Voice Impersonation Using Generative Adversarial Networks

Yang Gao*, Rita Singh, Bhiksha Raj
yanggao, rsingh, bhiksha@cs.cmu.edu
Electrical and Computer Engineering Department, Carnegie Mellon University

Introduction

- Voice Impersonation is a challenge problem requires convincingly convey the impression of having been naturally produced by the target speaker.
- Common voice transformation methods modifies the instantaneous characteristics of a source signal, such as pitch and spectral envelope. When trained, they are heavily reliant on the availability of *parallel* recordings of the source and target utterances.
- These methods are generally insufficient to capture unmeasurable, unquantifiable *style* in the general sense of the word.
- **Our goal: Learning style transformed voice impersonation model using unparalleled dataset and a *discriminate* learning mechanism.**

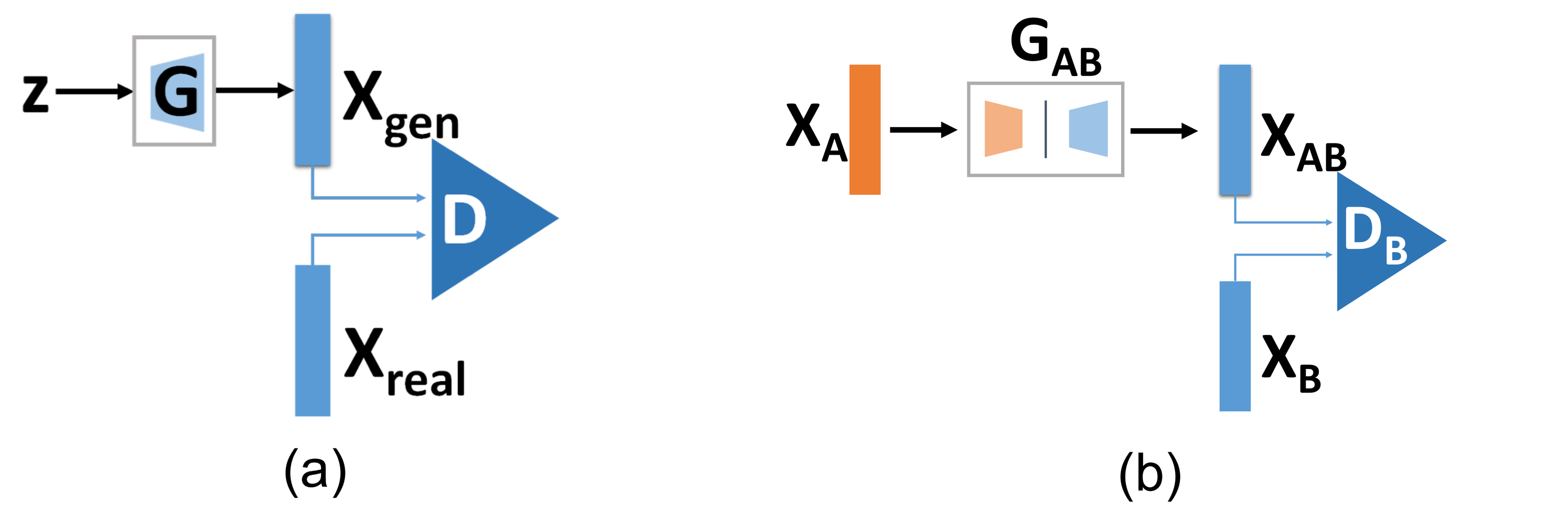

Fig. 1 A discriminative learning model: Generative Adversarial Network (GAN) (a) The architecture of the original GAN model; (b) Style transfer by GAN.

Our contributions

- We propose a **Generative Adversarial Network based Voice Impersonation Model (VoiceGAN)** to specifically address the end-to-end voice impersonation task.
- It can generate convincing samples of impersonated voice while intrinsically addressing the problem of speech durational variability.

VoiceGAN model

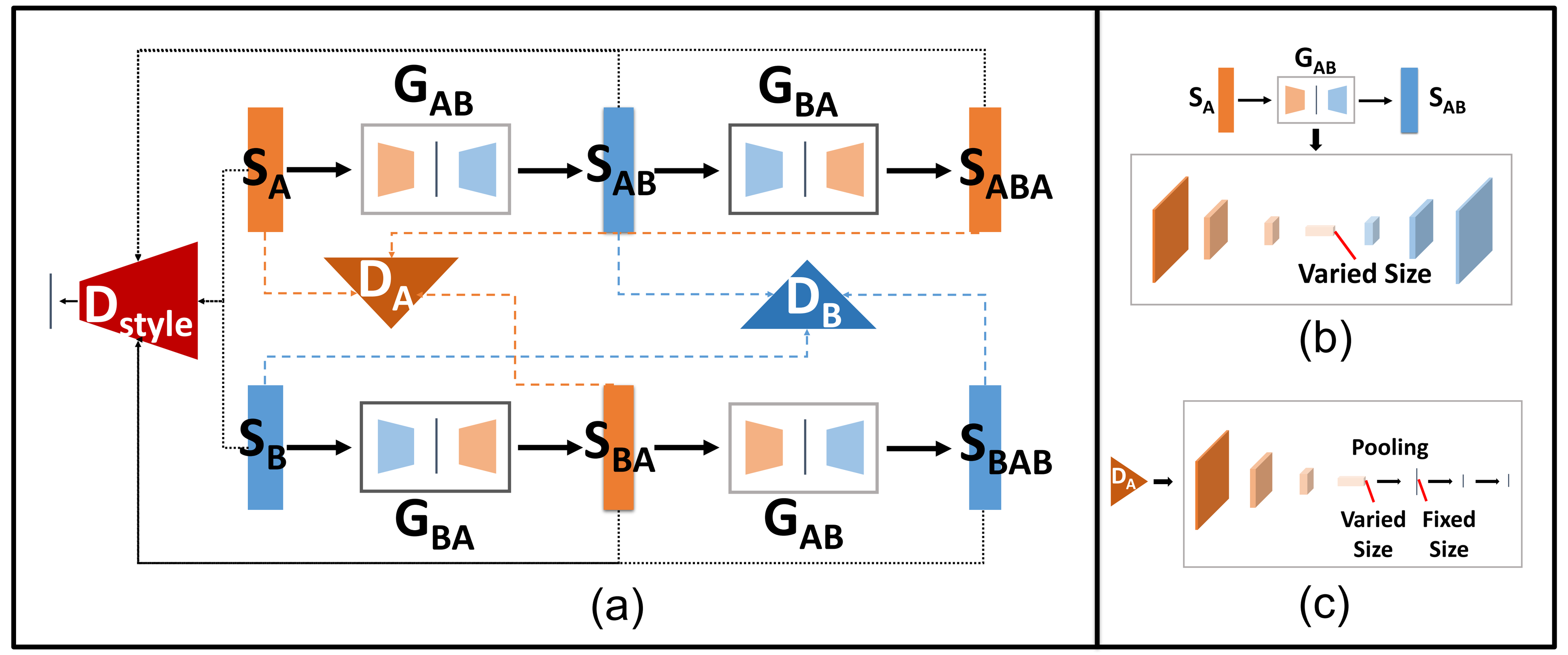

Fig. 2 (a) The architecture of the proposed VoiceGAN model; (b) Visualization of Generator G_A; (c) Visualization of Discriminator D_A .

- Retaining Linguistic Information
  - We modify our reconstruction loss as:

$$L_{CONST_A} = \alpha \cdot d(x_{ABA}, x_A) + \beta \cdot d(x_{AB}, x_A)$$

  - The term $d(x_{AB}, x_A)$ attempts to retain the structure of x_{AB} to keep the linguistic structure as x_A .
- Variable-length Input Generator and Discriminator
  - As shown in Fig 2 (b), the generator is of fully convolutional structure so it can handle varied length inputs.
  - As shown in Fig 2 (c), the discriminator has an adaptive pooling layer after the CNN layers and before the fully connected layers. This is a channel-wise pooling in which each channel's feature map is pooled into a single element. This conveys any variable-sized feature map into a vector of a fixed number of dimensions, with as many components as the number of channels.
- Style Embedding Models (D_S)
  - We add a second type of discriminator to our model to further extract the target style information.
  - The discriminator *Ds* determines if the original and transformed signals match the desired style.

Experimental results

- Visualization of Generated Results

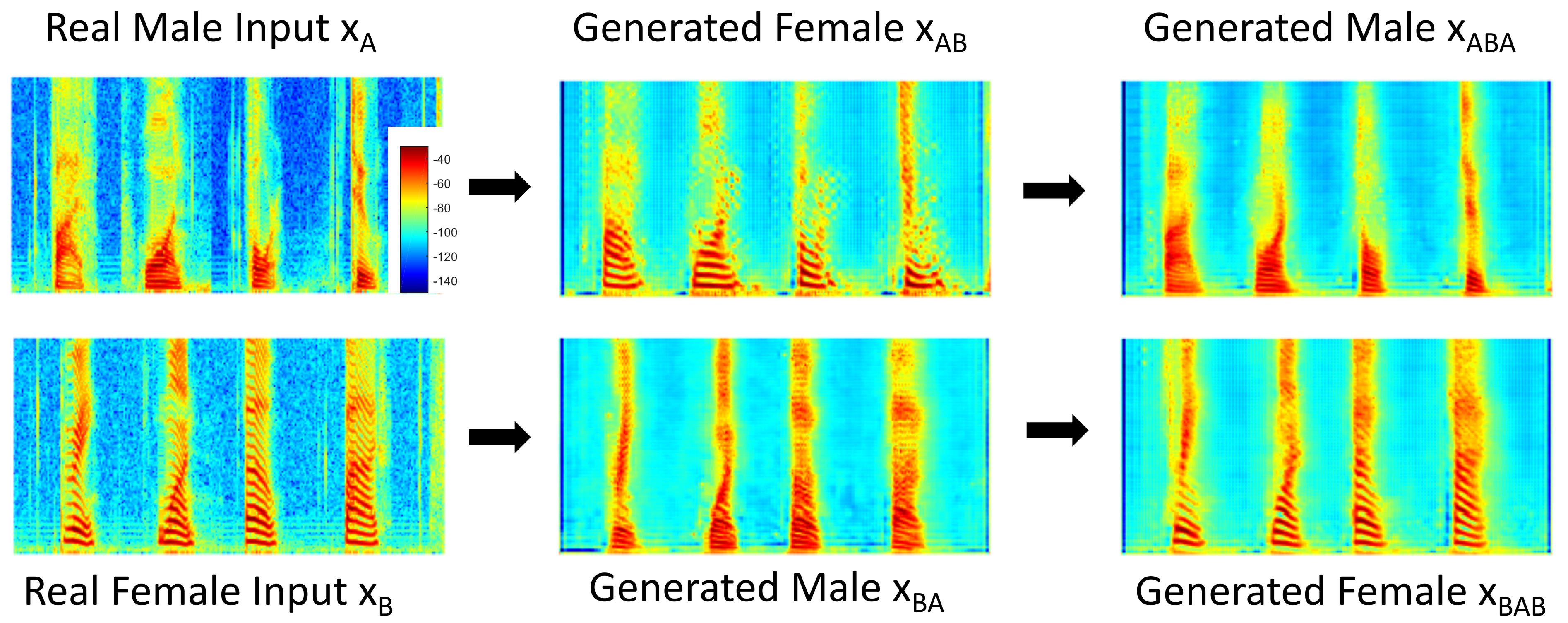

Fig. 2 Visualization of spectrograms generated from a speaker saying "3 1 oh 5" (first row) and "5 1 4 2" (second row). For each spectrogram, frequencies on the y-axis range from 0-4 kHz.

- Style Classification Test
  - We use an independently-trained CNN-based classifier to predict the style of our generated data. The classifier was trained on 800 utterances from speakers of both genders.
  - The results show that **100%** of indicates that our VoiceGAN network achieves **good style transfer** the generated data are classified as the target speaker's style, which performance.
- Speech Signal to Noise Ratio (SNR) test

Data (use GL-method)	A (*dB*)	B(*dB*)
Original signal	55.60±4.97	52.91±3.58
X_A and X_B	54.97±6.28	52.15±3.70
X_{AB} and X_{BA}	49.64±1.80	49.92±4.36
X_{ABA} and X_{BAB}	53.58±2.69	50.05±2.12

Table. 1 NISTSTNR TEST