Fredric Hsieh and Stephen Zhong

Leah Ding

CSC-481

26 April 2023

Evaluating Audio Conversion and Reversal

* Problem/Motivation
  + Cyber security issues: New audio conversion technology allows impersonation ending up with potential fraud or identity theft
  + Criminals can disguise their voices as other people in order to protect their identity from investigators
  + Misinformation can be shared through impersonation of someone trustworthy
* Related Works Review
  + AGAIN-VC: One Shot Voice Conversion
    - A recently developed version of the AGAIN tool (Activation Guidance and Adaptive Instance Normalization) which focuses upon voice conversion. Many voice conversion systems focus upon obscuring original speaker information, which the authors claim damages their sound quality, since speech signals are separated from information about the original speaker, forcing the tool to generate a voice on its own. In contrast, AGAIN-VC is an auto-encoder based model that does not seek to disentangle original information, and thus is able to achieve higher sound quality than comparable systems, speaker similarity is higher but synthesis quality is drastically improved.
  + Voice-AI
    - Focuses entirely upon converting a voice to a predetermined voice. We did not use it in research so we don’t know much about its inner mechanisms nor its limitations. It is used in our presentation as a demonstration of the power of voice changing. We would theorize that a tool like this would suffer in terms of obscuring speaker information and also requires a large quantity of target speaker speech, since it advertises its ability to convert a voice to that of celebrities or other well-known public features. It also has the ability to generate new voices, but this seems to be simple audio manipulation (pitch, tone etc.) rather than through machine learning methods.
  + CATCH YOU AND I CAN: Revealing Source Voiceprint Against Voice Conversion
    - We did not have source code access to this paper, so we only know what is outlined in the paper nor did we have the chance to experiment with its capabilities ourselves. The paper outlines the creation of REVELIO, a representation learning model which extracts the voiceprint of a speaker from a converted audio signal. The main part of this process is a “differential rectification algorithm”, which effectively seeks to eliminate as much representation in the voice clip as possible from the target speaker, while maximizing the influence of the original speaker. This is accomplished through the use of a “residual orthogonal block”, which splits components of the audio clip in order to find the relative importance of features in eliminating influence of the target speaker. The authors claim that the tool is effective in creating recognizable voices from voice clips that were converted through some popular machine learning methods, although as mentioned previously we were not able to test this ourselves.
  + Fraud through Voice-Cloning
    - This source is a simple news article which we used to prove that voice imitation is a real problem. It describes a fraud committed in 2021 in which the criminals impersonated a bank director in order to steal hundreds of thousands of dollars.
* Dataset Description
  + We used the existing dataset of the AGAIN-VC paper, for ease of use and access, as well as to make sure that data would work properly with it. The data consists of 4 people’s audio clips, between 2 and 10 seconds each. Each person has 10 audio clips of them saying various short sentences.
* Approach
  + Evaluation of capabilities of the AGAIN-VC tool
    - Can machine learning models differentiate between original and converted voices?
    - Can the original or converted speaker be identified following conversion?
      * How accurately can machine learning models determine original speakers following conversion?
  + Determining feasibility of reversal tools aimed at machine-learning based sound conversion
    - Determine an effective way to turn an audio clip back to its original form, with minimal information provided
      * Gain as much information possible from the smallest amount of provided information (preferably only the provided converted sound clip)
* Results
  + Distinguishing Converted and Original Voice Clips
    - Humans are able to distinguish original and converted voice clips effectively (sample size was small for this since we do not have an effective way to get this survey out to a large number of people)
    - Qualitatively, the more the original and target speaker differ, the lower the synthesis quality, in particular, accents or sex differences resulting in large differences in voice pitch cause lowered synthesis quality
    - Machine learning methods (a conventional neural network in our case) are also able to effectively distinguish the two classes. Even without much tuning of hyperparameters, it achieves over 80% accuracy in classifying the original and converted voices.
    - This could allow automation of determining conversion in audio clips, freeing up human resources.
  + Determining original speaker from converted voice clip
    - Humans are unable to accurately determine the original speaker from a converted voice clip
    - Random Forest
      * Random forest classifier was unable to accurately identify the original speaker from a converted voice clip after training upon converted voice data
    - CNN
      * CNN training on only original voice data is able to effectively identify original voices from converted voices clips, with 82% accuracy
      * The mistakes made in classification are primarily on the converted clips of p228, a woman. We hypothesize that her voice possesses features that are somehow overwritten by the conversion or are not weighted by the classifier and thus is mistakenly classified as one of the other speakers.
  + Reversion
    - We assumed that we possessed clips of the original voice speaker, which are required to train the model to recognize the original speaker.
    - Through classification as a converted clip then classification of the original speaker, the converted audio clip can be reverted to the original speaker by running it through AGAIN-VC again with the target speaker set as the original speaker
    - The twice-converted clip is recognizable as the original speaker, indicating success in our methods
* Discussion
  + Conversion
    - Voice conversion machine learning models are distinguishable relatively easily by humans, even those which are focused upon speech quality
      * They remain relatively effective against machine learning models that have not been trained on converted data, although models trained on converted data are able to distinguish them
    - AGAIN-VC’s focus upon audio quality at cost of speaker privacy possibly gives our classifiers an advantage compared to other voice converters, but those would be more easily distinguishable by humans and thus less useful
  + Reversal
    - Our set of tools identifying converted audio and original speaker are able to effectively reverse audio with no extra human input
    - However, our assumption that we possess original speaker data may not necessarily hold true under many circumstances and small amounts of speaker data or large numbers of potential speakers may cause issues
    - More sophisticated models are possible as demonstrated by REVELIO but these require original speaker data to remain in converted audio, making them potentially able to be circumvented by a more advanced voice conversion model
  + Future Steps
    - Additional training could benefit our models, our sample sizes are small and do not include converted voices in the case of the classifier intended to find the original voice. Including these data points may allow greater accuracy
    - Focusing upon pulling data from the converted clip alone, focusing upon things like pulling information of the original speaker like REVELIO, or identifying conversion without knowledge of the model used, may allow us to successfully reverse voices without the requirement of so much information on the original speaker and model