Эссе-отчет по аблационному эксперименту с kNN-VC

«Конвертация голоса с таргетами разной длины»

Введение

В ходе лабораторной работы была исследована модель голосовой конверсии kNN-VC . Цель эксперимента — установить, как влияет длина таргет-сэмпла на качество преобразования голоса и в какой момент исчезают нежелательные особенности, такие как акцент исходного говорящего.

Методика эксперимента

Для эксперимента были подготовлены исходный голос (запись эмоционально прочитанного монолога Быковы из сериала Интерны `source_voice.wav`) и таргет-голос (спокойная речь из одного из интервью Охлобыстина И.И.), нарезанный на сэмплы длиной 10, 20, 30, 40, 50 и 60 секунд соответственно.
С помощью предоставленного скрипта выполнена конвертация исходной записи с использованием каждого из таргет-сэмплов.
Стоит заметить, что таргет - это спокойный голос, в то время как copc - весьма эмоциональный. Но даже с этими различиями конвертация довольно хорошо звучит. Есть и более спокойные версии copc (source_voice.wav и source_voice_2.wav) - с ними модель справляется еще лучше.

Результаты и наблюдения

- 10 секунд:
Преобразование сохраняет много фонетических особенностей моего голоса, однако уже слышится картавость таргета. Четко прослушивается моя интонация, хотя тембр уже похож на таргет.Подобный короткий таргет излишне «копирует» стиль таргета, но плохо стирает признаки source. Итог: преобразование слушается неестественно, явно слышны искажения речи и неравномерность произнесения слов.

- 20 секунд:
Изменения становятся чуть более выраженными в сторону target-голоса. Тем не менее, некоторые интонационные паттерны и акцент сохраняются. Тембр копируется сильнее и субъект почти не узнается.

- 30 секунд:
Качество возрастает — мой акцент и особенности становятся менее заметны. Тембр и манера речи все больше напоминает таргет, но при определенных звуках еще можно услышать отголоски исходника.

- 40 секунд:

Голос звучит практически как у Быкова, особенности моей речи стерты сильнее. Акцент таргета преобладает. Картавость более ярко выражена.

- 50–60 секунд:
На этой длине таргета достигается лучшее качество: исчезли практически все индивидуальные особенности source, полностью принята манера и тембр таргет-голоса. Акцент, дикция — максимально близки к target-голосу, при сохранении первоначального текста.

Вывод и обсуждение

kNN-VC действительно чувствительна к длине target-файла. Короткие записи приводят к тому, что на выходе слишком хорошо сохраняются манера, акцент и другие особенности source и речь звучит неестественно — это особенно заметно при конвертации между разными языками или сильно разными акцентами. При увеличении длины таргета смешение признаков становится более сбалансированным, и лучшие результаты достигаются при 40–60 секундах по моим наблюдениям.
Оптимальной длиной target, дающей отчетливо «новый» голос без артефактов исходного говорящего, считаю >= 40 секунд.

Персональное впечатление:
Возможности нейронок в обработке речи, конечно, поражают.

Приложение:
https://yandex.ru/video/preview/16818224242897272372