Нейромережні технології при аналізі голосових повідомлень користувачів

Шевченко Н.Ю.
Донбаська державна машинобудівна академія

Одним з ключових біометричних параметрів людини є її голос, що має набір індивідуальних особливостей, які відносно легко піддаються вимірюванню, наприклад, частотні або амплітудні характеристики голосового сигналу [1].

Серед базових характеристик голосу можна виділити силу голосу або гучність, висоту голосу (у середньостатистичної людини діапазон до 1,5 октави, в повсякденному спілкуванні більшість використовує 3-4 ноти), тембр голосу – сукупність додаткових коливань або обертонів, які виникають поряд з основною частотою, тон голосу (емоційне забарвлення голосу), яке задається різними методами: зміною гучності, темпу, висоти. За переліченими характеристиками можна не тільки ідентифікувати користувача мобільного пристрою, але і «навчитися» визначати його настрій.

Отже, одним з перспективних напрямів прикладного використання методів аналізу голосу є визначення настрою користувача (власника мобільного пристрою) через аналіз його голосових повідомлень та, виходячи з цього, підбирати контент для контекстних повідомлень протягом доби.

Для визначення настрою користувача (власника мобільного пристрою) через аналіз його голосових повідомлень необхідно реалізувати наступний алгоритм (у концептуальному сенсі):

1) визначення характеристик голосу;

2) побудова нейромережевої моделі визначення настрою користувача.

Для визначення характеристик голосу будується спектрограма на множині частот і амплітуд за допомогою перетворення Фур’є. При цьому необхідно враховувати, що перетворення Фур’є – математична функція, яка націлена на ідеальний, незмінний звуковий сигнал, тому вимагає практичної адаптації.

Наприклад, можна аудіозапис поділити на невеликі відрізки, протягом яких звук не буде змінюватися, а далі застосувати перетворення до кожного з відрізків [2].

Далі будується спектрограма другого порядку для усунення зайвої інформації у вигляді гармонік, які не зручні для аналізу, бо дублюють інформацію. В результаті отримується єдиний пік (кепстр), що характеризує монотонну хвилю.

Для визначення саме настрою користувача використовується нейронна мережа – багатошаровий персептрон з алгоритмом зворотного поширення помилки. В якості активаційної функції в багатошаровому персептроні використовується сигмоїдальна активаційна функція, зокрема логістична.

Вхідними параметрами мережі виступають характеристики голосу, а в якості виходів – вектор можливих варіантів настрою користувача.

Для аналізу голосу та навчання нейромережі користувачеві задаються декілька ключових запитань (в текстовому або звуковому форматі), які дозволять отримати еталоні властивості голосу власника мобільного пристрою.

Далі формується база еталонних зразків для порівняння (протягом декількох тижнів, наприклад), користувачеві задаються питання протягом дня. Відбувається спектральний аналіз отриманих повідомлень. Користувач самостійно визначає власний настрій.

Процес порівнювання поточних зразків з еталонними складається з наступних стадій [2]: фільтрація шумів, спектральне перетворення сигналу, постфільтрація спектра, ліфтерінг, накладення вікна Кайзера, фільтрація.

Основним порівняльним параметром є міра подібності звукових фрагментів. Для її обчислення необхідно порівняти спектрограми цих фрагментів. При цьому спочатку порівнюються спектри, отримані в окремому вікні, а потім обчислені значення усереднюються. X [1..N] і Y [1..N] – масиви чисел, однакового розміру N, що містять значення спектральної потужності першого і другого фрагментів відповідно, міра подібності між якими обчислюється за формулою:

$$f_{xy} = \left| \frac{\sum_i (x_i - M_x)(y_i - M_y)}{\sqrt{\sum_i (x_i - M_x)^2} * \sqrt{\sum_i (y_i - M_y)^2}} \right|, \quad (1)$$

де Мх і Му – математичні очікування для масивів X [] і Y [].

Навчання нейронної мережі відбувається за алгоритмом зворотного розповсюдження помилки з наступними вхідними параметрами: значення f_{xy} за першою фразою, значення f_{xy} за другою фразою, значення f_{xy} за третьою фразою.

На виході нейронної мережі – вектор можливих варіантів настрою користувача: {гарний, середній, поганий}.

Найбільш оптимальною архітектурою є MLP 3x2x1 (помилка класифікації 0,05). Параметри вагових коефіцієнтів визначеної структури нейронної мережі використовуються для прогнозування настрою користувача.

Так, наприклад, при f_{xy}(phrase1)=87 та f_{xy}(phrase2)=90 та f_{xy}(phrase3)=95 настрій користувача характеризується як «середній».

Далі в залежності від варіанта настрою формується користувацький контент для підтримки гарного настрою, або підвищення настрою протягом доби, якщо він нижче «гарного».

Варіант створення та використання контенту: формування відокремленої бази даних повідомлень різноманітного семантичного наповнення, вивід повідомлення на екран мобільного пристрою. Повідомлення може супроводжуватися звуковим сигналом відповідного наповнення. База даних повинна поновлюватися щотижнево. При цьому повідомлення не повинні повторюватися.

Література

Спектральный компьютерный анализ голоса – метод ранней и дифференциальной диагностики нарушений голосовой функции [Електронний ресурс]. – URL: https://nikio.ru/спектральній-анализ-голоса (дата: 27.03.2021).

Идентификация пользователя по голосу / А. Желтов [Електронний ресурс]. – URL: https://habr.com/ru/post/144580/ (дата: 27.03.2021).