Секція: Технічні науки

Семенюк Вікторія Валеріївна

магістр

Донецького національного університету імені Василя Стуса

м. Вінниця, Україна

ПІДВИЩЕННЯ ЯКОСТІ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ ЕМОЦІЙНОГО СТАНУ ЛЮДИНИ ПО ЕМОЦІЙНО ЗАБАРВЛЕНИМ ФРАЗАМ

Емоційний стан людини – це важливий аспект міжлюдського розуміння. Завдяки емоційним ознакам ми можемо зрозуміти у якому стані знаходиться співрозмовник: чи має він гарний настрій, чи, навпаки, засмучений, нейтральний, чимось незадоволений. Людина може визначити емоційний стан співрозмовника за певними ознаками. Він сприймається як в цілому (візуально – вираз обличчя, поведінка і таке інше, на слух – інтонація вимови, мова), так і окремо – по одному з факторів.

Мета дослідження полягає у підвищенні якості автоматичного розпізнавання емоційного стану людини маючи тільки аудіо уривок (зразок – емоційно забарвлена фраза) речення цієї особи, тобто тільки по голосу.

Так як психічний стан (емоція) впливає на емоційну насиченість звучання голосу, то емоційну насиченість продукування мовного висловлювання можна виявити по зміні інтонаційних характеристик мови. Якщо емоційне напруження в голосі зростає, то зміщення характеристик мовного сигналу буде направлено в область високих частот. Там розташовуються позитивні емоції, а в нижній частині – негативні. Сила і динаміка зміни сили голосу у часі, також важливий акустичний засіб кодування невербальної інформації. Наприклад, смуток виражається

слабкою силою голосу, а в динаміці – це повільні наростання і спади; гнів – виражається збільшеною силою голосу, різкими злетами і обривами. Також, зміни акустичних властивостей голосу та мови можуть бути обумовлені біофізичними характеристиками людини. У розпізнаванні емоційних ознак також допомагають паралінгвістичні та екстралінгвістичні засоби спілкування. Отже, за сукупністю цих ознак можна виокремити певні особливості, які характерні тільки для виду певної емоції та автоматизувати процес розпізнавання.

Існує багато способів для ідентифікації емоцій за голосом: використання різних алгоритмів, математичних моделей, систем ознак з використанням класифікаторів, методів розпізнавання мови з попередньою обробкою сигналу, акустичне моделювання, мовне моделювання та комбінування. Але переважно розпізнавання, в незалежності від вибору способу, відбувається за принципом роботи з аудіофайлом. Це забезпечує роботу з записаною мовою безпосередньо, але швидкість розпізнавання все одно залишає бажати кращого. На вирішення цієї проблеми можна подивитися і з іншого боку.

На даний час в завданнях ідентифікації емоцій в мові найчастіше використовують нейронні мережі, тому що такий підхід до ідентифікації має гарний результат. У роботі з аудіофайлами переважно використовують рекурентну нейромережу. Завдяки тому, що цей тип нейронної мережі має внутрішню пам'ять та може її використовувати для того, щоб обробити довільні послідовності входів, вона найбільш задовольняє потребам розпізнавання аудіозаписів. Зараз це стандартний підхід і він дає великий відсоток позитивних результатів, але, якщо є потреба це пришвидшити – виникають проблеми. Тому ідея дослідження складається в експериментальному припущенні, що для такого роду розпізнання можна задіяти згорткові нейромережі. Такий тип мереж швидко та з великим відсотком розпізнає зображення, тому проблема полягає ще й в тому, щоб

інтерпретувати аудіо в зображення. Для швидкого перетворення доцільніше використовувати аудіо-fingerprint. Ця технологія має схожість з дактилоскопією людини і полягає у виділенні найбільш важливих рис, які характерні для аудіосупроводу, що стійкі до шумів і спотворень, і порівнянні цих характерних ознак із зразками вибірки [1]. Реалізується на основі спектрального аналізу.

Для навчання нейронної мережі використовувалися аудіозаписи, які містили характерні ознаки емоцій – еталонна база даних. База даних мала у собі аудіозаписи професійних акторів, які зображували певний вид емоційного стану. Використовувалася вибірка, що складалась з 4500 аудіофайлів з емоційними фразами. Використовувалися записи двадцяти чотирьох професійних акторів (12 жінок і 12 чоловіків).

За допомогою методу звукового відбитку аудіофайли еталонної бази даних були конвертовані у зображення. На вхід надходила частотна синусоїда сигналу, потім за допомогою скрипта з синусоїди було згенеровано спектрограму. Для роботи використовувалися аудіофайли з виділеною емоцією і для аналізу не знадобилось розбиття аудіофайлу на фрагменти, фрейми або вікна.

В експерименті проводилось два типи аналізу: емоції згруповано на 3 класи (позитивні, негативні, нейтральні) та емоції згруповано на 8 класів для спроби більш точної класифікації (агресія, спокій, відраза, страх, щастя, нейтральний стан, печаль, здивування). Були створені дві концептуальні моделі для кожного з підходів з різною кількістю вихідних нейронів. Але для чистоти експерименту моделі для обох підходів мали однакову структуру. Обсяг тестової вибірки склав 956 записів. На малюнку 1 показані результати правильного розпізнавання (у відсотках) за узагальненим тестуванням для групи з трьох класів (ліворуч) і для групи з восьми класів емоцій (праворуч).

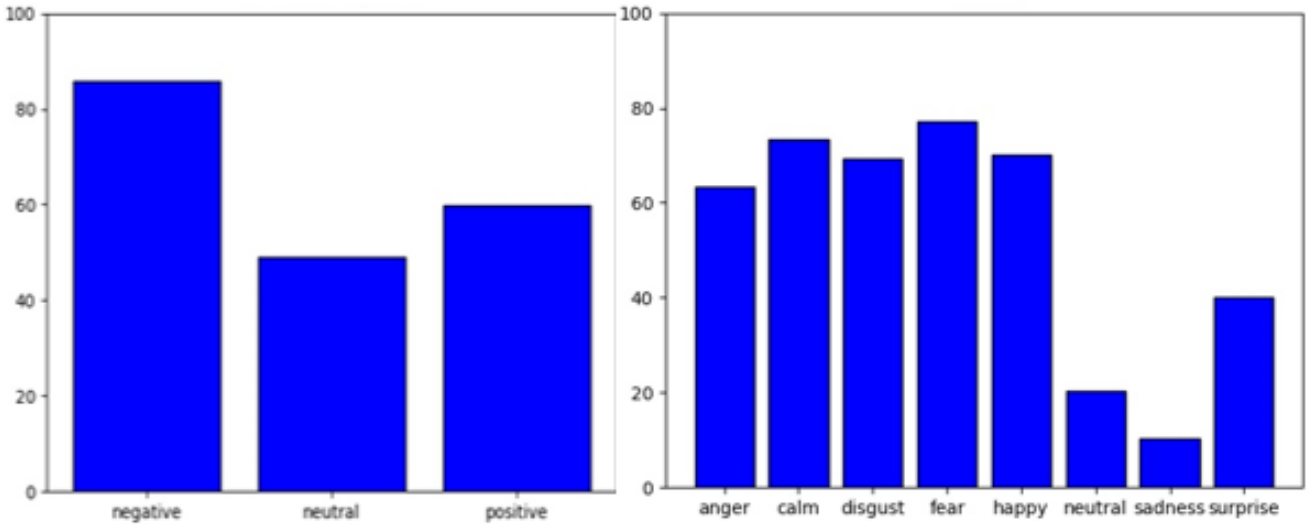

Рис 1. Результати правильного розпізнавання для 3 і 8 класів емоцій

Експериментальні дослідження показали, що для розпізнавання узагальнених класів емоцій найкраще визначаються негативні емоції, найгірші показники – у нейтральних емоцій. Для поліпшення точності розпізнавання можна застосувати додаткові характеристики фреймів сигналу, отримані алгоритмом MFCC [2].

Література

1. Grosche Peter, Müller Meinard and Serrà Joan Audio Content-based Music Retrieval. Multimodal Music Processing, Meinard Müller, Masataka Goto, and Markus Schedl, Eds. Dagstuhl Publishing, Wadern, Germany. April 2012. Vol. 3 of Dagstuhl Follow-Ups, chapter 9. P. 157-174.
2. Chakraborty Koustav, Talele Asmita, Upadhya Savitha Voice Recognition Using MFCC Algorithm. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, ISSN: 2349-2163. November 2014. P. 158-161.