UNIVERSITATEA TEHNICĂ
DIN CLUJ-NAPOCA

FACULTATEA DE ELECTRONICĂ, TELECOMUNICAȚII ȘI TEHNOLOGIA INFORMAȚIEI

MASTER TEHNOLOGII MULTIMEDIA

ACP 4

Speech Processing Applications

Oscar GAL II TM

Coordonator: Prof. dr. ing. Mircea GIURGIU

2022

str. Memorandumului nr. 28, 400114 Cluj-Napoca, România
tel. +40-264-401200, fax +40-264-592055, secretariat tel. +40-264-202209, fax +40-264-202280
www.utcluj.ro

Abstract— **This paper will present an implementation of a Tensorflow pre-trained deep convolutional neural network model to classify voices and figure out if a person has or hasn't Parkinson disease.**

I. INTRODUCTION

In modern days, speech recognition technologies are being used on a daily basis by a large percent of the population. Alexa, Siri, Cortana, and many other more are changing the way people interact with their devices, homes, cars and jobs. These kinds of vocal assistants allow people to talk to a computer or a device that interprets what that person is saying in order to respond to a question or a command.

With a long history of development and innovation, it was the introduction of these artificial intelligence voice-controlled assistants, or digital assistants, into the voice recognition market that changed the landscape of this technology in the 21st century.

It is estimated that by the end of 2022, 55% of the american households will be equipped with smart speakers and 50% of the searches will be voice activated, due to the fact that these types of searches are said to be the fastest growing mobile type search.

Voice Recognition Market size surpassed USD 2 billion in 2019 and is anticipated to grow at over 18% CAGR between 2020 and 2026. The demand for easy, faster, and convenient user authentication among various sectors will play a crucial role in driving the industry growth.Voice Recognition Market size surpassed USD 2 billion in 2019 and is anticipated to grow at over 18% CAGR between 2020 and 2026. The demand for easy, faster, and convenient user authentication among various sectors will play a crucial role in driving the industry growth.

Speech recognition applications have infiltrated into people's daily lives, not only in their homes, therefore we can list the following major domains:

- In the workplace
- Banking
- Marketing
- Healthcare
- IoT
- Language learning

The benefits of using speech recognition technology include incorporating simple tasks to increase efficiency, such as starting video conferences, schedule meetings, making payments, quickly gathering data, reminders, less paperwork, improved workflows, assisted guidance and navigation, and voice commands.

People with disabilities can benefit from speech recognition programs. For individuals that are Deaf or Hard of Hearing, speech recognition software is used to automatically generate a closed-captioning of conversations such as discussions in conference rooms, classroom lectures, and/or religious services.

Speech recognition is also very useful for people who have difficulty using their hands, ranging from mild repetitive stress injuries to involve disabilities that preclude using conventional computer input devices. In fact, people who used the keyboard a lot and developed RSI became an urgent early market for speech recognition. Speech recognition is used in deaf telephony, such as voicemail to text, relay services, and captioned telephone. Individuals with learning disabilities who have problems with thought-to-paper communication (essentially they think of an idea but it is processed incorrectly causing it to end up differently on paper) can possibly benefit from the software but the technology is not bug proof. Also the whole idea of speech to text can be hard for intellectually disabled person's due to the fact that it

is rare that anyone tries to learn the technology to teach the person with the disability.

This type of technology can help those with dyslexia but other disabilities are still in question. The effectiveness of the product is the problem that is hindering it from being effective. Although a kid may be able to say a word depending on how clear they say it, the technology may think they are saying another word and input the wrong one. Giving them more work to fix, causing them to have to take more time with fixing the wrong word.

Speech processing is the study of speech signals and processing methods of signals. These signals are usually processed in a digital representation. Aspects of speech processing include the acquisition, manipulation, storage, transfer and output of speech signals. Therefore, while the input is called speech recognition, the output is called speech synthesis.

Speech recognition uses a broad array of research in computer science, linguistics and computer engineering. Many modern devices and text-focused programs have speech recognition functions in them to allow for easier or hands-free use of a device. There are two types:

- Speech recognition is used to identify words in spoken language
- Voice recognition is a biometric technology for identifying an individual's voice

Speech recognition systems use computer algorithms to process and interpret spoken words and convert them into text. A software program turns the sound a microphone records into written language that computers and humans can understand, following these four steps:

1. analyze the audio;
2. break it into parts;
3. digitize it into a computer-readable format; and
4. use an algorithm to match it to the most suitable text representation

To meet these requirements, speech recognition systems use two types of models:

- Acoustic models. These represent the relationship between linguistic units of speech and audio signals.
- Language models. Here, sounds are matched with word sequences to distinguish between words that sound similar.

The accuracy of speech recognition may vary depending on the following factors:

- Error rates increase as the vocabulary size grows:

e.g. the 10 digits "zero" to "nine" can be recognized essentially perfectly, but vocabulary sizes of 200, 5000 or 100000 may have error rates of 3%, 7%, or 45% respectively.

- Vocabulary is hard to recognize if it contains confusing words:

e.g. the 26 letters of the English alphabet are difficult to discriminate because they are confusing words (most notoriously, the E-set: "B, C, D, E, G, P, T, V, Z — when "Z" is pronounced "zee" rather than "zed" depending on the English region); an 8% error rate is considered good for this vocabulary.

- Speaker dependence vs. independence:

  A speaker-dependent system is intended for use by a single speaker.

  A speaker-independent system is intended for use by any speaker (more difficult).

- Isolated, Discontinuous or continuous speech

With isolated speech, single words are used, therefore it becomes easier to recognize the speech.

Machine learning, a subset of artificial intelligence, refers to systems that can learn by themselves. It involves teaching a computer to recognize patterns, rather than programming it with specific rules. The training process involves feeding large amounts of data to the algorithm and allowing it to learn from that data and identify patterns. In the early days, programmers would have to write code for every object they wanted to recognize (e.g. human v. dog); now one system can recognize both by showing many examples of each. As a result, these systems continue to get smarter over time without human intervention.

Teaching a machine to learn to read a spoken language as humans do, is something that hasn't yet been perfected. Listening to and understanding what a person says is so much more than hearing the words the person speaks. As humans, we also read the person's eyes, their facial expressions, body language, and the tones and inflections in their voice. Another nuance of speech is the human tendency to shorten certain words (e.g. "I don't know" becomes "dunno"); we have said abbreviated words for so long that we do not pronounce them as precisely as when we learned them. This human disposition poses yet another challenge for machine learning in speech recognition.Natural language processing (NLP) is a division of artificial intelligence that involves analyzing natural language data and converting it into a machine-readable format. Speech recognition and AI play an integral role in NLP models in improving the accuracy and efficiency of human language recognition.

In speech recognition, the computer takes input in the form of sound vibrations. This is done by making use of an analog to digital converter that converts the sound waves into a digital format that the computer can understand. Advanced speech recognition in AI also comprises AI voice recognition where the computer can distinguish a particular speaker's voice.

II. MOTOR NEURONE DISEASE

Motor neurone disease (MND) is a neurodegenerative condition that affects the brain and spinal cord. MND is characterized by the degeneration of primarily motor neurons, leading to muscle weakness.

The presentation of the disease varies and can be muscle weakness, wasting, cramps and stiffness of arms and/or legs, problems with speech and/or swallowing or, more rarely, with breathing problems. Whichever area the disease starts, as the disease progresses the pattern of signs and symptoms becomes similar, with increasing muscle weakness in the person's arms and legs, problems swallowing and communicating and weakness of the muscles used for breathing, which ultimately leads to death. Most people die within 2-3 years of developing symptoms, but 25% are alive at 5 years and 5-10% at 10 years. The most common type of MND is amyotrophic lateral sclerosis (ALS). There are rarer forms of MND such as progressive muscular atrophy and primary lateral sclerosis, which may have a slower rate of progression.

MND is a disorder which can affect adults of any age. However, incidence is highest in people aged 55-79; onset below the age of 40 years is

uncommon. There are approximately 4,000 people living with MND in England and Wales at any one time. The cause of MND is unknown. About 5-10% of people with MND have a family history of the disease and several abnormal genes have been identified.

Parkinson's disease (PD) can affect speech in several ways. Many people with PD speak quietly and in one tone; they don't convey much emotion. Sometimes speech sounds breathy or hoarse. People with Parkinson's might **slur words, mumble or trail off** at the end of a sentence. Most people talk slowly, but some speak rapidly, even stuttering or stammering. Understanding written or spoken language presumably involves spreading neural activation in the brain. This process may be approximated by spreading activation in semantic networks, providing enhanced representations that involve concepts not found directly in the text. The approximation of this process is of great practical and theoretical interest. Although activations of neural circuits involved in representation of words rapidly change in time snapshots of these activations spreading through associative networks may be captured in a vector model. Concepts of similar type activate larger clusters of neurons, priming areas in the left and right hemisphere. Analysis of recent brain imaging experiments shows the importance of the right hemisphere non-verbal clusterization. Medical ontologies enable development of a large-scale practical algorithm to re-create pathways of spreading neural activations. First concepts of specific semantic type are identified in the text, and then all related concepts of the same type are added to the text, providing expanded representations. To avoid rapid growth of the extended feature space after each step only the most useful features that increase document clusterization are retained. Short hospital discharge summaries are used to illustrate how this process works on real, very noisy data. Expanded texts show significantly improved clustering and may be classified with much higher accuracy. Although better approximations to the spreading of neural activations may be devised a practical approach presented in this paper helps to discover pathways used by the brain to process specific concepts, and may be used in large-scale applications. “NLP has progressed to the point where speech recognition software can very accurately understand human voices, including different accents and pronunciations.” How are words and concepts represented in the brain? The neuroscience of language in general, and word representation in the brain in particular, is far from being understood, but the cell assembly model of language has already quite strong experimental support (Dehaene, Cohen, Sigman, & Vinckier, 2005; Pulvermuller, 2003), and agrees with broader mechanisms responsible for memory (Lin, Osan, & Tsien, 2006). In the cell assembly (or neural clique) model words (or general memory patterns) are represented by strongly linked subnetworks of microcircuits that bind articulatory and acoustic representations of spoken words. The meaning of the word comes from an extended network that binds related perceptions and actions, activating sensory, motor and premotor cortices (Pulvermuller, 2003). Various neuroimaging techniques confirm the existence of such semantically extended networks.

- Parkinson’s disease (PD) is a neurodegenerative disorder of growing prevalence and incidence related to the decrement of substantia nigra in the midbrain.
- The Unified PD Rating Scale (UPDRS) is also used in clinical PD evaluations. It assigns a standardized score in the interval 0–4 on 50 items, evaluating concepts such as cognition, speech, swallowing, handwriting, gait, posture, and numbness.
- Regarding voice features, one of the most common arti-facts to measure the

Parkinson's deterioration is based on the sustained phonation /ah/ for as long and as steadily as possible

- On one hand, an approach to motivate the patient to speak is the work of Ireland in which a chat-bot conversation enables speech monitoring. On the other hand, another widely-used method to collect speech recordings are the phone calls. Both methods collect a big amount of data in a non-supervised way.

Feature number	Description
1	Pitch
2	Jitter
3	Shimmer
4	Sharpness
5	Noise-Harmonics Ratio
6	Mucosal Wave Energy Average
7-20	Cepstral Description of the Glottal Source
21-34	Spectral Description of the Glottal Source Power Spectral Density
35-46	Biomechanics of the Vocal Folds (Body and Cover)
47-58	Temporal Description of the Glottal Source Contact and Open Phases
59-62	Glottal Gaps: Contact, Adduction and Permanent
63-65	Indicators of Neurological Alteration
66-72	Physical, Neurological and Flutter Tremor (Amplitude and Frequency)

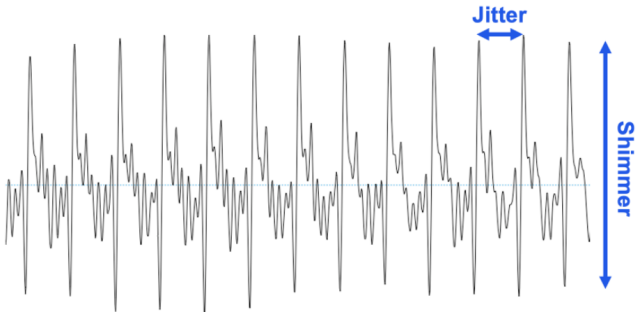
Classically the Vowel Space Area (VSA) and the Formant Centralization Ratio (FCR) have been proposed to characterize dysarthria in Parkinson's Disease (PD).

Nº (#)	Description
1	Diphthong /ah-ee/
2	Sequence /papapa/
3	Sequence /tatata/
4	Sequence /kakaka/
5	Sequence /pataka/
6	Sequence /pakata/
7	Triphthong /ah-ee-oo/
8	The five vowels /ah eh ee oh oo/

- Time domain estimates, as jitter, shimmer and open-closed phase quotients are also used in describing dysphonic voice
- The determination of the dysphonic grade is traditionally carried out by independent referees according to a subjective criterion on a given scale. One of the most popular is GRBAS (grade, roughness, breathiness, asthenia and strain)
- Under the functional point of view, the following main behaviors may be observed in the abnormal operation of the vocal folds: asymmetric vibration, contact defects and dystonia (hypo-, hyper-tension and tremor). Most organic larynx pathologies reproduce either one or another behavior, or all of them.
- As far as sustained vowels are concerned, this would not be a problem, but it becomes a major obstacle when segmental parameters are involved, as in the use of passages (either read or spontaneous).

The speech production system is not a rigid, mechanical machine, but composed of an assortment of soft-tissue components. Therefore, although parts of a speech signal might seem stationary, there are always small fluctuations in it, as vocal fold oscillation is not exactly periodic. Variations in signal frequency and amplitude are called jitter and shimmer, respectively. Jitter and shimmer are acoustic characteristics of voice signals, and they are caused by irregular vocal fold vibration. They are perceived as roughness, breathiness, or hoarseness in a speaker's voice. All natural speech contains some level of jitter and shimmer, but measuring them is a common way to detect voice pathologies. Personal habits such as smoking or alcohol consumption might increase the level of jitter and shimmer in voice. However, many other factors can have an effect as well, such as loudness of voice, language, or gender. As jitter and

shimmer represent individual voice characteristics that humans might use to recognize familiar voices, these measures could even be useful for speaker recognition systems.



III. RESULTS

The developed application serves the purpose of analyzing whether a patient does or does not have Parkinson, based on some parameters subtracted from their voice within an audio file;
The application consists of two parts:

- Backend, which is a Django Application
- Frontend, which is a React Application;

```
- algo
    |-- dataset
        |-- dataset_our_voices
            |...
        |-- dataset_voices
            |-- read_text
                |...
            |-- spontaneous_dialogue
                |...
        |-- processed_results.csv
        |-- trained_model.sav
        ("The 2 above files can be missing, but will be generated"
    |-- __init__.py
    |-- training_model.py
- media ("folder for django to store the saved media")
- parkinson ("django root app folder")
    |-- __init__.py
    |-- settings.py
    |-- urls.py
    |-- wsgi.py
    |-- asgi.py
- results ("django app")
    |-- migrations
        |...
    |-- __init__.py
    |-- admin.py
    |-- apps.py
    |-- models.py
    |-- urls.py
    |-- serializers.py
    |-- views.py
    |-- test.py
- API_Parkinson.postman_collection.json
- db.sqlite3
- manage.py
- requirements.txt
```

Backend

On the Backend part, the focus is on training an AI to predict whether the patient is healthy or suffers from Parkinson disease.

Through other attempts have been made at porting the functionality from Praat to Python, the library aims to provide a complete and Pythonic interface to the internal Praat code, accessing Praat's C/C++ code (which means that the algorithms and their output are identical to the ones in Praat) and providing efficient access to the program's data, while also providing an interface that looks like any other Python library.

In order to install Parselmouth, the following code must be ran into the terminal:

pip install praat-parselmouth

In order to do that, the work is based on an open source library, Parselmouth. A bank of audio files of both healthy and Parkinsons' sufferers have been used to train the AI, so for starters, the audio files will be processed in order to extract crucial data for the next steps. For each file, we take into consideration the following measurements:

- localJitter
- localAbsoluteJitter
- rapJitter
- ppq5Jitter
- localShimmer
- localdbShimmer
- apq3Shimmer
- aqpq5Shimmer
- apq11Shimmer
- hnr05
- hnr15
- hnr25
- hnr35

- hnr38

These measurements are taken with the help of the library previously mentioned, which allows the usage of Praat in Python code. Once the dataset is built, it will be exported into a CSV file for late usage. After creating the dataset, a Linear Regression object will be initialized, which will be given the previously processed data to learn.
Once the training is done, a dump of the Linear Regression shall be saved for future usage. In order to predict a result, the DatasetCreator class needs to be instantiated and the DatasetCreator.predict(wav_path="./path/to/file') method to be called'

All the files required for the algorithm to work can be found into the 'datasets' folder, where we can also find two more folders, one containing wav files of healthy and Parkinson patients' recordings, these being used to train the algorithm in the beginning - and the other one where other recordings for testing purposes can be found;

On the API side, we have the following:

- GET request on /results returns the results of all the audio files that have been previously analyzed
- POST on /results is the request used in order to process a new wav audio recording
- GET on /results/{{id}} is used to retrieve only one result, based on the id it has in the database
- PUT on /results/{{id}} has the usage of updating an already existent result based on the id it has been saved under in the database
- DELETE on /results/{{id}} has the purpose of deleting one of the results, in regards to the id it has;

Frontend

The frontend part of the project serves as an user interface. It is dedicated to medical researchers or doctors, who can add wav recording files of their patients and runs them through the algorithm that detects parkinson in one's voice. They are also provided with a list of all the audio recordings/ patients they have added in the app, each of them having their own 'mini medical file', once clicked on a patient's name. The recordings can be deleted or uploaded, while the individual card contains information about a certain person.

While this is just a prototype, the FE application can suffer serious improvements in the future, such as sending the results on the email, having multiple lists which could be filtered, opening graphics regarding the voice signal, authentification etc.

A React App serves as the front end, being the final user interface. "Create React App" is a popular way to create single page applications, providing a modern setup and no need for further configuration as certain tools, such as Babel or Webpack are preconfigured and hidden. To create the project, the only requirement is having Node.js installed and the NPX create-react app command run, together with the project name. The default script for running the project is accessed from the console by entering the NPM START command for development mode, which will automatically open the web browser on localhost, port 3000.

A single page app presents the user with a couple of inputs, for them to fill with name, age, sex and upload a .wav file to be processed by the backend. While the results are loading, a basic loader spinner will be displayed until the data are ready to be displayed in a modal.

The speech production system is not a rigid, mechanical machine, but composed of an assortment of soft-tissue components. Therefore, although parts of a speech signal might seem stationary, there are always small fluctuations in it, as vocal fold oscillation is not exactly periodic. Variations in signal frequency and amplitude are called jitter and shimmer, respectively. Jitter and shimmer are acoustic characteristics of voice signals, and they are caused by irregular vocal fold vibration. They are perceived as roughness, breathiness, or hoarseness in a speaker's voice. All natural speech contains some level of jitter and shimmer, but measuring them is a common way to detect voice pathologies. Personal habits such as smoking or alcohol consumption might increase the level of jitter and shimmer in voice. However, many other factors can have an effect as well, such as loudness of voice, language, or gender. As jitter and shimmer represent individual voice characteristics that humans might use to recognize familiar voices, these measures could even be useful for speaker recognition systems.

The code specified above can be also seen in the following repositories:

Backend:
https://github.com/galoscar07/api_parkinson
Frontend:
https://github.com/ioanac9/parkinson-interface

BIBLIOGRAPHY

[1] Lin TY. et al. (2014) Microsoft COCO: Common Objects in Context. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48
[2] "Speed/accuracy trade-offs for modern convolutional object detectors." Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z, Song Y, Guadarrama S, Murphy K, CVPR 2017
[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.- Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In the European Conference on Computer Vision, pages 21–37. Springer, 2016.
[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pages 91–99, 2015
[5] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. arXiv preprintarXiv:1605.06409, 2016.
[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640, 2015.