Конверсия голоса в задаче плавного перехода от автоматической диалоговой системы к оператору

Бибик Денис Андреевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: den-bibik@yandex.ru

Научный руководитель *— Соколов Евгений Андреевич*

В наше время во многих центрах обработки звонков происходит внедрение автоматических диалоговых систем, позволяющих автоматизировать диалог с клиентом. Однако, в независимости от того каким интеллектом обладает диалоговая система, могут возникнуть ситуации, когда необходимо осуществить переход от автоматической системы к оператору. Одна из идей бизнеса состоит в том, чтобы сделать этот процесс незаметным для пользователя. При этом возникает задача бесшовного переключения с голоса диалоговой системы на голос оператора.

Данная проблема может быть решена тремя способами: посредством синтеза голоса очень похожего на голос оператора, либо с помощью преобразования голоса оператора в голос похожий на голос диалоговой системы, либо с помощью комбинации этих двух подходов. Так, например, наиболее очевидным методом, вытекающим из второго подхода, является распознавание речи оператора (речь в текст) с последующим синтезом (текст в речь). Минусы такого метода очевидны: ошибки распознавания речи изменяют содержание текста, помимо этого возникает полная потеря информации об интонациях. Также данный метод создает большое время задержки.

В этой работе предложен метод с использованием свёрточных нейронных сетей, комбинирующий оба подхода. Суть предложенного метода заключается в применении алгоритма Artistic Style, предполагающего оптимизацию расстояния между весами нейронной сети по входным данным. В качестве функции генератора была использована предобученная нейронная сеть starGAN-VC, на вход которой подаются сырые аудиоданные (raw wav). При этом стилевая функция потерь оптимизировалась и по записи речи пользователя, и по записи синтезированной речи. В результате подбора параметров выяснилось, что оптимальный результат получается при использовании матриц Грама со всех свёрточных слоев, кроме последних трех. Также в результате экспериментов стало понятно, что для данной се-

ти использование контентной функции потерь не улучшает качество работы метода.

Предложенный в данной работе метод позволяет изменять голос оператора и голос диалоговой системы таким образом, что переключение между ними малозаметно. Для работы методу достаточно 150мс контекста и собранных статистик матрицы Грама голоса оператора и синтезированного голоса, что открывает возможности для использования алгоритма в реальном времени.

Литература

1. Stargan-vc: Non-parallel many-to-many voice conversion with star generative adversarial networks / Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, Nobukatsu Hojo // CoRR. — 2018.
2. A neural algorithm of artistic style. / Gatys Leon A., Ecker Alexander S., Bethge Matthias. // Nature Communications — 2015.
3. Transition conversations from bot to human. `https://docs.microsoft.com/en-us/azure/bot-service/bot-service-design-pattern-handoff-human?view=azure-bot-service-4.0`.