Method on the Arctic Dataset given by (Huang et al., 2020) is provided in the ESPNet toolbox. It uses a single speaker TTS model trained on the M-AILABS speech dataset as a pretrained network to fine tune with the 832 utterances given in the CMU Arctic dataset. As mentioned in the previous section, the model structure is similar to a Transformer + Tacotron based TTS model with an encoder and a decoder, to predict the final target mel spectrogram data. We retrain this model with multi-speaker input from the CMU Arctic dataset by pairing the 932 utterances from the target speaker to random utterances of the source speaker for 2000 epochs to generate audio voiced with the target speaker.

Evaluation:

5.3. ASR

The ASR model is based on the Deep Speech 2 (Amodei et al., 2016) architecture. The model was trained on 11,940 hours of labeled speech containing 8 million utterances. The model was benchmarked to have a 4.42% WER on read Wall Street Journal (WSJ) corpus. We evaluate this ASR model on the LibriSpeech dataset.

Evaluation: The average Word error was computed on the test split of the LibriSpeech dataset consisting of 5 hours of audio with 17 male speakers and 16 female speakers.

5.4. TTS

We retrain the pretrained TTS Tacotron 2 (Shen et al., 2018) based model on the CMU arctic dataset for the target speaker by replacing the original wavenet vocoder[4] by a Parallel Wavegan to obtain a model consistent with the baseline voice conversion model for comparison. The monotonic attention obtained during training can be seen in fig. 6.

Evaluation:

5.5. Sensitive Text Redaction

We use the sensitive text redaction model as proposed in (Yang & Liang, 2018). Given any input text this model can redact multiple classes of sensitive text such as name, phone number, street address, url, credit card number, and url. It allows for custom rules to be defined as well. Consider the task of replacing all sensitive information in any given text input with the token 'BLANK'.
If the input text is : "David drives a car. His phone number is 412-555-5555. His portfolio site is https://github.com/david12"
The output after text redaction is : "BLANK drives a car. His phone number is BLANK. His portfolio site is BLANK"

Evaluation: The model is evaluated using precision, recall and F-score.

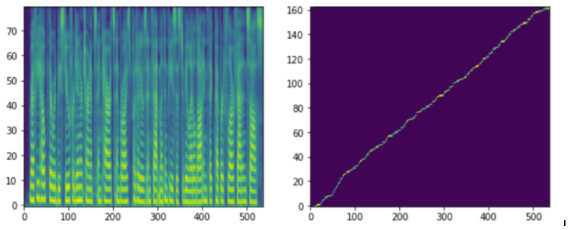

Figure 6. Waveform and Attention generated during TTS training

6. Results

On comparing the outputs obtained from the multi-speaker source to single target speaker end-to-end baseline approach, retrained for 2000 epochs on multi-speaker CMU arctic input with the two step, first ASR, then TTS approach, we see that the approach of applying ASR followed by TTS significantly outperformed the end-to-end voice conversion baseline model.

Avg WER for Baseline	Avg WER for our Model
0.53	0.17

For the purpose of evaluation, we considered WER of the final audio generated by both the basline and ASR-TTS model and compared their performance using the DeepSpeech2 ASR model.
The WER of Deepspeech2 ASR model on Librispeech dataset is 5.33. Our experiments saw that the majority of the error occuring in the ASR-TTS model for voice conversion were introduced through the ASR model itself. We also did subjective evaluation of the audio files whose paths are being attached here:

Sample outputs generated
For the ground truth transcript: "JACOB BRINKER, WHO WAS HIS ROADMATE, BROUGHT THE NEWS" of the Source audio. The following outputs were generated:

Output of Baseline End-to-End Voice Conversion: End-to-End-VC
Output of ASR followed by TTS: ASR-TTS
Output ASR followed by text-redaction followed by TTS: ASR-Redaction-TTS

7. Conclusion and Future Directions

Our results in this project show that it is possible to create a highly accurate user de-identification voice conversion model while redacting sensitive information contained in the input speech by applying a two step approach of first applying automatic speech recognition and performing NLP based text redaction on the transcript obtained, followed by an application of Text-to-Speech conversion. We also showed that this approach is much more effective than applying a direct end-to-end voice conversion model. Another major observation was that of the errors seen in the output audio, the automatic speech recognition contributed significantly higher to the errors than the text-to-speech conversion. Future work in the direction of building an automated speech recognition system with a much lower word-error rate will significantly improve the accuracy of our model.

References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pp. 173–182, 2016.

Huang, W.-C., Hayashi, T., Watanabe, S., and Toda, T. The sequence-to-sequence baseline for the voice conversion challenge 2020: Cascading asr and tts. *arXiv preprint arXiv:2010.02434*, 2020.

Kominek, J. and Black, A. W. The cmu arctic speech databases. In *Fifth ISCA workshop on speech synthesis*, 2004.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783. IEEE, 2018.

Tobing, P. L., Wu, Y.-C., and Toda, T. Baseline system of voice conversion challenge 2020 with cyclic variational autoencoder and parallel wavegan. *arXiv preprint arXiv:2010.04429*, 2020.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., et al. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*, 2018.

Yang, Z. and Liang, Z. Automated identification of sensitive data from implicit user specification. 2018.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

Zhao, Y., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., Ling, Z., and Toda, T. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. *arXiv preprint arXiv:2008.12527*, 2020.