CS 753 : Automatic Speech Recognition Project

Voice Conversion using GANs

Varun Bhatt (140260004)
Arka Sadhu (140070011)
Karan Taneja (15D070022)

Problem Statement

We are given a audio input from speaker '*s*' from which we obtain the spectral frames $X_s = \{x_{s,n}\}_{n=1}^{Ns}$ where N_s is the number of spectral frames.

We are also given a target speaker '*t*' for which we have to generate spectral frames $X_t = \{x_{t,n}\}_{n=1}^{Nt}$ and therefore the audio output.

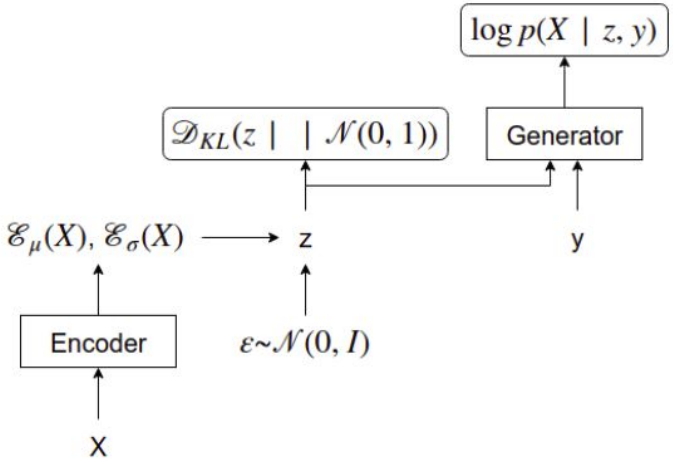
The audio output should be such that it sounds like target speaker '*t*' speaking the linguistic content of X_s.

The most common way to evaluate voice conversion systems currently is to get subjective opinion from listeners.

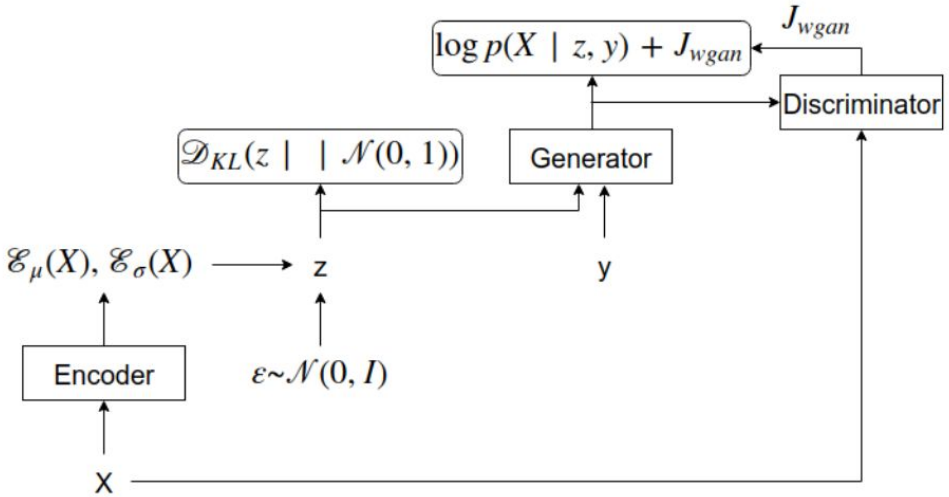
Related Work

- Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. **Voice conversion from non-parallel corpora using variational auto-encoder**, 2016.
- Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. **Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks**, 2017.
- Sanjeev Arora, Yingyu Liang, and Tengyu Ma. **A simple but tough-to-beat baseline for sentence embeddings**. 2017.

Architecture of C-VAE and VAWGAN

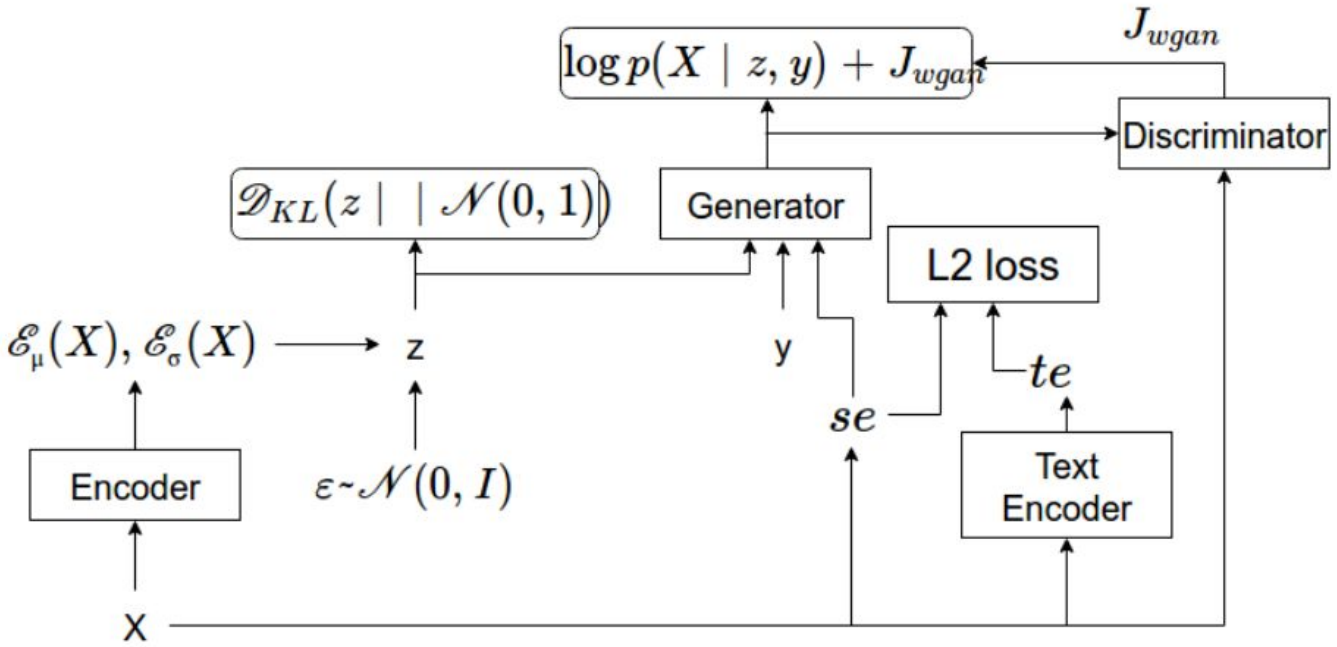


C-VAE

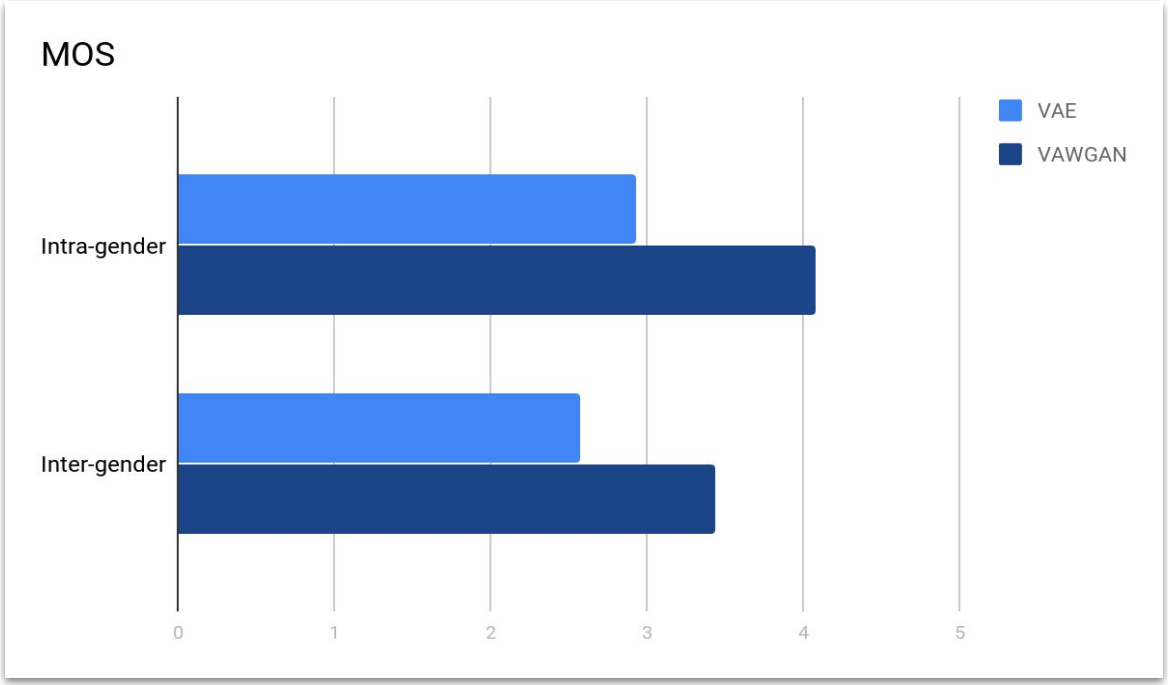


VAWGAN

Sentence Embedding: VAWGAN-S

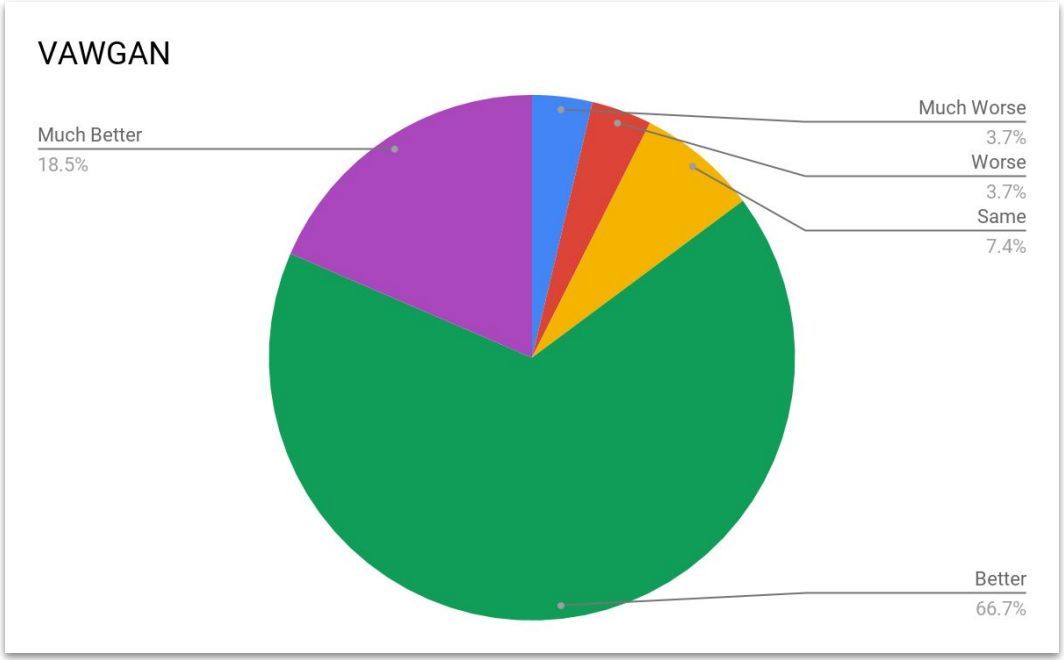


Experiment Setup and Results



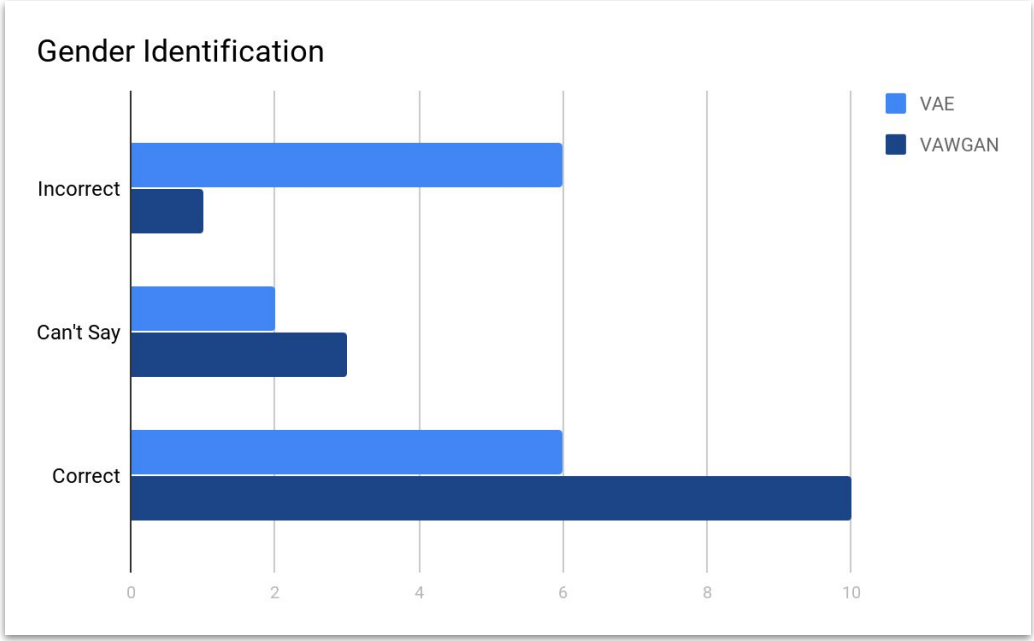
Mean Opinion Scores for the VAE Baseline and VAWGAN

Experiment Setup and Results



Difference between simple VAE baseline and VAWGAN

Experiment Setup and Results



Gender Identification correctness for VAE Baseline and VAWGAN

Discussion

- VAWGAN clearly outperforms VAE baseline model in both inter-gender and intra-gender conversion tasks. Inter-gender in general more difficult than intra-gender.
- Overall VAWGAN ratings given are better than VAE baseline which is consistent with the results presented in paper.
- A lot of gender identification errors are made in inter-gender conversion in VAE baseline model. VAWGAN reduced the error rate a lot.
- VCC 2018 has much longer sentences as compared to VCC 2016 and it directly affects conversion, decreasing the MOS.

Conclusion

- The results of voice conversion using VAE and WGAN given in the paper have been replicated.
- A novel method of extending it by using sentence embeddings as a representation of content has been developed.
- In future, a better speaker representation like i-vectors can be used for improving encoding and generation.