УДК 004.934.5

ПОДХОД К ОТБЕЛИВАНИЮ ЭКСПРЕССИВНОГО РЕЧЕВОГО СИГНАЛА СРЕДСТВАМИ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

Казакова С.А. (ИТМО)
Научный руководитель – кандидат технических наук Рыбин С.В. (ИТМО)

Введение. В данной работе рассмотрена группа методов аугментации данных, основанных на генеративных подходах, обозначенная как голосовые преобразования. Цель голосовых преобразований – изменить одну из основных характеристик аудиофрагмента (текст, голос, эмоцию, стиль), сохранив при этом остальные. Задача может быть решена применением модифицирующего ресинтеза, основанного, например, на архитектуре потока [1]. В рамках исследования был проведен ряд исследований о применимости такого подхода для удаления эмоций из экспрессивных образцов речи.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №623088 «Разработка русскоязычного персонифицированного эмоционального диалогового агента».

Основная часть. Основная задача генеративных моделей – создание новых образцов данных того же типа, что и данные из обучающего множества. Реализация этой идеи связана с концепцией генеральной совокупности (распределения данных) – предположением о существовании совместного распределения вероятностей, включающего бесконечное число пар признак-метка, в том числе образцы из обучающего и тестового множеств [2].

Потоки – это гибкие генеративные модели, реализующие отображение некоторого базового распределения на другое, обычно сложное распределение. Если базовое распределение является нормальным, то модель называется нормализующим потоком.

В процессе обучения модель получает случайные переменные из исходного распределения и выполняет некоторое нелинейное инвертируемое преобразование, зависящее от некоторых дополнительных параметров (например, вкраплений диктора, стиля или эмоций). Затем модель вычисляет функцию плотности вероятности. Учитывая обратимость этого преобразования, потоки представляются высокоэффективной архитектурой для задачи аугментации данных.

Принимая на вход нормальное распределение $N(0,1)$, вкрапления диктора, текст и другие необходимые параметры, поток генерирует признаки для вокодера. Если признаки эталонного образца отправить в поток в обратном направлении, их можно разобрать на составные части, а затем заменить один или несколько элементов. Если модель достаточно надежна и была обучена на большом количестве данных, повторное прохождение через тот же поток позволит получить новое акустическое представление для вокодера с контролируемыми изменениями желаемых характеристик. По сравнению с вариационными автокодировщиками (VAE) и генеративными адверсарными сетями (GAN), потоки дают более точную оценку плотности распределения данных, поскольку они обучаются путем максимизации их совместного логарифмического правдоподобия [3].

В рамках аналогичного подхода возможно реализовать систему, где модифицирующий ресинтез применяется исключительно для модификации экспрессивной составляющей, сохраняя параметры диктора и содержания речи. Подобный подход является решением актуальной задачи предобработки акустических данных для систем биометрической аутентификации, так как эмоциональные компоненты речи часто являются причиной ошибок второго рода в текстонезависимых системах [4].

Выводы. Проведен анализ подходов модифицирующего ресинтеза и разработан метод эмоционального отбеливания речевых данных.

Список использованных источников:

1. Kazakova S.A., Svischev A.N., Zorkina A.A., Rybin S.V., Kocharyan A.M. Expressive Audio Data Augmentation Based on Speech Synthesis and Deep Neural Networks // 2023 IEEE International Conference "Quality Management, Transport, and Information Security, Information Technologies" (IT&QM&IS), 2023, pp. 123–126.
2. Kevin P. Murphy. Probabilistic Machine Learning An Introduction. MIT Press, 2022. 828 p.
3. Rafael Valle, Kevin Shih, Ryan Prenger, Bryan Catanzaro. Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis // arXiv preprint arXiv:2005.05957. – 2020, pp. 1–10.
4. Rusko M., Trnka M., Sakhia D., Stelkens-Kobsch T., Finke M. Weaknesses of voice biometrics – sensitivity of Speaker verification to emotional arousal. // 25th International Congress on Sound and Vibration, 2018, pp. 1–8.