В. В. Василевский Р. Д. Рында

АЛГОРИТМ ИЗМЕНЕНИЯ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ АУДИОФАЙЛА

Реализован алгоритм изменения эмоциональной окраски речи. Алгоритм базируется на модели Wav2Vec2, дообученной с использованием датасета, в котором один человек озвучивает фразы с различными эмоциями. На основе полученной модели разработано графическое приложение, позволяющее преобразовывать эмоции аудиозаписей.

Введение

Работа посвящена алгоритму, разработанному для анализа и изменения эмоциональной окраски аудиофайлов, с применением предобученной модели Wav2Vec2.

Дообучение модели выполнялось на датасете, озвученном одним спикером с нейтральным, радостным, грустным, удивлённым и раздражённым эмоциональными окрасами речи.

I. Алгоритм работы

Входной аудиосигнал сначала разбивается на фиксированные аудиофреймы $x \in \mathbb{R}^M$, где M — размерность входного вектора-фрейма. Для извлечения признаков используется дообученная модель Wav2Vec2, которая преобразует каждый фрейм в скрытый вектор $h \in \mathbb{R}^N$ по формуле

$$h = \tanh(Wx + b), \tag{1}$$

где $N \in [768, 1024]$ выбирается исходя из требований к качеству и скорости. После обработки всех фреймов получается последовательность $\{h_i\}_{i=1}^{T}$, (T — количество фреймов в сигнале), передаваемая в блок коррекции эмоций.

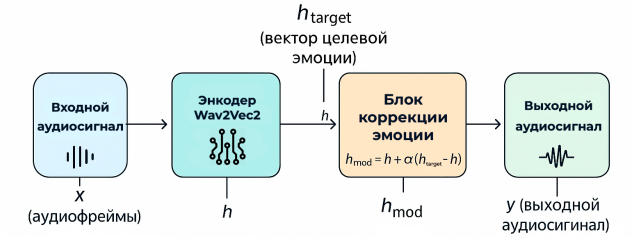

Рис. 1 – Схема работы системы

Исходное представление h смещается в направлении вектора целевой эмоции h_{target}, вычисляемого как

$$h_{\text{target}} = \frac{1}{K} \sum_{j=1}^{K} h_j^{(\text{target})}. \tag{2}$$

Далее формируется дельта-вектор $d = h_{\text{target}} - h$, который проходит через нейронную сеть построенную на архитектуре MLP(multilayer perceptron):

$$d' = W_2 \,\text{ReLU}(W_1 d + b_1) + b_2, \tag{3}$$

где $W_1 \in \mathbb{R}^{K \times N}$ и $W_2 \in \mathbb{R}^{N \times K}$. Полученный вектор масштабируется пользовательским коэффициентом $\alpha \in [0, 1]$ и складывается с исходным представлением: $h_{\text{mod}} = h + \Delta, \quad \Delta = \alpha\, d'$.

Для обеспечения плавности во времени к $\{h_{\text{mod},i}\}$ дополнительно применяется нормализация и, при необходимости, одномерная свёртка.

Рис. 2 – Внутренняя логика блока коррекции эмоций

Последовательность модифицированных векторов $\{h_{\text{mod},i}\}$ подаётся на вход вокодеру, который синтезирует выходной аудиосигнал y. На финальном этапе выполняется нормализация амплитуды и сглаживание возможных артефактов, полученных в процессе преобразования.

II. Выводы

Предложен алгоритм, построенный на основе дообученной модели Wav2Vec2 и модуле коррекции эмоций, позволяющий изменять эмоциональную окраску аудиофайлов. Разработанная модель интегрирована в приложение, построенное на базе фреймворка qt.

1. Baevski, A., Zhou, H., Mohamed, A., Auli, M. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations // Advances in Neural Information Processing Systems. – 2020.
2. Eyben, F., Wöllmer, M., Schuller, B. Introduction to the special issue on affective computing // IEEE Transactions on Affective Computing. – 2016.

Василевский Владислав Валерьевич, студент кафедры информационных технологий автоматизированных систем, БГУИР, vlad.vasilevskiy.07@gmail.com.

Рында Роман Дмириевич, студент кафедры информационных технологий автоматизированных систем, БГУИР, romarynda1@gmail.com.

Научный руководитель: Езовит Алексей Владимирович, ассистент кафедры информационных технологий автоматизированных систем, БГУИР, a.ezovit@bsuir.by.