REPUBLIC OF TÜRKIYE
ALTINBAŞ UNIVERSITY
Institute of Graduate Studies
Electrical and Computer Engineering

CALL CENTER FOR DEEP LEARNING BASED ATTRACTIVE SPEECH RECOGNITION

Hasnaa Mohammed Hussein Jasim ALALAF

Master of Science

Supervisor

Prof. Dr. Osman Nuri UÇAN

Istanbul, 2023

CALL CENTER FOR DEEP LEARNING BASED ATTRACTIVE SPEECH RECOGNITION

Hasnaa Mohammed Hussein Jasim ALALAF

Electrical and Computer Engineering

Master's Thesis

ALTINBAŞ UNIVERSITY

2023

The thesis titled CALL CENTER FOR DEEP LEARNING BASED ATTRACTIVE SPEECH RECOGNITION, prepared by HASNAA MOHAMMED HUSSEIN JASIM ALALAF and submitted on 09/04/2023 has been **accepted unanimously** by the majority of votes for the degree of Master of Science in Electrical and Computer Engineering.

Prof. Dr. Osman Nuri UÇAN

the Supervisor

Thesis Defense Jury Members:

Prof. Dr. Osman Nuri UÇAN	Department Of Engineering and Architecture, Altinbas University	_____
Asst. Prof. Dr. Ayça KURNAZ TÜRKBEN	Department Of Engineering and Architecture, Altinbas University	_____
Asst. Prof. Dr. Serdar KARGIN	Department Of Biomedical Engineering, Arel University	_____

I hereby declare that this thesis/dissertation meets all format and submission requirements of a Master's thesis.

Submission date of the thesis to Institute of Graduate Studies: ____/____/____

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Hasnaa Mohammed Hussein Jasim ALALAF

Signature

DEDICATION

I devote and pledge this research work to my supervisor who is salient for guiding me through whole research work as well as my family for always assisting me in my hard time.

PREFACE

This thesis is made as a completion of the master education in Electrical and Computer Engineering. Several persons have contributed academically, practically and with support to this master thesis. I would therefore firstly like to thank my supervisor Prof. Dr. Osman Nuri UÇAN for him time, valuable input, and support throughout the entire master period.

Finally, I would like to thank my family and friends for being helpful and supportive during my time studying Electrical and Computer Engineering at Altinbas university.

ABSTRACT

CALL CENTER FOR DEEP LEARNING-BASED ATTRACTIVE SPEECH RECOGNITION

ALALAF, Hasnaa Mohammed Hussein Jasim

M.Sc., Electrical and Computer Engineering, Altınbaş University,

Supervisor. Prof. Dr. Osman Nuri UÇAN

Date: April / 2023

Pages: 48

Numerous sorts of businesses utilize call center technologies to manage client calls. The majority of this center's central staff handles incoming calls and routes them to the appropriate locations. Owing to a high call volume, clients are placed on hold for unusually extended times. Bringing additional personnel to the facility is a feasible solution to the problem. This may not be a thorough or cost-effective solution, and it certainly increases the cost. This research aims to use the Nurul network, K nearest neighbors, Support vector machine Classifier, Random Forest Classifier, Logistic regression Classifier, and decision tree-based machine learning algorithm to detect speech events in noisy urban environments using an open-source dataset (Gender Recognition by Voice) downloaded from the Kaggle website. Moreover, the Random Forest Classifier (99.5%) and the Decision Tree Classifier (98%) yield extremely accurate models. The interactive Turkish voice recognition system created for the thesis was examined across several variables, such as age, gender, department, duration, and ambient noise level.

Keywords: Call Center System, Deep Learning, Speech Recognition.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

ASR : Automatic Speech Recognition

HMM : Hidden Markov Model

NN : Nural Network

LSTM : Long Short-Term Memory

1. INTRODUCTION

Speech analytics permits the extraction of data from recorded talks. By studying client feedback, sentiment analysis is one use of speech analytics that tries to learn about a product or service. In addition, the findings of this study can assist contact center personnel in establishing rapport with callers and resolving any potential issues. Analysis of sound sensitivity can be conducted in two primary ways. The first modeling kind is acoustic, whereas the second is linguistic. In linguistic modeling, converting audio files to text files and then analyzing their content are required steps. As a result of this form of modeling, we can observe that specific phrases and expressions are utilized more frequently in particular settings. It is possible to predict certain outcomes based on the frequency with which certain words or phrases are used in spoken texts. Several research have demonstrated that linguistic modeling can be used to determine the emotional tone of a recorded speech. Throughout their experiments, the researchers modified their models while keeping dyslexia, semantic impairment, and word function in mind. Obtaining a human transcript of every audio recording, however, can be time-consuming and expensive. It is a daunting technical problem to develop a high-quality, automatic text-to-speech system that can convert audio recordings to text. Even if a single letter is swapped in the original text, a translation into a different language may vary substantially. Using the acoustic properties of audio data, acoustic modeling is performed. These are the measurable characteristics of human speech, such as rate of delivery and pitch, that a computer can determine. Together, these alterations can convey important information about emotions. We employ acoustic homes since they are also commonly used in other sorts of business and psychological evaluations, which informs our approach to sentiment analysis. The accuracy of the results, however, is highly dependent on the quality of the sound recordings used. Given that we use audio data as our acoustic homes, the sound quality can have a significant impact on our ability to determine the right values for those acoustic houses that would result in inferior models. The second issue is that random noise has always existed in actual history. Ideal sentiment analysis in real-time interactions may not be able to discern between seemingly unrelated noises based on training data. This strategy merits more examination and analysis due to its simplicity of implementation and effectiveness.

1.1 AIM OF STUDY

The primary goal of this study is aimed at employing the Nurul network, K- nearest neighbors, Support vector machine Classifier, Random Forest Classifier, Logistic regression Classifier, decision tree-based machine learning algorithm, for the detection of speech events in noisy urban environments.

i. Using a shared network in a NN acoustic model to develop Automatic speech recognition (ASR) system by using Different algorithm in deep learning.
ii. the first attempt to use deep learning techniques to develop a speech recognition system for East-to-West Uzbek and its numerous dialects.
iii. Researchers examined cutting-edge position-embedding strategies for Uzbek voice recognition using a convolution-augmented transformer.
iv. Several cutting-edge analytic techniques were applied to the suggested Automatic speech recognition (ASR) system for learning Uzbek, resulting in performance and overall system improvements.

.

1.2 THESIS STRUCTURE

The Thesis Remaining Chapters are Structured as Follows:
Chapter 2: Related Works with Literature Review
Chapter 3: The Proposed Protocol and Methodology
Chapter 4: Simulation In additional Results
Chapter 5: Conclusion In additional Future work

2. LITERATURE REVIEW

The earliest popularity surveys of voices date back to the turn of the 20th century. As the first mass-produced voice-recognition toy, Radio Rex was released to the market in the 1920s. In 1938, during the New York City World's Fair, scientists from bell labs unveiled their speech synthesis device to the public. Typically, system developers for automated voice recognition concentrate on phonetic-acoustic concepts.

If you need to convey your message quickly and accurately, there is no replacement for speech. A rising number of people are investing the time and energy necessary to master voice commands for usage with a variety of smart devices. At least 702 of the 7097 languages still in use today have been confirmed as being spoken by their speakers [1]. For a language to be deemed "alive," it must have at least one native speaker. While there are numerous languages spoken around the world, not everyone uses them in the same way. The majority of the world's population, however, speaks only one of the 23 official languages, which include Mandarin Chinese, English, Spanish, and Hindi. These languages are good for training artificial intelligence systems for Text-to-Speech (TTS), Automated Speech Recognition (ASR), natural language processing, and computational linguistics due to their extensive use and wealth of data. Yet, there is little funding to support research into obscure languages and the development of specialized technologies. Hence, finding equivalent solutions for languages with limited resources is a challenging and time-consuming endeavor.

ASR is an important and active research topic due to its wide range of theoretical and practical applications in fields such as security [2], education [3], smart healthcare [4,5], and smart cities [6], as well as the development of interfaces and technological tools that streamline voice processing. After the aural data has been converted to text, text matching can be used to further refine the transcription of the identified speech signal. A comprehensive framework for deep semantic learning is utilized by ASR to convert spoken words to text. Automatic speech recognition (ASR) research involves knowledge from numerous areas [7].

Recent years have witnessed considerable advancements in ASR due to the implementation of deep learning techniques. Google, Apple, Facebook, Microsoft, Amazon, and International Business Machines (IBM) have created state-of-the-art speech recognition systems for English, European, and Asian languages, but the development of automatic speech recognition (ASR) systems for most Central Asian languages, including Uzbek, is in its infancy. This is due to the absence of a central source of Uzbek speech samples and the existence of regional dialects [8]. Moreover, there are insufficient voice corpora for Central Asian languages to construct an ASR system. If a first strategy to categorizing sounds for the ASR system is not developed, a huge speech corpus may provide difficulties for the computer and dramatically reduce test results. Obtaining enough voice data for training and testing is a significant obstacle in the development of an automated speech recognition system. Only the 7,000 most widely spoken languages in the world have access to these training data at present.

ASR will become increasingly prevalent in human-computer interactions. Installations of ASR have had a substantial positive impact on numerous fields, including healthcare and government [10], because they increase productivity while reducing the demand for human capital. This is the case due to the rapid expansion of the sectors they serve. It is quite difficult to create an ASR system that adjusts for individual variances in accent and voice. People who speak multiple languages tend to have significantly different accents than those who speak only one. As aspects such as culture, gender, language, and speech rate are considered, it becomes increasingly challenging to acquire sufficient training data for the ASR model.

Both the Google Assistant and Siri on the iPhone can automatically detect conversations in over 40 and 35 languages, respectively [11]. Most current ASR systems [12,13,14,15] decode spoken words using a Gaussian Mixture Model, Hidden Markov Model, or Deep Neural Network (DNN) (GMMs). DNNs are essential for the development of ASR systems [16,17], which is largely attributable to the development of specific neural network models and training and classification techniques. These methods have been used to a wide range of problems, including feature extraction [20, 21], audio signal classification [22, 23], text recognition and TTS [24, 25], disordered speech processing [26], and vocabulary-based

short and long speech recognition [27]. Yet, the efficacy of ASR systems is highly dependent on the speech samples used to train DNN models.

ASR systems have recently switched from DNN-based hybrid modeling to End-To-End (E2E) modeling [28] [29,30,31,32]. For hybrid models, language, pronunciation, and acoustics must be optimized separately. End-to-end automatic speech recognition (ASR) systems educate the system's pronunciation, acoustic, and language modeling components to convert an input speech sequence into an output token sequence.

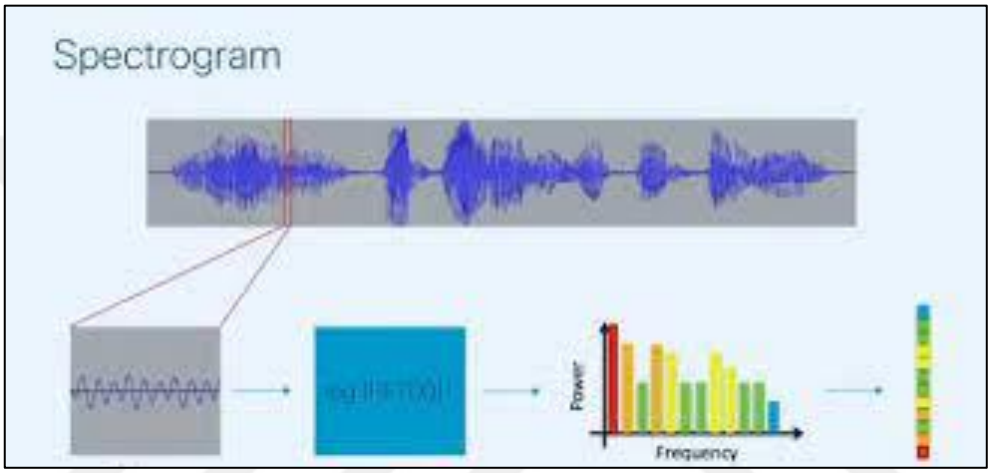

Figure 2.1: Speech Recognition By Algorithm [33].

An adaptive deep neural network Utilizing Hidden Markov Model (DNN-HMM) techniques, the most advanced automatic speech recognition system for Uzbek has been developed. For the test set [33], recently published ASR results on Uzbek speech corpus data have been made available, attaining a WER of 17.4%.

This research advances the creation of a vocal activity detection (VAD) pipeline related to the E2E transformer, as overly long speech segments are a prevalent problem in experimental ASR systems. The suggested pipeline combines the accuracy of In a Speech Segmented [34] with the maximal segmentation length of an energy-based VAD function. Using cutting-edge cooking techniques and meticulously constructed models, our major goal was to establish a new standard for the speech community.

HMM and GMM-based models are the industry standard in voice recognition. Given that the HMM is a stochastic model, it is calibrated with a small number of fixed states. Any phase of an incoming voice signal can conceal an infinite number of states. For the HMM

to function, the input voice signal must be capable of being represented by a well-defined parametric random process. However, advancements in speech technology have led to significant improvements in TTS and ASR [35,36], as well as a variety of problems with speech analysis and recognition. DNNs differ from conventional ANNs in that they conceal a substantial amount of data between the input and output layers. The neuron is the fundamental component of all neural networks, but especially deep neural networks (DNN). Between neurons in the same layer and those in the layer below, there exist complete connections. HMMs and GMMs have been mostly replaced by DNN-based acoustic models in ASR systems. Due of its completely linked feature, the DNN's inability to acquire structural localization data based on feature distance is a significant drawback. If you attempt to train a DNN using stochastic gradient descent (SGD), you will encounter the problem of vanishing gradients [37].

In the academic research field, raw speech has mostly replaced manually created characteristics. They employed DNN models to collect task-relevant data for these attributes [38]. These characteristics have been effective for a variety of speech-related difficulties, including emotion detection [39], automatic speaker recognition [40], and speaker identification [41]. Due to the success of end-to-end (E2E) models trained on raw speech samples, researchers in the field of speech technology have recently concentrated on convolutional neural networks (CNNs). Networks trained with raw voice data can recognize representations that are more discriminative, contextual, and general [38]. It has been demonstrated that CNNs may approach structural localization in the feature space [42], in part by pooling inside a biased frequency zone to reduce translational variance and account for disturbances and minute changes in the feature space. By changing their past interpretation of the speech signal, they can take advantage of the considerable links between speech frames. According to Soltau et al. [43], CNNs have a substantial impact on the performance of ASR systems since these models cannot withstand the strain of semi-clean data.

In contrast, RNNs are useful for applications that must function in noisy environments because they improve identification accuracy by accumulating additional contexts. The inability of RNNs to learn temporal dependencies is hampered by gradient fading and over-inflating concerns. Long short-term memory (LSTM) uses a memory block as a

specialized component to regulate the flow of data, hence eliminating the previously listed issues [42]. Due to their sensitivity to static input, LSTM-RNNs are capable of producing feature-specific target pauses. Least-squares-transfer-memory (LSTM) is a complex network component with a memory function that enables it to store data for extremely long durations. Input, forget, and output gates are the three types of LSTM gates that enable selective retrieval of previously stored data. The input entry of the cell unit determines if data enters the unit, the forgetting entry determines if data from the previous point is forgotten, and the output entry determines if data from the previous point is transmitted to the next point. For reliable acoustic modeling results, there must be little latency between data collection and analysis. The two-way LSTM architecture was created to better analyse input sequences and derive conclusions.

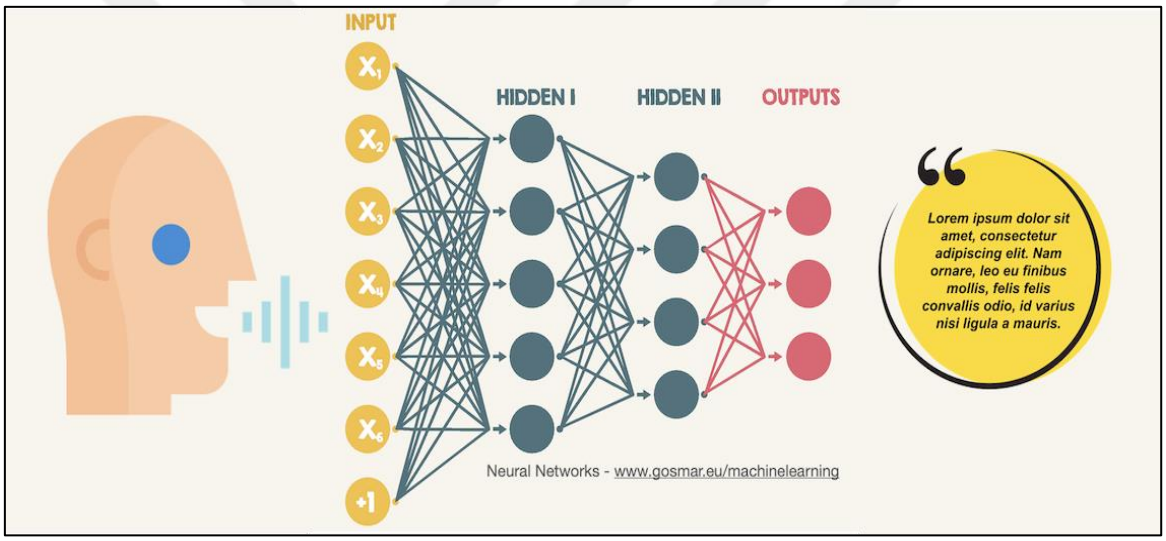

Figure 2.2: Neural Network With Speech Recognition

End-to-end (E2E) trained ASR models, which can instantly turn an incoming acoustic speech input into a word sequence or grapheme, are the subject of significant research. Google and Microsoft, two giants in the IT industry, are devoting significant resources to developing end-to-end automatic speech recognition (ASR) systems. In 2017, Prabhavalkar et al. studied the RNN transducer (RNN-T), attention-based models, and a novel model that combines the RNN transducer with attention. Additionally assessed was the CTC-trained system's direct grapheme sequence output. Utilizing nearly 12,500 hours of training data, they demonstrated that sequence-to-sequence techniques can compete on dictation test sets with a powerful baseline system. Rao et al. [30] examined the efficiency

of sequence-to-sequence RNNs in the same year. T's This model comprises of a decoder partially initialized by a recurrent neural network (RNN) language model trained on text input and an encoder initially initialized by a convolutional tree classifier (CTC) (CTC). Combining linguistic and textual data can improve E2E ASR performance. Using a CTC pre-trainer on the RNN-T encoder increases WER by 5%, while utilizing an 8-layer encoder instead of a 5-layer encoder increases WER by 10%. Google's He et al. [31] developed an E2E ASR that is approximately 20% faster than real-time on a Google Pixel phone when compared to a strong embedded baseline system. Li et al. of Microsoft's Speech and Language Group developed an RNN-T model encoder trained with CTC, or Cross-Entropy, in 2020. (CE). The RNN-T encoder's WER increased by 12.0% when using the model with future context, but decreased by 11.6% when using the CE initialization, relative to when using the zero-lookahead model. The algorithm was taught with over 65 thousand hours of Microsoft data transcriptions.

Despite the enormous study into speech recognition, only a percentage of the outcomes have proven useful to Central Asian languages. Using the Kaldi toolset and a DNN model, Mambare et al. [44study] produced ASR for Kazakh speakers. By 2021, artificial intelligence (AI) researchers have developed an end-to-end (E2E) ASR model for the Kazakh language employing an encoder-decoder with an attention architecture. This system is comprised of three components: a collaborative network, a prediction model, and an encoder. We supplied it data from a 300-hour Kazakh voice corpus to train it. Even though the corpus comprises 300 hours of Kazakh speech, it is not available to the public and cannot be utilized to train accurate E2E ASR models. [46] According to Khazanov et al. Over 332 hours of audio and 153,000 utterances from speakers of diverse ages, gender, and geographic areas make up the open-source Kazakh voice corpus, which tries to overcome these limits. Beginning with official papers, online publications, blogs, and encyclopedias, they acquired Kazakh-language content from a range of additional sources. The retrieved dialogue was then uttered using a web-based, mobile-friendly speech-recording application. In the past, academics have attempted to create speech recognition systems for the Uzbek language. Specifically, it was possible to acquire a preliminary Uzbek speech corpus [8] with 3,500 distinct utterances. Based on hidden Markov models and deep neural networks, Musaev et al. [33] presented an ASR system for Uzbek in 2021. Over 105 hours of voice transcription data were used to train the suggested model. There

are no comparable free or inexpensive Uzbek voice corpus datasets. Therefore, it is crucial to create an ASR system employing a large corpus of publicly available Uzbek speech.

3. METHODOLOGY

The flow diagram of the interactive call system developed in this study is shown in Figure 3.1. given. in diagram shown "sound analysis" blocks deep learning based a with the approach call its center caller of the person his speech analysis by inference provides.

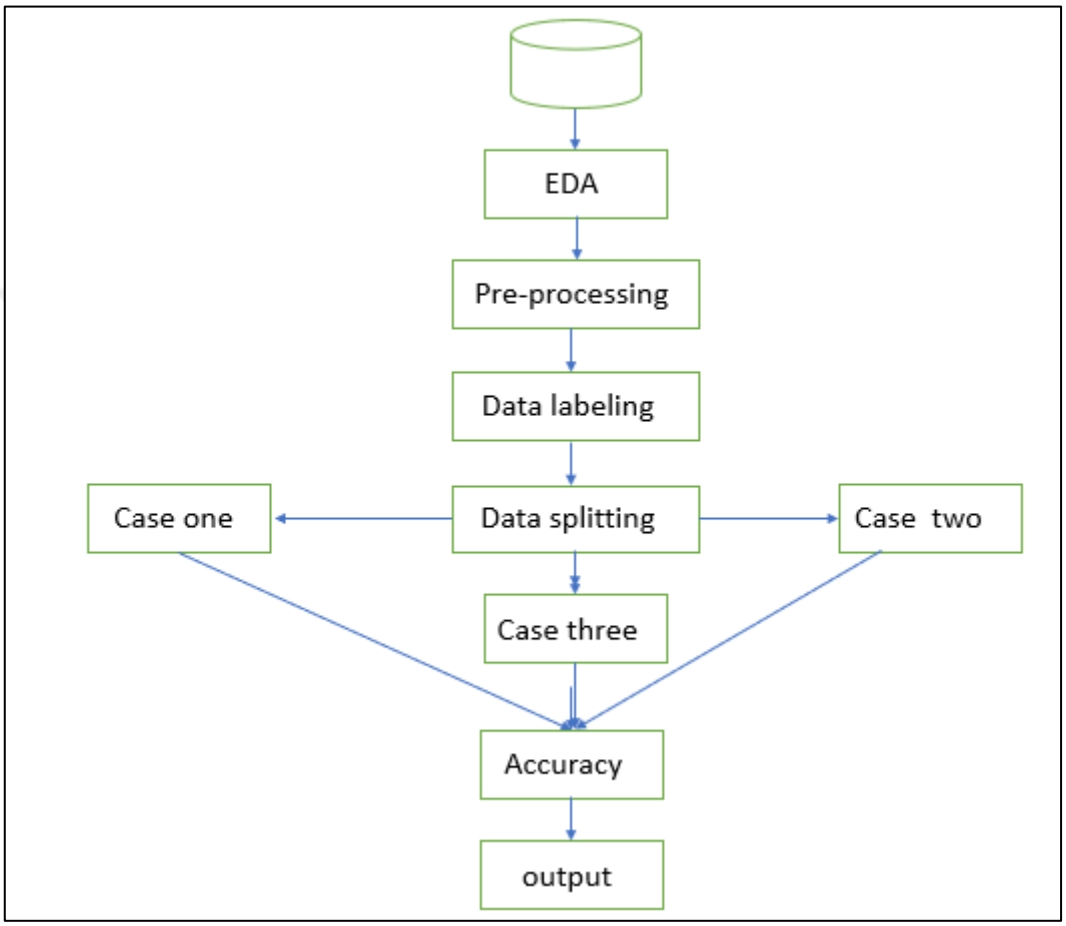

Figure 3.1: Suggested Of The System Flow Scheme.

3.1 DATASET DESCRPTION

Gender Recognition by Voice and Speech Analysis

This database was created to identify a voice as male or female, based upon acoustic properties of the voice and speech. The dataset consists of 3,168 recorded voice samples, collected from male and female speakers. The voice samples are pre-processed by acoustic

analysis in R using the see wave and tune R packages, with an analyzed frequency range of 0hz-280hz (human vocal range). We can see dataset in this link: https://www.kaggle.com/datasets/primaryobjects/voicegender.

3.2 MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

Mel-frequency Cepstral (MFK), short-term in the power spectrum of an audio signal is to be shown. MFKK, linear non- a Mel-scale frequency on you a log strength It is based on the linear cosine transform of the spectrum. MFKK, an audio signal, Using MFK analysis (Xu et al., 2004) derived from a cepstral type representation collected are the coefficients. Mel-frequency and Cepstral between difference, human hearing system of

inspired by much greater response than linear frequency bands in normal cepstral in MFCC found Mel-scale intermittent frequency are bands. Figure 3.2, MFCC attribute extraction process shows.

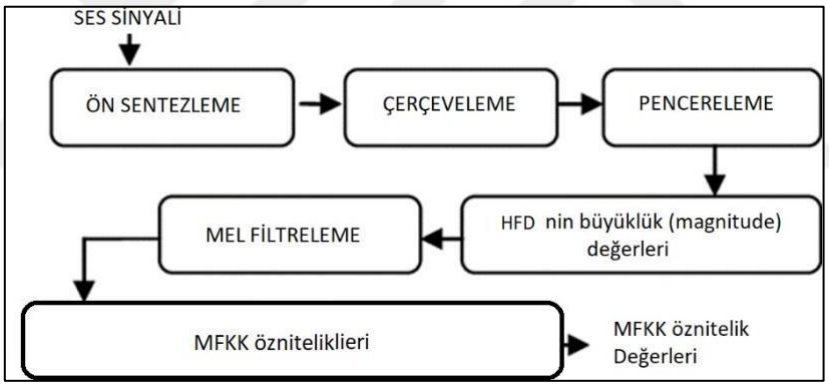

Figure 3.2: MFCC Algorithm Flow Diagram

MFCC algorithm, sound from the signal attributes to take out oriented the following seven from my step consists of:

i. Front synthesizing: Front synthesizing, sound signal higher to frequencies raise It's a filter for. Compared to low frequencies, higher frequencies Due to its small amplitudes, balancing the frequency spectrum there are benefits. This is also when the fast Fourier transform operation is applied. emerge coming out digital problems prevent, and signal-noise rate (SNR) improve for used. Front emphasis, x to the signal equation in 3.1 as applies:

$$y(\ ) = x(\ ) - ax(\ -1) \longrightarrow \quad (3.1)$$

Typical values of the filter α coefficient are between 0.95 and 0.97. pre- emphasis filter has a significant impact on modern systems, as it simply modern in applications of HFD digital their problems prevent for average using normalization obtainable.

ii. Framing: To divide the signal into short-term frames, after the pre-emphasis, framing my name used. This my name, in the signal your frequencies in time change used because. The basic logic of framing is that frequency lines the application of HFD to the entire speech signal to prevent signal loss in logical is not. Typical frames period 20ms-40ms are among and There is 50% overlap between consecutive frames. Using framing, sound in the signal your frequencies short a duration along is fixed we can assume thus, Fourier transformation by doing frequency to the lines oriented good a prediction get we can.

iii. Windowing: The framing next, each to the square, sound signal Hamming window, below stated applied as follows:

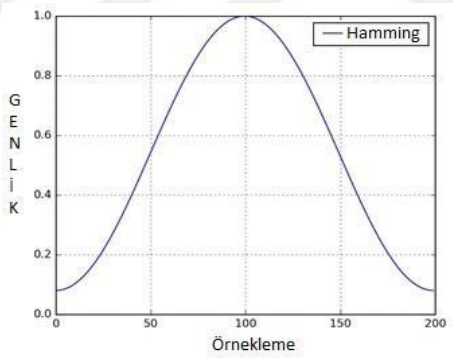

Figure 3.3: Hamming Window

The reason for applying the Hamming window is the fast Fourier transform (HFD) by your signal still is conjecture prevent and spectral leak is to reduce.

iv. fast Fourier transform: One per frame to calculate the frequency spectrum N- point HFD was performed. Typically, the N value is 256 or 512. Each After removing the HFD from the frame, the power spectrum is shown in equation 3.3. as calculated:

v. Mel Filter: The Mel Filter is the final step and is typically triangular into 40 Mel- scale filters. calculated by applying filters. Mel-scale filter stack, human ear sounds imitate perception.

Each filter, center at frequency a to reaction owner the one which... triangle is in the form and linear aspect to zero falls. Figure 3.4, Mel-scale a filter pile of shows.

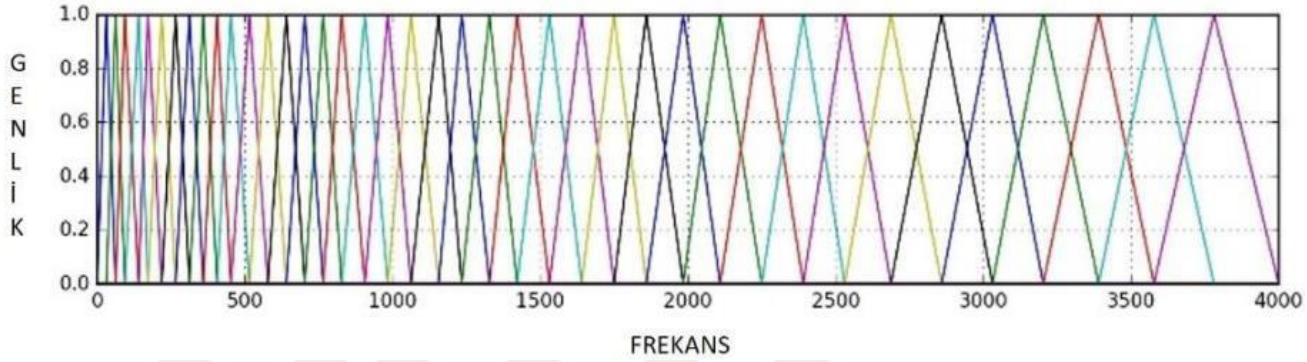

Figure 3.4: Mel-Scale Filter Pile

vi. MFCC Attributes: Filter stack output is highly correlated, this is posing a problem for many machine learning algorithms, so the filter correlating the stack coefficients and compressing the representation of the filter for Discrete Cosine Transformation AKD (Discrete Cosine transform (DCT)) is applied. auto sound in the recognition system general aspect 2-13 coefficients used.

3.3 HIDDEN MARKOV MODEL (HMM)

The hidden Markov model is the most widely used in modern voice recognition systems. It is technical and a useful model for time-varying spectral features. method. End a few front yearly times within, almost all OST It is used as a basic framework in systems (Evermann et al., 2004; Soltau and arc., 2005; Matsoukas and arc., 2006). Modern HMM based sound recognition systems, research at Carnegie Mellon University and IBM in the 1970s group by was created.

Hidden Markov models, time Series data to model for It is a special type of Bayesian network used. These are the probability on the observation sequences. they represent their distribution (Ghahramani, 2001; Murphy, 2002). Saklı Markov model, name, two descriptive from the feature gets. Of these in the first, SMM, It can be thought of as a

double stochastic model. Also, as a random model known stochastic a model, time inside Some random of variables or represents the evolution of systems. deterministic that can only develop in one direction (random not) your models on the contrary, stochastic models a quantity random develops. A in SMM stochastic a model, a soap opera observation aspect while being observed other a stochastic model has been hidden (is in condition). Hidden stochastic model, Only can be evaluated with observations (Cappé et al., 2006). Second defining feature assumes that the hidden state satisfies the first-order Markov property, other words, the state at time t depends only on the state at time t − 1 and from him before all from situations is independent. observations same in time It also fulfills a first-order Markov property of states, a other in words, a Considering the situation, corresponding to this observation, other all from situations and from observations is independent (Ghahramani, 2001).

The working principle of the HMM based voice recognizer is shown in Figure 3.5.Here, the input is an attribute of the spoken word using the MFCC algorithm. converted to order sound is the signal.

$$Y_{1:T} = \{ \ldots \} \tag{3.2}$$

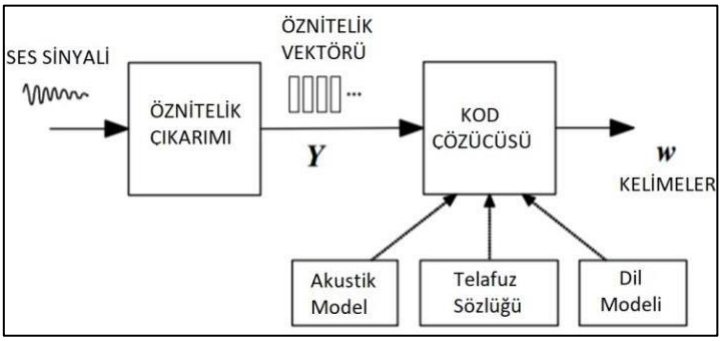

Figure 3.5: Hidden Markov Model

Code solvent, a word string on works.

$$w_{1:L} = \{ \ldots \} \tag{3.3}$$

Y producing possibility of highest the one which... equation in 3.4 shown:

$$w = \underset{w}{arg\,max}\{ P(Y \mid w) P(w) \} \tag{3.4}$$

p(Y |w) possibility of, acoustic model with and P (w) tongue with the model determines. Each a sound unit of, acoustic model thanks to simple a sound aspect is shown, for example, a cat, /k/, /to/,

/D/, /I/ shaped is displayed.

Any a w word for, opposite incoming model words dictionary aspect to create for phonemes by combining synthesized.

of model's parameters, said your words sound signal including education data is estimated using Usually language model, each word probability is only N - one to the premise connected the one which... a N – gram from the model occurs.

In Figure 3.6 a Situations in SMM and observations dependencies between A graphical representation is shown. Here, the observations are in circles starting with O and hidden. situations if S with starting with circles shown.

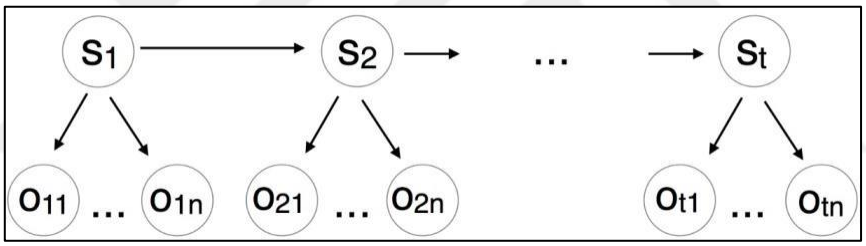

Figure 3.6: Situations Dependencies Between And Hidden Markov In The Model Observations (Lisson 2014)

3.3.1 HMM Elements

a. Situation Set: S = {S $_1$, S $_2$, ..., S $_N$}, here |S| = n. t at the time situation qt _ = S $_i$ ∈ S is displayed.

b. Set of observation symbols: V = {v $_1$, v $_2$, ..., v $_M$ }, here |V | = M.t.at thetime observation symbol he $_t$ is displayed with .

c. Transition Possibilities: A = {$_{aij}$ }, where {a $_{ij}$ } = P[q $_t$ +1 = $_{Sj}$ |q $_t$ = S $_i$],one ≤ I, j ≤N ($_{Si}$ from the state S $_j$ to the state transition possibility of).

d. Emission Possibilities: B = {$_{bj}$ (k)}, here b $_j$ (k) = p[O $_t$ = v $_k$ | q $_t$ = S $_j$], one ≤ j≤ N, one ≤ k ≤ M (Sj from the state v k symbol

e. Initial State Probabilities: π = {π $_i$} where π $_i$ = P(q $_1$ = Si), 1 ≤ $_i$ ≤ N(Theinitial

state S_i being possibility of)

Considering the elements of an SMM, SMM is expressed as λ = (A, B, π) can be, here N and M, indirect aspect A and B by representation is done. A SMM, Moreover, requires a set of observation symbols: O = O1O2...OT. An SMM λ and an observation sequence, O, Considering the first-order Markov properties of states and observations, a rearrange the joint distribution of a set of states and observations

3.3.2 Forward Directional Variables

Forward probability (αt(i)), partial observation at time t, considering the λ model is the probability of the O1O2...Ot and Si state. In other words, the state sequence, t to the time until observed, at time t a Si in case of to the end don't melt is the probability.

$$\alpha_t(i) = (O_1 O_2 O_3 \ldots , q_t = S_i |\lambda) \quad (3.5)$$

All time in your steps, all your situations Further the possibility of calculate for dynamica programming approach used. Forward possibility (αt(i)), the following as, dynamic using programming inductive aspect

This dynamic programming approach for advanced variable calculation, length T and can be viewed as a state lattice of N dimension. In Figure 3.7 the state lattice presents visually. The dynamic programming approach is advanced in the state lattice. variable HE (TN2) timely calculation provides.

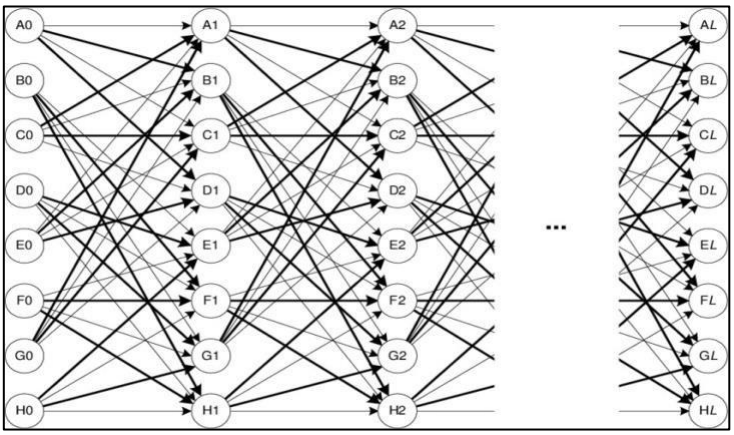

Figure 3.7: Situation Cage Example.

3.3.3 Back Directional Variables

Back directional possibility (βt(i)), I status and t the time and λ model eyelash before taken, observation of the series t + one from time to the end until the one which... is the probability. This, t + one

from time to the end until a observation the series, t at the time a Si in case of starting from the possibility of observing.

Forward directional to the variable similar way, all time in your steps all your situations backA dynamic programming approach is used to calculate the directional probability. Back directional probability (βt(i)) inductively using dynamic programming.

3.4 ARTIFICIAL BORDER NETWORKS

To give general information about the general structure of a neural network, border into their nets this in the section Location will be given. in 1989 Dr. Robert Hecht-Neilsen border their webs "… a set of simple, each other high level connected process from the elements formed a computer system" aspect defined (Caudill, 1989). A classic yet simple node in a neural network type, McCulloch-Pitts is the node (McCulloch and pits, 1943). This your node or Representation of a neuron by the name McCulloch and Pitt prefer to express them. It can be seen in 3.8.

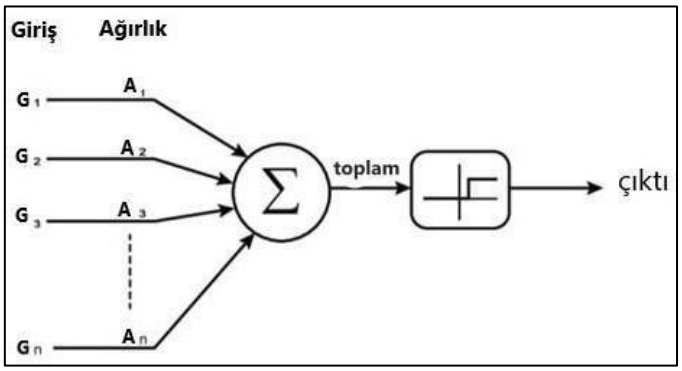

Figure 3.8: Classical A Border Network Node The One Which... Mcculloch-Pitts Neuron (Chakraborty And Sarddar)

A border in your network, Figure in 3.9 seen as three sort layer has. Login layer, are the nodes through which we feed the input data. The output layer is the input layer given in the

input layer. data in line with network by produced is the output. new frequencies prediction to make Because we want it, in our project, the output layer's output size is the input layer's size, that is, it must be equal to the number of nodes in this layer. The middle layer is the hidden layer is named. Simple way expression to make if necessary hidden layer size and number, your network What until information distinguish that you can determines. This border your network basis is the structure.

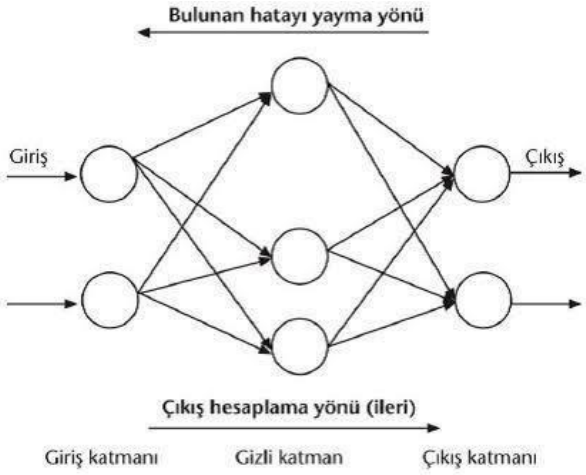

Figure 3.9: Simple A Feedforward Nerve Your Network Sample Structure. (SA) (Derinogrenme.Com, 2017)

Now we will explain how the network can learn. Each layer is a node relates to the series. From each node in a layer, it is divided into one, as shown in Figure. 3.9. There is a terminal connecting each of the nodes in the next layer. each node, from the ends incoming to the inputs based on previously defined a calculation it does. McCulloch-Pitts at the node made calculations basis aspect a sigmoid is the function. We add the inputs and output if the total input is above a certain threshold, we produce 1 as the value, 0 otherwise. As we shall see later, these nodes are more complex. There are representations of the McCulloch-Pitts neuron to understand the basics of neural networks. It is a good starting point for In Figure 3.9, in combination with these nodes, tips arrows with shown. to these ends more widespread so to speak your network weights is called.

Basis aspect made work, a neuron print out own value with bump and nextthe result is to feed the next node to which its end is connected by inputting it. produced output in line with this weights by updating network which entry of your data which distinguish it should

produce output data We can teach you how. Update weights back to process spread (back propagation) First Name is given.

3.4.1 Back Spread

The back propagation part of the training of a neural network takes place where the learning takes place. is the place. This is where the entire network is used to get the correct output for a given input. along their weight has updated is the place. Back spread of the algorithm inner to the structure oriented studies (Rumelhart et al., 1985) and the concept in this section We will make a general statement about Backpropagation returns each exit node to the expected output with we compared from the output starts. in hand we did output with expected output between difference in many ways computable. This calculation used for classic a method, is the mean square error. The function used to calculate the error is our loss.is called our function. With the back propagation algorithm of the loss function

It must be differentiable for it to work. Next time he re-entered the network when given expected to the output more near output values get to do now in the network all We want to update the weights. The ends coming to the output node of the loss-to-work function We start by taking the partial derivative with respect to Each derivative is the output of the loss function, each input indicates how much it depends on its weight. The weights are now updated and the error of the activation functions in the nodes of the previous layer with the same method can be updated. This update again up to the input layer to update the whole network. again makes.

3.5 MACHINE LEARNING

Thanks to the advances in computing capacity over the last 10-15 years, machine in learning new a area emerge out. This to the field deep learning is called. Concept aspect deep learning, machine to learn inside many include the algorithm. However, in terms of neural networks, deep learning is generally output and There are too many hidden layers between the input layer, where each layer consists of many sets of nodes. are seen as very large models consisting of layers. In the deep learning field, the main reason for the progress experienced is the graphics processing units (GPUs) of deep learning. It is applied

for general purpose calculation purposes neural networks Two main advantages of GPU computing over CPUs are found. The first is rapid access to the processing unit of large amounts of data, thus energetic data additional burden decreases. Other supremacy if of GPU made calculation

types are optimized. GPUs are well optimized for computational matrices been taken to functions has. Artificial neural your networks their training for large extent matrix to transactions need from being heard because, GPU more fast process to be made facility gives.

3.5.1 Support Vector Machines

Support vector machines are a form of machine learning based on the notion of structural risk minimization. The field of pattern recognition is just one use of this technology [47]. SVM estimates the decision function by constructing a support vector linear model. If training data can be divided into two groups along a linear dimension, SVM can decide which of two hyper planes divides the information most effectively [48]. To construct an ideal hyper plane of separation, SVM frequently incorporates non-linear mappings, as seen in Figure 2. The input patterns are mapped into higher dimensional feature space via SVM in a non-linear fashion. It is possible to classify data sets using a linear SVM if they are linearly separable [49]. Support vectors are the patterns that are optimized for and lie along the margins.

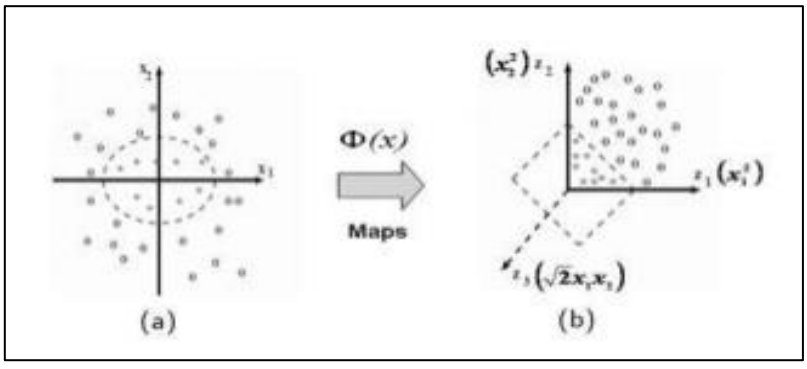

Figure 3.10: Example for SVM.

After being (converted into) support vectors, training patterns are evenly distributed along the separation hyperplane. Support vectors, which are utilized as training samples to determine the ideal hyperplane, are the most challenging patterns to categorize [50]. These

are the patterns that provide the most information about the categorization task. [51] When developing machines having a variety of non-linear decision surfaces in the input space, the kernel function is utilized to generate the inner products.

3.5.2 Decision trees

The rules that this model generates support the choice to employ it. The decision tree method is simple and easily understood. It demonstrates that machine learning is applicable to the analysis of symbolic objects [52] and is not restricted to the field of statistics. It is crucial to employ machine learning techniques in the event of a traffic accident because their major function is to provide actionable recommendations. The most well-known methods for creating the best tree are the ID3 algorithm [53], designed for nominal characteristics, and its successors C4.5 and C5 [54], which also handle quantitative elements.

3.5.3 K-Nearest Neighbors

As a second experiment, we employ the k-nearest neighbors technique to machine learning. The second technique, on the other hand, is unaffected by background noise. Several pattern recognition applications [55, 56] utilized it. Unfortunately, it has exceptionally high storage and processing requirements for large volumes of data. Its algorithm is based on the notion of similarity. It is a learning algorithm that is believed to be extremely user-friendly. Once the algorithm votes on its most similar samples based on a specified distance, the class of the new sample is determined by a majority vote among these k most similar samples (nearest neighbors) [57]. There is a substantial association between the metric employed and KNN performance; hence, selecting a good k value is crucial [58].

3.5.4 Random Forest

The proposed method for supervised speaker diarylation in dyadic psychotherapy is based on the algorithm for random forests. Random forests are an algorithm for machine learning that is simple to understand and is now employed in the field of psychotherapy [60]. While there has been an increase in interest in machine learning in general, the random forest technique uses decision trees as its basis as a machine learning classifier. Random forest,

also known as an ensemble learner, integrates the outputs of a predetermined number of decision trees into a single predictive model. It is helpful for resolving classification and regression problems. The ensemble approach of generating classification decisions enhances accuracy since it polls all of the trees in the set, as opposed to just one .The primary benefit of the random forest method is its resistance to multicollinearity in input data and variables that do not influence classification accuracy. This is a pertinent finding given that we do not know which features will be significant for which pair, and the likelihood that speech quality are closely related. Each random forest tree is created using a unique selection of variables and data points from the training set. Bagging is the method of randomly picking a subset of data items for use in training individual trees without replacement. With this bagging procedure, the out-of-bag error rate may be computed, providing for an approximation of the generalization error. We can determine, for each item in the learning set, which trees do not employ that item when acquiring knowledge. These classifiers are often known as "out of the bag" classifiers. The error produced by a classifier when evaluated using only the training data is known as the out-of-bag error. For our purposes, it is advantageous to be able to compute the generalization error without using data from outside the learning set. Applying this method on unprocessed, real-world audio from psychotherapy sessions and computing the out-of-bag error on the learning data, we can evaluate the overall strength of the prediction in each dyad and determine for which dyads the diarylation performed well and poorly. Hence, we examine the relationship between the dyadic out-of-bag error and the dyadic speaker error, as evaluated by a separate test set [61].

3.5.5 Logistic Regression

Logistic Regression is a classification that serves to solve the binary classification problem. The result is usually defined as 0 or 1 in the models with a double situation.

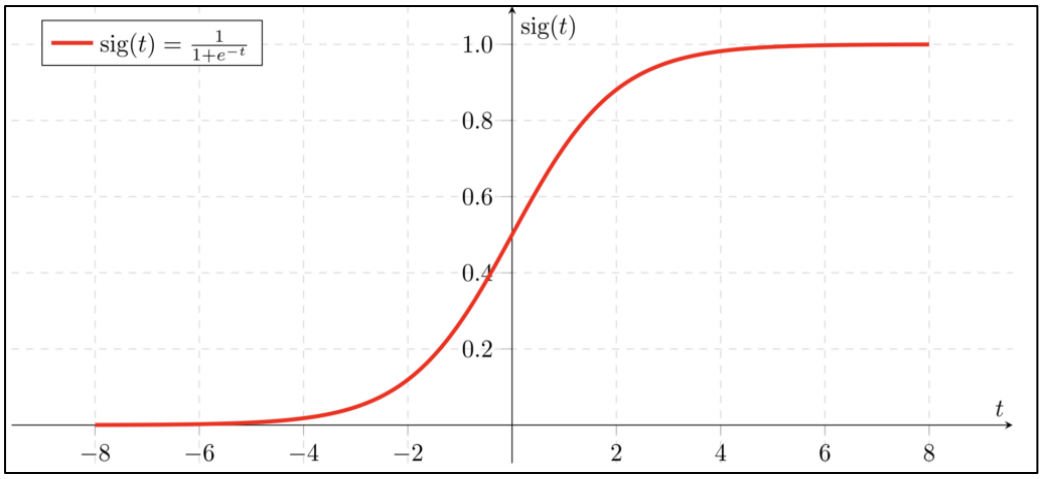

Figure 3.11: Example Logistic Regression.

3.6 GOOGLE CLOUDY SOUND RECOGNITION

Google Cloudy, use of easy a application interface with (API- Application interface) deep learning based border network models by applying of developers her voice to the text enables conversion. This API recognizes 120 languages and variants globally. We recommend in the system, more more process to be done in the name of, spoken phrases to the text convert for Google cloudy sound recognition API (Application program Search face) we used Google by algorithm used, this Nural network described earlier in the chapter, Nurul network, K- nearest neighbors, Support vector machine Classifier, Random Forest Classifier, Logistic regression Classifier, decision tree-based machine learning algorithm. The following Flow diagram, this function in our system How that we apply shows.

The Google Cloud structure can also operate over real-time streaming. Theisman audio encodings in architecture, including FLAC, AMR, PCMU and Linear-16 is supported. Moreover, automatic tongue perception option also available. With this together without the need for additional sound reduction processes within the scope of the noise immunity system, many in the environment noisy sound data can operate. Unsuitable content filtering feature thanks to Some in languages text in the results unsuitable contents can be filtered. Within the scope of automatic punctuation, dot, comma, etc. Punctuation marks can also be added. Google Cloud structure with model selection feature can include models that we have trained ourselves as well as pre-trained models. Thanks to the speaker parsing

module, which of the people in the speech consumption to what you do about automatic predictions gives.

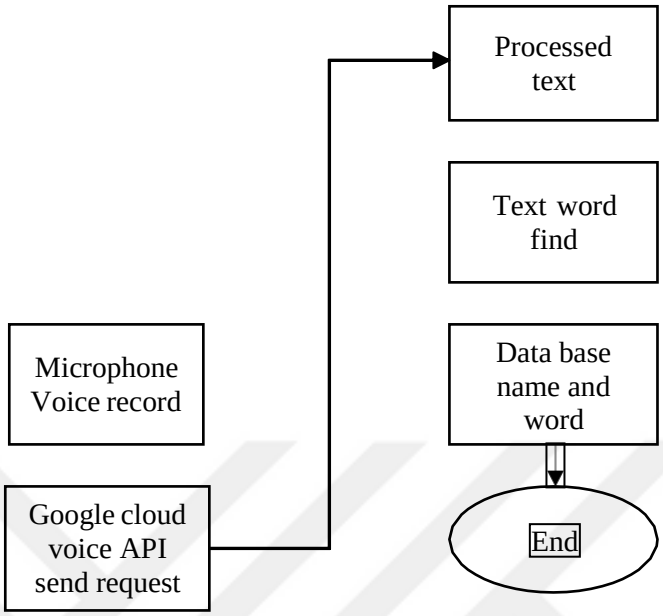

Figure 3.12: Google Cloudy Speech API Search Your Face In Our System Use Flow Scheme.

As shown in the flowchart, the program first scans the authorization code. using Google cloudy sound to its API connects and Initialization process performs. It then records the audio using the microphone and sends it as a request to the Google Cloud. sends it. Split first by taking Google Cloud rendered text in pre-trained structure process it does. More Then dividing text in the database name and in words by comparing results returns.

4. RESULTS

We'll give a quick overview of the information before we delve in deeper. Using several parameters, we have a dataset that can determine the gender of a speaker's voice. The best results can be achieved if we study how humans perform. The eardrum is the final stop for sound waves that pass through the ear canal. The ossicles in the middle ear conduct the vibrations from the eardrum to the cochlea in the inner ear. The snail-shaped inner ear, often known as the cochlea, plays an important role in hearing and balance. Thousands upon thousands of microscopic hair cells can be found in the cochlea. To relay auditory information to the brain, hair cells convert sound waves into electrical impulses. The brain identifies the sound and lets you know what it is.

Neurons in the brain undertake operations to categorize sounds; our proposal would mimic this process utilizing machine learning and custom-built neural network techniques to achieve the same result.

Following the importation of the libraries and the loading of the dataset, we carry out EDA, also known as exploratory data analysis, on the dataset. In order to analyze and comprehend the dataset, as well as its characteristics and target classifications and Distribution of target variables. The following Figure 4.1 describes distribution of target variables for both male and females.

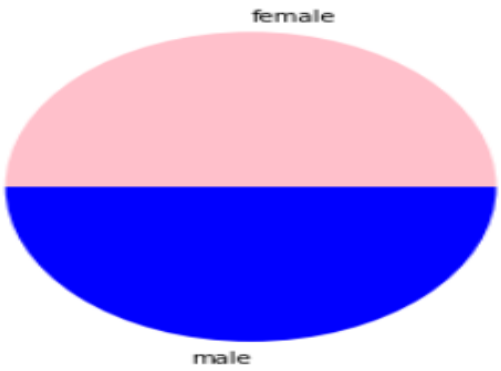

Figure 4.1: Distribution Of Target Variables.

Plots are generated from analyses of correlations between characteristics, making it easy to see how variables are related. Figure 4.2 shows correlation matrix.

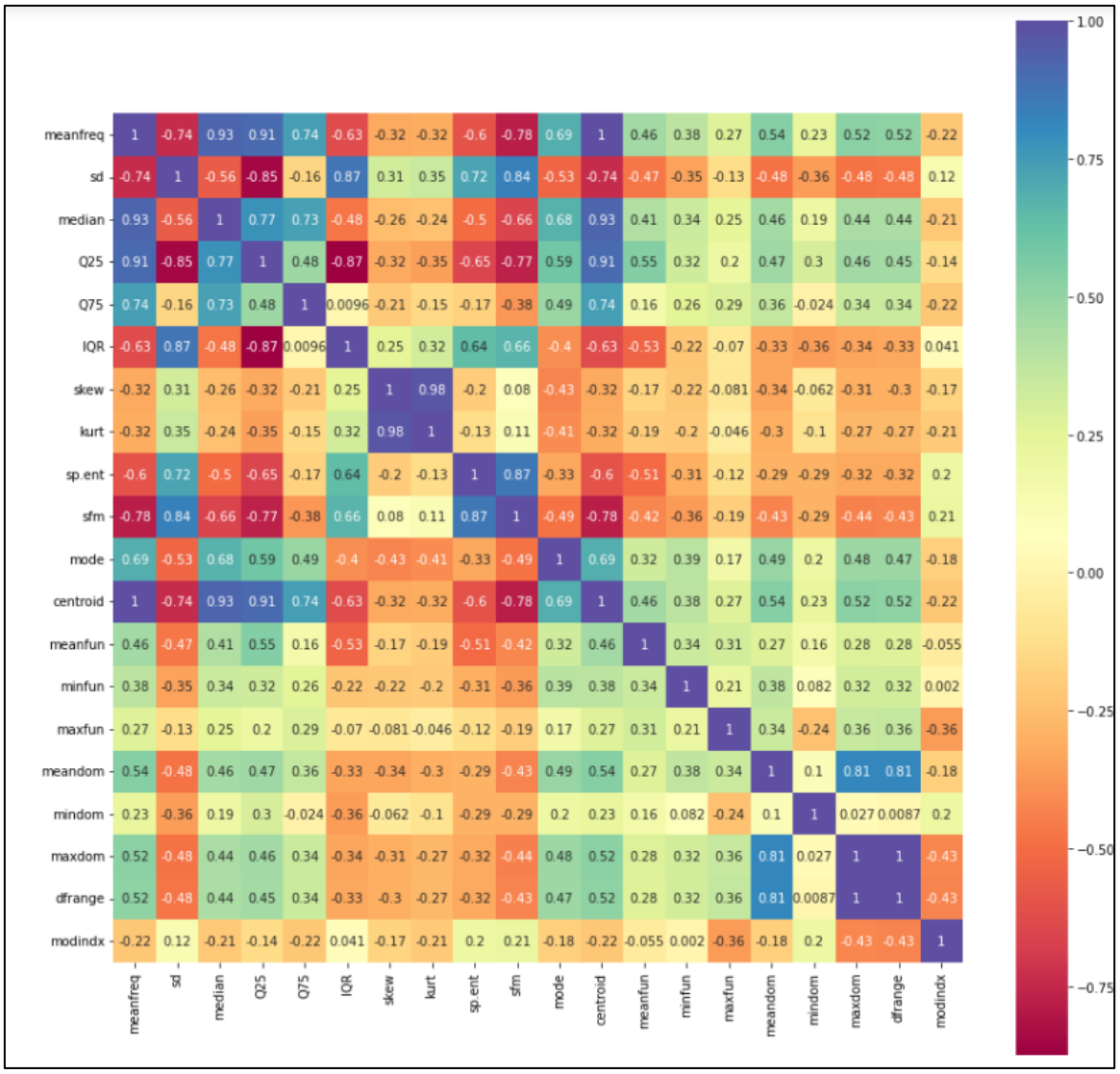

Figure 4.2: Correlation Matrix

In order to conduct the real classification, a model consisting of two stages of classification was built.

4.1 CASE ONE

Using the machine learning tools, we begin by utilizing a collection of supervised classification algorithms. This is the initial step. In order to get the data ready for training the models, they need to be segmented into two parts: one for training 80%, and the other for the actual experiment and test 20%. After the step of categorizing the algorithms, we proceed to evaluate the performance of each one in order to determine the significance of the most effective of the bunch.

4.1.1 Decision Tree Classifier

With this classifier the predicted array of first 10 values were as ['female' 'male' 'female' 'female' 'male' 'male' 'female' 'female' 'male' 'female']. The confusion matrix is as shown in tablev4.1.

Table 4.1: Confusion Matrix Of 1-1 Decision Tree Classifier

confusion	female	male	All
female	285	16	301
male	16	317	333
All	301	333	634

Each algorithm is analyzed utilizing assessment techniques that are among the most effective available. When we have finished running trials on all of the algorithms, we will be able to evaluate each one based on how efficient and accurate it is. The findings that are the best available can be suggested. The effectiveness of this method is outlined in Table 4.2.

Table 4.2: Decision Tree Accuracy Score

precision	**recall**	**f1-score**	**support**
female	0.95	0.95	301
male	0.95	0.95	333
accuracy	0.95	0.95	634
macro avg	0.95	0.95	634
weighted avg	0.95	0.95	634

The score for the accuracy of the decision tree was recorded as 94.95%, and the findings were used as the test dataset.

4.1.2 Random Forest Classifier

This algorithm's prediction as prepare for comparison was as ['female' 'male' 'female' 'female' 'male' 'male' 'female' 'female' 'male' 'female'] for first ten values. In table 4.3 we can show the confusion matrix.

Table 4.3: Confusion Matrix Of Random Forest Classifier

confusion	female	male	All
female	295	5	300
male	6	328	334
All	301	333	634

As discussed earlier, the algorithm's efficacy and precision must be understood prior to completing the classification process. In each algorithm, measurement is performed using the same instruments. The Random Forest algorithm yielded the outcomes indicated in Table 4.4.

Table 4.4: Random Forest Scores

precision	**recall**	**f1-score**	**support**
female	0.98	0.98	300
male	0.98	0.98	334
accuracy	0.98	0.98	634
macro avg	0.98	0.98	634
weighted avg	0.98	0.98	634

The result for accuracy using Random Forest was 98.26%. When compared to the method used in the earlier scenario, the accuracy in this situation is significantly higher.

4.1.3 Logistic Regression Classifier

The prediction result with this method was as ['female' 'male' 'female' 'female' 'male' 'male' 'male' 'female' 'male' 'female'] for first ten elements. The confusion for Logistic Regression is showed in Table 4.5.

Table 4.5: Confusion Matrix Of Logistic Regression

confusion	female	male	All
female	257	11	268
male	44	322	366
All	301	333	634

To illustrate our point, we compared the performance of five different algorithms in the first scenario, and then compared those results to those produced by a neural network. Using Table 4.6, we will determine the efficacy of the classification system and the accuracy with which it performs.

Table 4.6: Logistic Regression Scores

precision	**recall**	**f1-score**	**support**
female	0.85	0.966	268
male	0.97	0.88	366
accuracy	0.97	0.98	634
macro avg	0.92	0.92	634
weighted avg	0.92	0.91	634

A score of 91.32% was obtained for the accuracy of the logistic regression. As compared to the accuracy of the first two algorithms, we can safely state that this one has a lower accuracy.

4.1.4 KNN Classifier

The first ten elements based on KNN algorithms pred parameter was ['female', 'male', 'female', 'male', 'male', 'male', 'male', 'female', 'female', 'female']. Table 4.7 contains the confusion matrix for the KNN algorithm.

Table 4.7: Confusion Matrix of KNN

confusion	female	male	All
female	223	62	285
male	78	271	349
All	301	333	634

Table 4.8: KNN Scores

precision	**recall**	**f1-score**	**support**
female	0.74	0.78	285
male	0.81	0.78	349
accuracy	0.97	0.78	634
macro avg	0.78	0.78	634
weighted avg	0.78	0.78	634

The accuracy score for the KNN algorithm was the same as 77.91 percent. This method provided a classification accuracy that was lower than that of its competitors. Accuracy is the measure of how efficient a classification method is with a given packet of data, and it is included in all classification algorithms.

4.1.5 SVM Classifier

The array of SVM classifier for predicted firs elements was ['female', 'male', 'female', 'female', 'male', 'male', 'female', 'female', 'male', 'female']. The algorithm is also used widely for classification. The confusion matrix with our dataset is showed in table 4.9.

Table 4.9: Confusion Matrix of SVM

confusion	female	male	All
female	258	5	263
male	43	328	371
All	301	333	634

Table 4.10: SVM Scores

	Precision	**recall**	**f1-score**	**support**
female	0.86	0.98	0.91	263
male	0.98	0.88	0.93	371
accuracy	--	--	0.92	634
macro avg	0.92	0.93	0.92	634
weighted avg	0.93	0.92	0.92	634

According to Table 4.10, SVM has a success rate of 92.42%. Its accuracy is above average when compared to earlier algorithms, but it is not the greatest.

4.2 CASE TWO

To aid with categorization, we are developing a neural network. We are using a duplicate of the same data, but it is fresh. The 'label' includes strings, but numbers are needed in mathematics, so we'll change these two Male=1 and Female=0 now that we have the feature set and the set of dependent variables. The dependent and independent factors must be isolated. The characteristics are listed in the first 20 columns of the data set, and the dependent variable, which may be either 1 (Male) or 0, appears in the final column

(Female). Separate the data into a training set and a testing set, the former to be used in teaching the neural network, the latter to be used in validating its training.

Seeing the actual process through which the model acquires knowledge is always crucial. Hence, education occurs at each period. The model can figure out how much of a hit it will take based on the real-world outcome, and it will alter its emphasis accordingly.

The variables used are optimizer='Adam,' loss='binary cross entropy,' and so on. We were able to get a 97.75% accuracy rate using the Neural Network on our train dataset. Furthermore, the Neural Network's accuracy on the test dataset was found to be 96.68%.

4.3 CASE THREE

we are performing an optimization method for our proposed model. The optimization will go throw cleaning the data and reentering them into the algorithms.

The average range for a human's fundamental frequency is between 0.085 Khz and 0.255 Khz, so we'll look for the variable that has this frequency information and remove it if we know it's an outlier.

Based on our research, the fundamental frequency of a typical adult male will be between 85 Hz and 180 Hz, and that of a typical adult female will be between 165 Hz and 255 Hz.

In the next experiment with the algorithms after optimizing the system, the results were as Random Forest Classifier=99.45% and Decision Tree Classifier=99.18%.

the Cross validation with same classifier as first time after the optimization was 96.37% and improved to become 98.98%

.

5. CONCLUSION

5.1 CONCLUSION

Numerous sorts of businesses utilize call center technologies to manage client calls. Most this center's central staff handles incoming calls and routes them to the appropriate locations. Owing to a high call volume, clients are placed on hold for unusually extended times. Bringing additional personnel to the facility is a feasible solution to the problem. This may not be a thorough or cost-effective solution, and it certainly increases the cost. About methodology, this paper intends to use three case studies to illustrate In the first, we implemented the Nurul network, K-nearest neighbors, Support vector machine Classifier, Random Forest Classifier, Logistic regression Classifier, and decision tree-based machine learning method to the publicly accessible dataset (Gender Recognition by Speech) on the Kaggle website. Now that we have the feature set and the dependent variable established, we may alter the Male=1 and Female=0 values, since mathematics demands numbers, but "label" contains texts. To conduct an accurate analysis of a problem, it is essential to identify the dependent and independent elements. The attributes are in the first 20 columns, and the dependent variable, which is either 1 (Male) or 0, is in the last column (Female) (Female). Build a training set and a testing set from the data, then use the training set to instruct the network and the testing set to evaluate its learning. In a third case, we employ an optimization technique to further enhance the efficacy of the model we propose. As part of the process of enhancing the detection of speech events in chaotic metropolitan contexts, the data will be cleaned up. this allows for a considerable degree of accuracy to be reached Comparing the results of the decision tree and random forest classifiers, the former achieves an accuracy rate of 99.18%, while the latter achieves an accuracy rate of 99.45%. The interactive Turkish voice recognition system created for the thesis was examined across several variables, such as age, gender, department, duration, and ambient noise level.

5.2 FUTURE WORK

In our future study, the knowledge gap between humans and computers will reduce in the Uzbek ASR system, and the problem of inadequate resources for the different variations of Uzbek, which currently suffer from a high error rate, will be resolved. In addition, we

intend to enhance our methodology by employing low-rank approximation techniques for comparison analysis.

REFERENCES

[1] De Lima, T.A.; Da Costa-Abreu, M. A survey on automatic speech recognition systems for Portuguese language and its variations. Comput. Speech Lang. 2020, 62, 101055.

[2] Chen, Y.; Zhang, J.; Yuan, X.; Zhang, S.; Chen, K.; Wang, X.; Guo, S. SoK: A Modularized Approach to Study the Security of Automatic Speech Recognition Systems.

[3] Xia, K.; Xie, X.; Fan, H.; Liu, H. An Intelligent Hybrid–Integrated System Using Speech Recognition and a 3D Display for Early Childhood Education. Electronics 2021, 10, 1862.

[4] Ahmad, A.; Mozelius, P.; Ahlin, K. Speech and Language Relearning for Stroke Patients-Understanding User Needs for Technology Enhancement. In Proceedings of the Thirteenth International Conference on eHealth, Telemedicine, and Social Medicine (eTELEMED 2021), Nice, France, 20 July 2021.

[5] Sodhro, A.; Sennersten, C.; Ahmad, A. Towards Cognitive Authentication for Smart Healthcare Applications. Sensors 2022, 22, 2101.

[6] Avazov, K.; Mukhriddin, M.; Fazliddin, M.; Young, I. Fire Detection Method in Smart City Environments Using a Deep-Learning-Based Approach. Electronics 2021, 11, 73.

[7] Khamdamov, U.; Abdullayev, A.; Mukhiddinov, M.; Xalilov, S. Algorithms of multidimensional signals processing based on cubic basis splines for information systems and processes. J. Appl. Sci. Eng. 2021, 24, 141–150.

[8] Musaev, M.; Khujayorov, I.; Ochilov, M. Automatic recognition of Uzbek speech based on integrated neural networks. In World Conference Intelligent System for Industrial Automation; Springer: Cham, Switzerland, 2021; Volume 1323, pp. 215–223.

[9] Qian, Y.; Zhou, Z. Optimizing Data Usage for Low-Resource Speech Recognition. IEEE/ACM Trans. Audio Speech Lang. Processing 2022, 30, 394–403.

[10] Świetlicka, I.; Kuniszyk-Jóźkowiak, W.; Świetlicki, M. Artificial Neural Networks Combined with the Principal Component Analysis for Non-Fluent Speech Recognition. Sensors 2022, 22, 321.

[11] Templeton, G. Language Support in Voice Assistants Compared. Available online: https://summalinguae.com/language-technology/language-support-voice-assistants-compared/ (accessed on 21 April 2021).

[12] He, L.; Niu, M.; Tiwari, P.; Marttinen, P.; Su, R.; Jiang, J.; Guo, C.; Wang, H.; Ding, S.; Wang, Z.; et al. Deep learning for depression recognition with audiovisual cues: A review. Inf. Fusion. 2021, 80, 56–86.

[13] Yu, C.; Kang, M.; Chen, Y.; Wu, J.; Zhao, X. Acoustic modeling based on deep learning for low-resource speech recognition: An overview. IEEE Access 2020, 8, 163829–163843.

[14] Aldarmaki, H.; Ullah, A.; Zaki, N. Unsupervised Automatic Speech Recognition: A Review. Speech Commun. 2022, 139, 76–91. [Google Scholar] [CrossRef]

[15] Ayvaz, U.; Gürüler, H.; Khan, F.; Ahmed, N.; Whangbo, T.; Abdusalomov, A. Automatic Speaker Recognition Using Mel-Frequency Cepstral Coefficients Through Machine Learning. CMC-Comput. Mater. Contin. 2022, 71, 5511–5521.

[16] Yu, J.; Zhang, S.X.; Wu, B.; Liu, S.; Hu, S.; Geng, M.; Liu, X.; Meng, H.; Yu, D. Audio-visual multi-channel integration and recognition of overlapped speech. IEEE/ACM Trans. Audio Speech Lang. Processing 2021, 29, 2067–2082. [Google Scholar] [CrossRef]

[17] Deena, S.; Hasan, M.; Doulaty, M.; Saz, O.; Hain, T. Recurrent neural network language model adaptation for multi-genre broadcast speech recognition and alignment. IEEE/ACM Trans. Audio Speech Lang. Processing 2018, 27, 572–582. [Google Scholar] [CrossRef][Green Version]

[18] Wali, A.; Alamgir, Z.; Karim, S.; Fawaz, A.; Barkat Ali, M.; Adan, M.; Mujtaba, M. Generative adversarial networks for speech processing: A review. Comput. Speech Lang. 2021, 72, 101308. [Google Scholar] [CrossRef]

[19] Zhang, W.; Chang, X.; Qian, Y.; Watanabe, S. Improving end-to-end single-channel multi-talker speech recognition. IEEE/ACM Trans. Audio Speech Lang. Processing 2020, 28, 1385–1394. [Google Scholar] [CrossRef]

[20] Mukhiddinov, M. Scene Text Detection and Localization using Fully Convolutional Network. In Proceedings of the 2019 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 1–5 November 2019. [Google Scholar]

[21] Reddy, V.R.; Rao, K.S. Two-stage intonation modeling using feedforward neural networks for syllable based text-to-speech synthesis. Comput. Speech Lang. 2013, 27, 1105–1126.

[22] Bhattacharjee, M.; Prasanna, S.R.M.; Guha, P. Speech/Music Classification Using Features from Spectral Peaks. IEEE/ACM Trans. Audio Speech Lang. Processing 2020, 28, 1549–1559.

[23] Koutini, K.; Eghbal-zadeh, H.; Widmer, G. Receptive Field Regularization Techniques for Audio Classification and Tagging with Deep Convolutional Neural Networks. IEEE/ACM Trans. Audio Speech Lang. Process. 2021, 29, 1987–2000.

[24] Ibrahim, A.B.; Seddiq, Y.M.; Meftah, A.H.; Alghamdi, M.; Selouani, S.A.; Qamhan, M.A.; Alshebeili, S.A. Optimizing arabic speech distinctive phonetic features and phoneme recognition using genetic algorithm. IEEE Access 2020, 8, 200395–200411.

[25] Mukhiddinov, M.; Akmuradov, B.; Djuraev, O. Robust text recognition for Uzbek language in natural scene images. In Proceedings of the 2019 International Conference on Information Science and Communications Technologies (ICISCT), Chongqing, China, 1–5 November 2019.

[26] Kourkounakis, T.; Hajavi, A.; Etemad, A. FluentNet: End-to-End Detection of Stuttered Speech Disfluencies with Deep Learning. IEEE/ACM Trans. Audio Speech Lang. Process 2021, 29, 2986–2999.

[27] Narendra, N.P.; Rao, K.S. Parameterization of Excitation Signal for Improving the

Quality of HMM-Based Speech Synthesis System. Circuits Syst Signal Process. 2017, 36, 3650–3673.

[28] Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.R.; Jaitly, N.; Kingsbury, B. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Mag. 2012, 29, 82–97.

[29] Prabhavalkar, R.; Rao, K.; Sainath, T.N.; Li, B.; Johnson, L.; Jaitly, N. A Comparison of Sequence-to-Sequence Models for Speech Recognition. Interspeech 2017, 2017, 939–943.

[30] Kanishka, R.; Haşim, S.; Rohit, P. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 193–199.

[31] He, Y.; Sainath, T.N.; Prabhavalkar, R.; McGraw, I.; Alvarez, R.; Zhao, D.; Gruenstein, A. Streaming end-to-end speech recognition for mobile devices. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6381–6385.

[32] Li, J.; Zhao, R.; Meng, Z.; Liu, Y.; Wei, W.; Parthasarathy, S.; Gong, Y. Developing RNN-T models surpassing high-performance hybrid models with customization capability. arXiv 2020, arXiv:2007.15188.

[33] Musaev, M.; Mussakhojayeva, S.; Khujayorov, I.; Khassanov, Y.; Ochilov, M.; Atakan Varol, H. USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments. arXiv 2021, arXiv:2107.14419.

[34] Giannakopoulos, T. Pyaudioanalysis: An open-source python library for audio signal analysis. PLoS ONE 2015, 10, e0144610.

[35] Khamdamov, U.; Mukhiddinov, M.; Akmuradov, B.; Zarmasov, E. A Novel Algorithm of Numbers to Text Conversion for Uzbek Language TTS Synthesizer. In Proceedings of the 2020 International Conference on Information Science and

Communications Technologies (ICISCT), Tashkent, Uzbekistan, 4–6 November 2020; IEEE: New York, NY, USA, 2020.

[36] Makhmudov, F.; Mukhiddinov, M.; Abdusalomov, A.; Avazov, K.; Khamdamov, U.; Cho, Y.I. Improvement of the end-to-end scene text recognition method for "text-to-speech" conversion. Int. J. Wavelets Multiresolution Inf. Process. 2020, 18, 2050052.

[37] Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; pp. 249–256.

[38] Latif, S.; Qadir, J.; Qayyum, A.; Usama, M.; Younis, S. Speech technology for healthcare: Opportunities, challenges, and state of the art. IEEE Rev. Biomed. Eng. 2020, 14, 342–356.

[39] Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Epps, J. Direct modelling of speech emotion from raw speech. In Proceedings of the Intespeeech 2019, Graz, Austria, 15–19 September 2019.

[40] Palaz, D.; Doss, M.M.; Collobert, R. Convolutional neural networks-based continuous speech recognition using raw speech signal. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Queensland, Australia, 19–24 April 2015; pp. 4295–4299.

[41] Muckenhirn, H.; Doss, M.M.; Marcell, S. Towards directly modeling raw speech signal for speaker verification using CNNs. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4884–4888.

[42] Passricha, V.; Aggarwal, R.K. A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition. J. Intell. Syst. 2020, 29, 1261–1274.

[43] Soltau, H.; Kuo, H.K.; Mangu, L.; Saon, G.; Beran, T. Neural network acoustic models for the DARPA RATS program. Interspeech 2013, 2013, 3092–3096.

[44] Mamyrbayev, O.; Turdalyuly, M.; Mekebayev, N.; Alimhan, K.; Kydyrbekova, A.;

Turdalykyzy, T. Automatic recognition of Kazakh speech using deep neural networks. In Asian Conference on Intelligent Information and Database Systems, Yogyakarta, Indonesia, 8–11 April 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 465–474.

[45] Mamyrbayev, O.; Oralbekova, D.; Kydyrbekova, A.; Turdalykyzy, T.; Bekarystankyzy, A. End-to-End Model Based on RNN-T for Kazakh Speech Recognition. In Proceedings of the 2021 3rd International Conference on Computer Communication and the Internet (ICCCI), Nagoya, Japan, 25–27 June 2021; pp. 163–167.

[46] Khassanov, Y.; Mussakhojayeva, S.; Mirzakhmetov, A.; Adiyev, A.; Nurpeiissov, M.; Varol, H.A. A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline. arXiv 2020, Chungsoo Lim Mokpo, Yeon-Woo Lee, and Joon-Hyuk

[47] Chang, "New Techniques for Improving the practicality of a SVM-Based Speech/Music Classifier," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1657-1660, 2012.

[48] Hongchen Jiang, JunmeiBai, Shuwu Zhang, and Bo Xu, "SVM-Based Audio Scene Classification," IEEE International Conference Natural Language Processing and Knowledge Engineering, Wuhan, China, pp. 131-136, October 2005.

[49] Lim and Chang, "Enhancing Support Vector MachineBased Speech/Music Classification using Conditional Maximum a Posteriori Criterion," Signal Processing, IET, vol. 6, no. 4, pp. 335-340, 2012.

[50] Md. Al Mehedi Hasan and Shamim Ahmad. predSuccSite: Lysine Succinylation Sites Prediction in Proteins by using Support Vector Machine and Resolving Data Imbalance Issue. International Journal of Computer Applications 182(15):8-13, September 2018.

[51] Hend Ab. ELLaban, A A Ewees and Elsaeed E AbdElrazek. A Real-Time System for Facial Expression Recognition using Support Vector Machines and kNearest Neighbor Classifier. International Journal of Computer Applications 159(8):23-29,

February 2017.

[52] Quinlan, J.R, “Simplifying decision trees”, International Journal of Man-Machine Studies, vol. 27, no. 3, pp. 221- 234, 1987.

[53] Quinlan, J.R. “Induction of decision trees”, Mach Learn vol. 1, pp. 81–106, 1986, https://doi.org/10.1007/BF00116251.

[54] Quinlan, J. R., & Cameron-Jones, R. M. “FOIL: A midterm report”, In European conference on machine learning pp. 1-20, 1993.

[55] Alkhateeb, F., Baget, J-F, et Euzenat, J., “Extending SPARQL with regular expression patterns (for querying RDF)”, Journal of web semantics, vol. 7, no 2, pp. 57-73, 2009.

[56] LI, J. et WANG, J., “System and method for automatic linguistic indexing of images by a statistical modeling approach”. U.S. Patent no. 7, pp. 394,947, 2008.

[57] Wang, X. H., Liu, A., & Zhang, S. Q. “New facial expressionrecognition based on FSVM and KNN”, OptikInternational Journal for Light and Electron Optics, vol. 126, no. 21, pp. 3132-3134, 2015.

[58] Cherif, W. Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis. Procedia Computer Science, vol. 127, pp. 293-299, 2018.

[59] Cherif, W., Madani, A., & Kissi, M. “A combination of low-level light stemming and support vector machines for the classification of Arabic opinions”, In 2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA), pp. 1-5, 2016. IEEE.

[60] Husain, W., Xin, L. K., Rashid, N. A., and Jothi, N. (2016). “Predicting Generalized Anxiety Disorder among women using random forest approach,” in Proceedings of the 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur

[61] Sun, B., Zhang, Y., He, J., Yu, L., Xu, Q., Li, D., et al. (2017). “A random forest

regression method with selected-text feature for depression assessment," in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC, Cham