GANs for Speech Recognition Data Augmentation

UNIVERSITY OF CAMBRIDGE

Tianyu Wu

Supervised by Phil Woodland & Chao Zhang

Objectives

Investigate the use of Generative Adversarial Networks (GANs) for acoustic data generation given an initial small training set.

- Build a trainable generator for phone units and context windows.
- Develop the training pipeline to improve the pre-trained acoustic model's performance with augmented data.

Background

Performance of a speech recogniser improves with increased training data size, **but**

- collecting a large matched training dataset is difficult,
- manually transcribing speech data is expensive.

Generative Adversarial Networks are a method which can generate simulated data [1].

Methodology

- Perform frame level speech data generation.
- Train separate GANs for each phonetic unit.
- Continue training the pre-trained acoustic model with augmented data (supervised training).

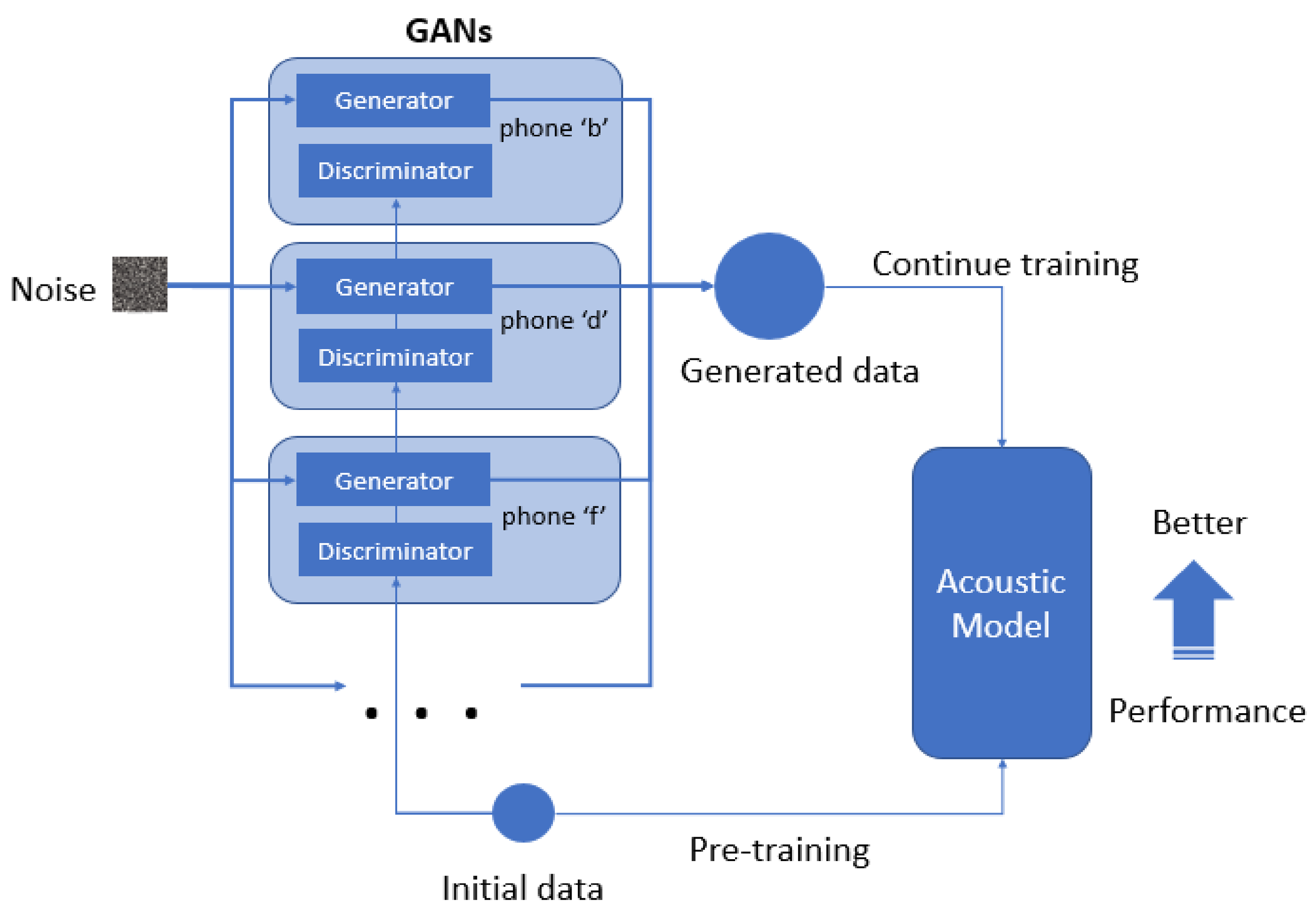


Figure 1: Architecture of whole system

Generative Adversarial Networks

GANs are a framework to estimate generative models via an adversarial-process:

- Discriminator is trained to perform classification between true data and fake data.
- Generator is trained to generate fake samples to fool the discriminator.

The configuration of our GAN:

- Use deep convolutional structure.
- Use spectral normalization technique [2].
- Condition the model on phone states [3].
- Generate data on the speech feature level (FBANK).

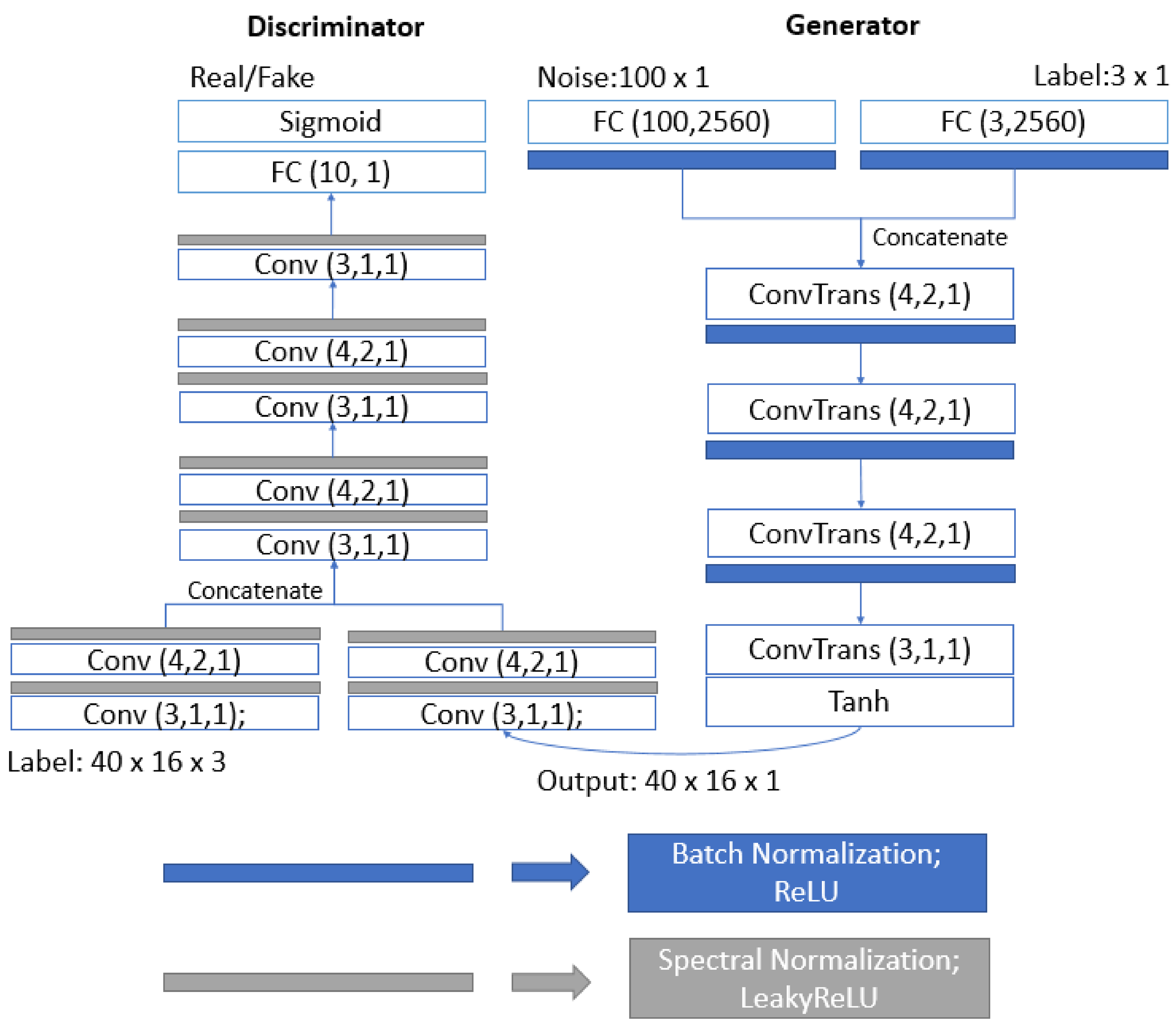
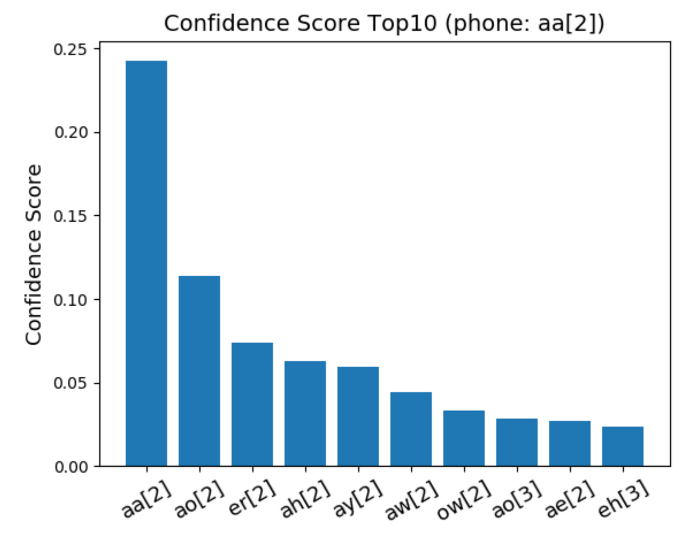


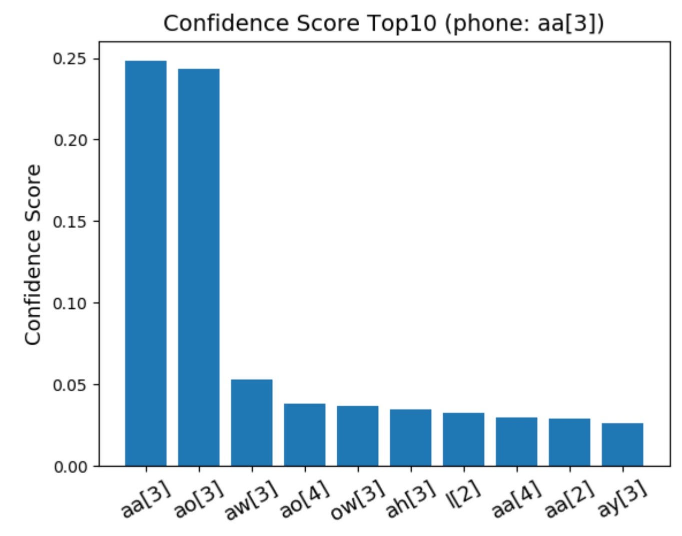
Figure 2: Structure of proposed GAN. Conv(4,2,1) means that the Convolutional layer uses a 4x4 kernel with stride 2 and zero padding 1.

Generated Fake Data

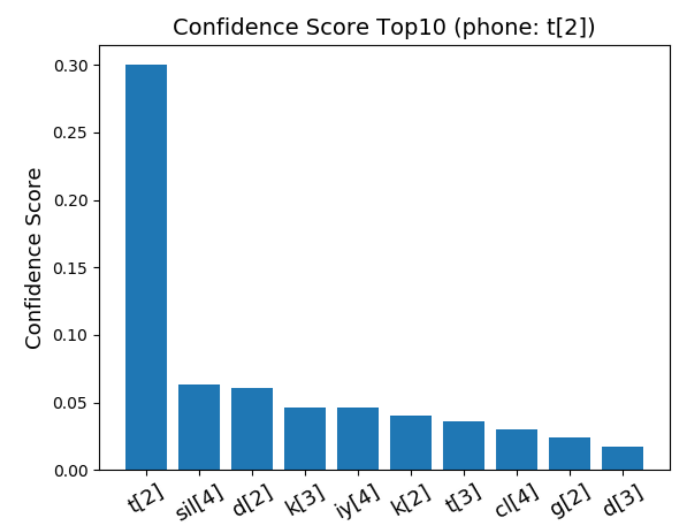
Evaluate the generated data based on the pre-trained acoustic model.



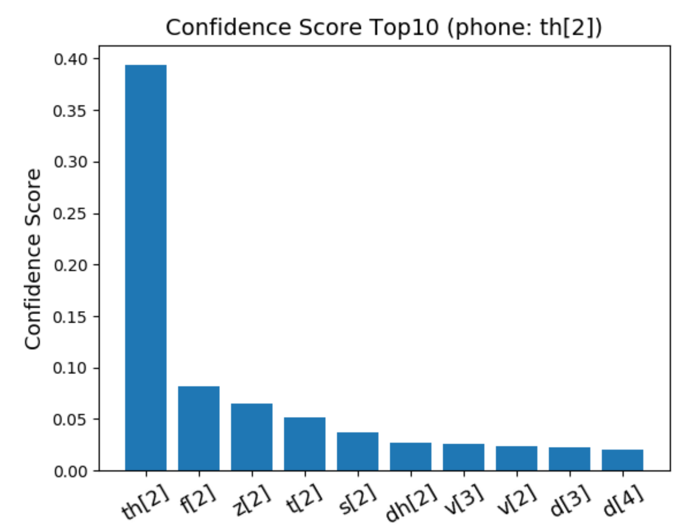
(a) **Fidelity test for fake phone aa[2]**



(b) **Fidelity test for fake phone aa[3]**



(c) **Fidelity test for fake phone t[2]**



(d) **Fidelity test for fake phone th[2]**

Table 1: Top1, Top3 and Top5 Classification Accuracy

Phone	Generated fake data			Test set data		
	%top1	%top3	%top5	%top1	%top3	%top5
aa[2]	30.2	60.7	74.1	43.6	82.1	91.4
aa[3]	28.8	60.8	71.3	52.0	87.8	95.0
t[2]	35.2	57.8	68.0	71.5	90.4	95.4
th[2]	45.1	69.6	78.2	42.8	69.6	78.2

Upcoming Work

- Improve acoustic model performance with augmented data.
- Extend to triphone system.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," pp. 1–9, 2014.

[2] T. Miyato, K. Toshiki, K. Masanori, and Y. Yuichi, "Spectral Normalization For Generative Adversarial Networks," 2018.

[3] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," pp. 1–7, 2014.