РАЗРАБОТКА АЛГОРИТМА КОНВЕЙЕРНОЙ ОБРАБОТКИ «РЕЧЬ → ТЕКСТ → ЭМОЦИЯ»

Нугманова М.А.
Digital technologies and artificial intelligence research institute, Uzbekistan
E-mail: Miss.abduazimova@gmail.com

Аннотация. В данной статье предлагается алгоритм конвейерной обработки «Речь → Текст → Эмоция» для полноценного понимания эмоционального состояния с помощью смысловой составляющей текста. Данный алгоритм конвеерной обработки состоит из двух основных модулей: первый модуль — Автоматическое распознавание речи (ASR), преобразующий речевой сигнал в текст; второй модуль — Обработка естественного языка (NLP), обеспечивающий определение эмоции на основе полученного текстового представления.

Ключевые слова: конвейер, сигнал, признак, ASR, NLP.

Annotatsiya. Ushbu maqolada “Nutq → Matn → Emotsiya” ketma-ket ishlov berish algoritmi taklif etiladi, u matnning ma’no jihatidan tahlili orqali insonning emotsional holatini to‘liq tushunishni ta’minlaydi. Taklif etilgan konveyer algoritmi ikki asosiy moduldan iborat: birinchi modul — Nutqni avtomatik aniqlash tizimi (ASR), u nutq signalini matnga aylantiradi; ikkinchi modul esa — Tabiiy tilni qayta ishlash tizimi (NLP), u olingan matnli tasvir asosida emotsiyani aniqlashni amalga oshiradi.

Kalit so‘zlar: konveyer, signal, belgi, ASR, NLP.

Abstract. This paper proposes a pipeline algorithm “Speech → Text → Emotion” designed for comprehensive understanding of a speaker’s emotional state through the semantic interpretation of text. The proposed pipeline consists of two main modules: the first module — Automatic Speech Recognition (ASR), which converts the speech signal into text; and the second module — Natural Language Processing (NLP), which determines the emotion based on the obtained textual representation.

Keywords: pipeline, signal, feature, ASR, NLP.

Введение.

Задача распознавания эмоций по речи на протяжении многих лет привлекает внимание исследователей. В классических подходах данный процесс реализовывался исключительно на основе спектральных и просодических признаков, извлекаемых из акустического сигнала. Однако человеческие эмоции выражаются не только через тембр и интонацию голоса, но и посредством содержания высказывания. Например, даже если фраза «я очень счастлив» произнесена в абсолютно нейтральной интонации, её семантическое содержание однозначно указывает на эмоцию радости.

ISSN 3030-3907

Следовательно, для полноценного понимания эмоционального состояния необходимо учитывать не только акустический сигнал, но и смысловую составляющую текста.

Прежде всего, непрерывный речевой сигнал $s(t)$ преобразуется в цифровую форму. Сигнал представляет собой непрерывную функцию времени, однако компьютер способен обрабатывать его только в дискретном виде. В связи с этим осуществляется дискретизация сигнала:

$$s[n],\ n = 0,1,\ldots,N-1$$

Здесь $s[n]$обозначает n-ю выборку речевого сигнала, N — общее количество выборок, а F_s — частота дискретизации. Обычно основная информация речевого сигнала человека содержится в диапазоне частот от 0 до 8 кГц. Согласно теореме Найквиста–Шеннона, для полного восстановления информации, содержащейся в сигнале, частота дискретизации должна быть как минимум в два раза выше максимальной частоты спектра. Поэтому в системах автоматического распознавания речи (ASR) обычно применяются частоты 16 кГц или 22,05 кГц. На практике также могут использоваться более высокие частоты, такие как 44,1 кГц или 48 кГц, однако это значительно увеличивает вычислительные затраты и, как правило, является избыточным для задачи определения эмоций. В результате процесса дискретизации непрерывный акустический сигнал преобразуется в цифровую форму и становится готовым для обработки компьютерными алгоритмами.[1]

На основе дискретного сигнала формируются акустические признаки для ASR. Эти признаки сегментируются по временным кадрам, каждый из которых длится, как правило, от 20 до 40 миллисекунд. При этом кадры частично перекрываются (например, окно анализа длительностью 25 мс и шаг 10 мс), что обеспечивает плавный переход во временной области. Для каждого кадра вычисляются такие признаки, как лог-Мел спектр или коэффициенты мел-частотной кепстральной трансформации (MFCC).[2] В результате формируется следующая матрица:

$$X \in \mathrm{R}^{T \times D}$$

Где T — количество временных кадров, D — размерность признакового пространства. Например, если длительность речи составляет 1 секунду, а шаг анализа равен 10 мс, то $T \approx 100$. Размерность признаков D обычно принимается равной 13 (MFCC), 40 или 80 (лог-Мел спектр). Таким образом, матрица X отражает временное акустическое представление речи и служит входными данными для модели ASR.

ISSN 3030-3907

При преобразовании речи в текст с помощью модуля ASR широко применяются два основных подхода: CTC временная классификация на основе нейронных сетей и Seq2Seq модель преобразования одной последовательности данных в другую. Модель CTC вычисляет вероятность транскрипции следующим образом:

$$P_{\text{CTC}}\left(w_{1:M} \mid X\right) = \sum_{\pi \in \text{B}^{-1}\left(w_{1:M}\right)} \prod_{t=1}^{T} P\left(\pi_t \mid X, t\right)$$

Здесь $w_{1:M} = \left(w_1, w_2, \ldots, w_M\right)$ — последовательность текстовых токенов, $\pi = \left(\pi_1, \pi_2, \ldots, \pi_T\right)$ — скрытая последовательность, содержащая также специальные символы *blank*. Оператор $\text{B}(\cdot)$ сокращает *blank*-символы и повторяющиеся элементы, формируя итоговый текст. Каждый элемент $P\left(\pi_t \mid X, t\right)$ представляет собой вероятность появления символа на временном кадре. Таким образом, модель CTC суммирует вероятности всех возможных вариантов выравнивания. Преимущество данного подхода заключается в том, что даже при различии длин речевого сигнала и текста модель способна корректно их сопоставлять.[3] Например, сигнал из 100 кадров может соответствовать последовательности из 10 токенов. CTC-модель учитывает все возможные соответствия и вычисляет итоговую вероятность.

В отличие от этого, подход Seq2Seq основан на иной концепции. В нём декодер предсказывает каждый новый токен, опираясь на ранее сгенерированные токены и акустический контекст:

$$P_{\text{S2S}}\left(w_{1:M} \mid X\right) = \prod_{m=1}^{M} P\left(w_m \mid w_{1:m-1}, X\right)$$

Где w_m — токен в позиции m, $w_{1:m-1}$ — последовательность предыдущих токенов, а $P\left(w_m \mid w_{1:m-1}, X\right)$ — акустический контекст. Такой подход напрямую сопоставляет речь и текст, обеспечивая более высокую степень семантической целостности. Например, в зависимости от контекста модель может выбрать правильный вариант между словами «olma» и «alma». В моделях Seq2Seq вероятностная последовательность выравнивается по длине, что позволяет формировать семантически цельные предложения.

Для дальнейшего повышения качества результатов ASR используется интеграция с языковой моделью.[4] Метод объединения языковой модели с моделью распознавания речи на этапе декодирования вычисляет итоговую вероятность следующим образом:

$$\log P\left(w_{1:M} \mid X\right) \approx \log P_{\text{ASR}}\left(w_{1:M} \mid X\right) + \alpha \log P_{\text{LM}}\left(w_{1:M}\right) + \beta M$$

где P_{ASR} — вероятность, рассчитанная акустической моделью (CTC или Seq2Seq), P_{LM} — вероятность языковой модели(CTC или Seq2Seq), α — вес

ISSN 3030-3907

языковой модели, β — штраф за длину, M — количество токенов. Вероятность языковой модели P_{LM} формируется на основе статистических закономерностей языка. Например, последовательность «men kitob o‘qiyapman» имеет более высокую вероятность, чем «men kitob yeyapman», поскольку языковая модель учитывает естественные сочетаемости слов. Таким образом, применение метода объединения языковой модели с моделью распознавания речи делает результаты ASR более естественными и семантически корректными.[5]

После получения текста из речи активируется модуль NLP. Текст подаётся на вход трансформерной модели, и извлекается эмбеддинг специального символа CLS:

$$h = f_{\Theta}(T),\ h \in \mathrm{R}^{H}$$

Где h — общий эмбеддинг текста, f_{Θ} — отображающая функция модели BERT или XLM-R, H — размерность скрытого вектора. Данный эмбеддинг отражает семантику всего предложения и далее используется для классификации эмоций.

Дискретные эмоции определяются с использованием softmax-классификатора:

$$p_k = \frac{\exp((W_c h + b_c)_k)}{\sum_{j=1}^{K} \exp((W_c h + b_c)_j)},\ k = 1, \ldots, K$$

Где $K = 5$ — количество классов эмоций: радость, грусть, гнев, страх и нейтральное состояние. Функция softmax нормализует вероятности для всех классов в диапазоне от 0 до 1, при этом их сумма равна 1. Например, если на вход подаётся предложение «men juda xursandman» («я очень счастлив»), то на выходе softmax вероятность для класса «радость» может составить 0.85, а для остальных классов — значительно меньшие значения.

Дименсиональные эмоции оцениваются посредством регрессии:

$$\hat{r} = \tanh(W_r h + b_r),\ \hat{r} = [\hat{v}, \hat{a}]^T$$

Где $\hat{v}$ — валентность (степень позитивности или негативности), $\hat{a}$ — уровень возбуждения (интенсивность эмоционального состояния). Функция tanh ограничивает значения в интервале $[-1,1]$. Например, состояние радости обычно характеризуется значениями $\hat{v} \approx +0.8$, $\hat{a} \approx +0.7$; состояние грусти — $\hat{v} \approx -0.7$, $\hat{a} \approx -0.5$. Таким образом, дименсиональная модель отображает эмоции в координатном пространстве.

При обучении модели ключевую роль играют функции потерь. Для дискретных эмоций используется кросс-энтропия:

ISSN 3030-3907

$$L_{\text{disc}} = -\sum_{k=1}^{K} y_k \log p_k$$

Где y_k — истинный класс, представленный в виде one-hot-вектора. Для дименсиональных эмоций применяется коэффициент согласованной корреляции (Concordance Correlation Coefficient, CCC):

$$\text{CCC}\left(u, \hat{u}\right) = \frac{2\rho_{u,\hat{u}}\sigma_u\sigma_{\hat{u}}}{\sigma_u^2 + \sigma_{\hat{u}}^2 + (\mu_u - \mu_{\hat{u}})^2}$$

Значение CCC, близкое к 1, свидетельствует о высоком качестве аппроксимации; значение около 0 указывает на отсутствие согласованности. Для валентности и возбуждения коэффициенты рассчитываются отдельно, а итоговая дименсиональная функция потерь записывается в виде:

$$L_{\text{dim}} = \left(1 - \text{CCC}\left(v, \hat{v}\right)\right) + \left(1 - \text{CCC}\left(a, \hat{a}\right)\right)$$

Общая функция потерь объединяет обе задачи:

$$L_{\text{text}} = \lambda_1 L_{\text{disc}} + \lambda_2 L_{\text{dim}},$$

Где λ_1 и λ_2 — весовые коэффициенты, обычно выбираемые в диапазоне 0.5–1.0 для обеспечения баланса между задачами.

Так как модуль ASR может допускать ошибки, для повышения устойчивости результата применяются *N-best* гипотезы. Например, система ASR может выдать первую гипотезу «men kitob o‘qiyapman» («я читаю книгу») и вторую — «men kitob yeyapman» («я ем книгу»). Каждой гипотезе присваивается коэффициент достоверности q_i, сумма которых равна 1. Тогда ожидаемое значение дискретных вероятностей вычисляется как:

$$\bar{p} = \sum_{i=1}^{N} q_i \ \text{softmax}\left(W_c f_\Theta\left(T^{(i)}\right) + b_c\right)$$

а для дименсиональных выходов:

$$\bar{r} = \sum_{i=1}^{N} q_i \ \tanh\left(W_r f_\Theta\left(T^{(i)}\right) + b_r\right)$$

Таким образом, учитываются различные варианты транскрипции, что делает итоговую оценку эмоции более стабильной. Этот метод повышает устойчивость к ошибкам и обеспечивает надёжность системы в практических условиях.

В заключение можно отметить, что разработанный алгоритм конвейерной обработки «Речь → Текст → Эмоция» позволяет определять эмоциональные состояния в речи с учётом семантической информации. Такой подход обеспечивает более высокую точность по сравнению с методами, основанными исключительно на акустических признаках, поскольку содержательная часть текста зачастую точнее отражает эмоцию.

ISSN 3030-3907

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Rabiner, L. R., & Schafer, R. W. *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.
2. Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). Stanford University.
3. Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. (2016). *Listen, Attend and Spell.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4960–4964.
4. Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). *Attention-Based Models for Speech Recognition.* Advances in Neural Information Processing Systems (NeurIPS), 28, 577–585.
5. Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). *Recurrent Neural Network Based Language Model.* In *INTERSPEECH 2010*, 1045–1048.

ISSN 3030-3907