

АДАПТАЦИЯ СИСТЕМ АВТОМАТИЧЕСКОЙ ВЕРИФИКАЦИИ ДИКТОРА К ДОМЕНУ ШЕПОТНОЙ РЕЧИ

Хмелев Н.А. (ИТМО)

Научный руководитель – кандидат технических наук, Новоселов С.А.
(ИТМО)

Введение. Современные системы автоматической верификации диктора достигли значительных успехов в условиях работы с вокализованной речью, однако их эффективность резко снижается при обработке шепотной речи. Это обусловлено фундаментальными различиями в акустических характеристиках: шепот лишен основной частоты (F_0), имеет ограниченный динамический диапазон и искаженную формантную структуру, что нарушает работу стандартных алгоритмов автоматической верификации диктора.

Основная часть. В данной работе предложен комплексный подход для адаптации системы на базе wav2vec 2.0[1] к условиям шепотной речи. Для улучшения качества работы систем автоматической верификации диктора по голосу, используются следующие методы:

1. Обучение модели распознавания диктора с использованием аугментации данных синтезом шепота с использованием Praat[2].
2. Доменная нормализация оценок с применением детектора шепота на уровне дикторских эмбедингов.

Для оценки качества подходов используются как классические базы данных, такие как VoxCeleb1-O и VOiCES, а также специализированные корпуса шепотной речи (wTIMIT[3], CHAINS[4], MSP-AVW[5]). Для специализированных корпусов шепотной речи собраны протоколы тестирования, исключающие текстозависимые сравнения.

В результате оценки качества замечено, что качество адаптированных моделей в классических сценариях сравнений нормальной против нормальной речи не деградирует. Значение метрики EER в условиях сравнения шепота против шепота уменьшается с 9.97% до 8.17% и с 5.24% до 4.60% на данных MSP-AVW и wTIMIT соответственно. В условиях сравнения шепота против нормальной речи значение метрики EER снижается с 17.36% до 11.24% и с 8.99% до 8.94% на данных MSP-AVW и wTIMIT соответственно.

В смешанных сравнениях важную роль играет доменная нормализация. Несмотря на улучшение качества работы с шепотом, адаптированная модель все еще подвержена доменному сдвигу. Итоговое решение с использованием адаптированной модели и применением доменной нормализации обеспечивает снижение значений по метрике EER: с 20,71% до 8,11% и с 18,82% до 6,53% на данных MSP-AVW и wTIMIT соответственно.

Помимо адаптации, в данной работе разработан алгоритм детектирования шепота на уровне дикторских эмбедингов. В качестве модели классификатора используется линейный дискриминантный анализ. Обученный алгоритм обеспечивает качество по метрике F1: 96.7% и 99.96% на данных MSP-AVW и wTIMIT соответственно.

Выводы. Проведенные исследования продемонстрировали эффективность применения биометрических моделей на основе wav2vec2.0 для обработки шепотной речи. Адаптация с использованием Praat улучшает качество работы систем автоматической верификации диктора, без значительного снижения качества работы с вокализованной речью. Классические линейные алгоритмы машинного обучения позволяют качественно детектировать шепот на уровне дикторских эмбедингов.

Список использованных источников:

1. Baevski A. et al. wav2vec 2.0: A framework for self-supervised learning of speech representations //Advances in neural information processing systems. – 2020. – Т. 33. – С. 12449-

12460.

2. Boersma P. Praat: doing phonetics by computer [Computer program] //http://www.praat.org/. – 2011.

3. Lim B. P. Computational differences between whispered and non-whispered speech. – University of Illinois at Urbana-Champaign, 2011.

4. Cummins F. et al. The CHAINS corpus (characterizing individual speakers)(2006) [Электронный ресурс]

5. Tran T., Mariooryad S., Busso C. Audiovisual corpus to analyze whisper speech //2013 IEEE International Conference on Acoustics, Speech and Signal Processing. – IEEE, 2013. – C. 8101-8105.