Enhanced Marathi Speech Recognition Using Double Delta MFCC and DTW

Rajashri G. Kanke[1], Ramnath M Gaikwad[2], Manasi R. Baheti[3]

[1]Research Student, Department of CS & IT,
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India,
[1]*csit.rgk@bamu.ac.in*

[2]Research Student, Department of CS & IT,
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India,
[2]*ramnath1254@gmail.com*

[3]Assistant Professor, Department of CS & IT,
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India
[3]*mrb.csit@bamu.ac.in*

Abstract:
This paper describes the technique Mel-Frequency Cepstral Coefficients (MFCC), Delta MFCC, and Double Delta MFCC for Extract the Features and DTW for Pattern Matching of Automatic Speech Recognition (ASR) for Marathi. These studies present a speaker-independent Speech Recognition System for Marathi. The Dataset of Created of Speech being natural data and Speech disordered people's speech data in the Marathi language. The dataset of speech samples was created with samples of Marathi Digits and words with and without speech disorder. 'PRAAT' was used for these recordings. Various feature extraction techniques are available, but MFCC is widely used and here, Double Delta MFCC is used to increase the recognition rate along with DTW for Pattern Matching. Speech being natural interaction medium technology can be used to develop small interfaces which can be used for various applications to help interaction between human and computer systems. This research aims to have good interaction between the Human and Computer Systems and, the patient suffering from Speech disorders peoples. That means the "Voice Technology" has been improved in the Marathi Language.

Keywords:Automatic Speech Recognition (ASR), Double Delta Mel-Frequency Cepstral Coefficients (DD MFCC), Dynamic Time Warping (DTW), Small Vocabulary Marathi Speech (SVMS)

I. INTRODUCTION

This study represents Small vocabulary Automatic Speech Recognition for the Marathi language using Mel-Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW) techniques. The database plays important role in recognition, in this work, a sort vocabulary database was required, and so a database was created in the research lab. The experimental work is carried out using Python by using some libraries. Many libraries are used in python for Speech recognition systems. LIBROSA library package is used in this work to read the wave File, Extract Features, Pattern Matching and Plot Diagram. Once the database is available, the next is to apply the feature extraction step; which can be done using various techniques. Here, the MFCC technique is used for Feature Extraction, being the most widely used and having similarity to the human auditory system. MFCC has 12 Cepstral coefficient features with one coefficient energy feature and Delta has extracted the 26 features Double Delta MFCC has 39 Cepstral coefficient features to extract the speech features. [12] After the extraction of features, the pattern-matching technique is to be used. DTW technique is implemented for this. Although the database used is small, and techniques are also widely used, once the recognition is obtained, such systems can be used to develop interfaces for some small, useful applications/platforms to address various requirements, such as hands-free applications, IVRS, specific devices for small vocabulary, teaching aid, etc.

II. METHODOLOGY

Speech files are preprocessed in order to process the feature extraction. Signal modeling and pattern matching are the two main tasks carried out by the voice recognition system. Speech signals are transformed into a set of parameters through a process. Speech when recorded and saved to the system becomes digitized. It is needed to preprocess speech file(s). The technique followed figure by us for the proposed work is shown in Figure no 1. [1]

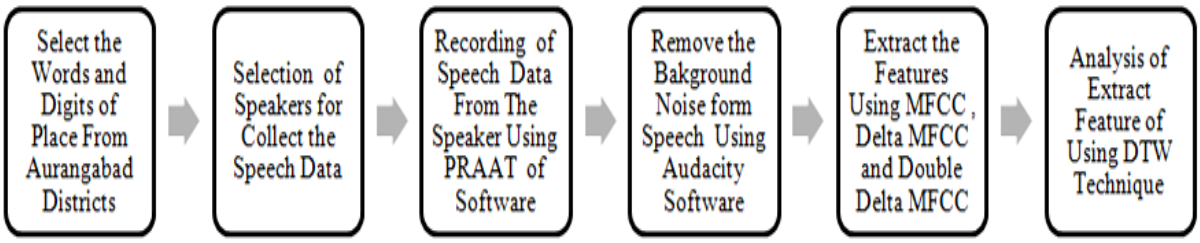


Fig.1: Methodology for Proposed Work [1]

III. IMPLEMENTATION

A. DATABASE USED FOR WORK

The speech data will be collected from Normal Speakers and Speech Disorder Speakers in the Marathi language. The selected speakers will be from Aurangabad District. They would be comfortable with reading & speaking the Marathi language the speaker is classified on the essential gender. We are choosing a small vocabulary of 10 common Digits and 5 Isolated Words in Marathi speech data. For small vocabulary speech recognition, the vocabulary size was kept to a maximum of 10 digits one target language chosen is the Marathi language. While maintaining enough veracity for models developed using the data to possibly be fully utilized for some applications, we want to have a limited vocabulary to ensure that the data-collecting collecting method was lightweight. Additionally, we want the dataset to be accessible in ways that are comparable to those of collections like the 10 frequent numbers and the 5 Words. The recording environment is noisy Environment so very difficult to collect the speech data. Speech data acquisition is the first step to the word-building of a speech recognition system the speech training data is utilized to train the algorithm to affect recognition accuracy. We describe here the measures taken for collected speech data, to develop a small vocabulary Marathi speech recognition system (SVMS) as a part of voice. To achieve high audio quality, the recording was done in the lab so with noisy sound and echo. The sampling frequency for all recordings was set to 16000 Hz. [1] the speech data was collected with the help of microphone 'PRAAT' software using the Mono channel. We are using noise-canceling software 'Audacity' to remove the noise and echo. The small vocabulary data size is a total of 10 speakers and each speaker is to collect one word in ten utterances. The total size of the database is 1500 Audio files. The following table is collected from the Marathi speech database and IPA (International Phonetic Alphabet) format. Table no I and table no II represent the used database of Marathi digits and isolated words.

TABLE I: MARATHI SPEECH DATABASE OF DIGITS

Sr. No	Devnagari Digits	IPA Format	Frequency (Hz)	No. of Utterance
1	एक	/eːkə/	16000	10
2	दोन	/d̪oːnə/	16000	10
3	तीन	/t̪iːnə/	16000	10
4	चार	/cɑːrə/	16000	10
5	पाच	/pɑːcə/	16000	10
6	सहा	/səɦɑː/	16000	10
7	सात	/sɑːt̪ə/	16000	10
8	आठ	/ɑːʈʰə/	16000	10
9	नऊ	/nəuː/	16000	10
10	दहा	/d̪əɦɑː/	16000	10

TABLE II: MARATHI SPEECH DATABASE OF WORDS

Sr. No	Devnagari Words	IPA Format	Frequency (HZ)	No. of Utterance
1	आज	/ɑːʄə/	16000	10
2	मला	/Məlaː/	16000	10
3	तुला	/T̪Ulaː/	16000	10
4	आहे	/ɑːɦeː/	16000	10
5	पाणी	/Pɑːɳiː/	16000	10

IV. EXPERIMENTAL WORK

A. *FEATURES EXTRACTION FOR SPEECH RECOGNITION*

The purpose of feature extraction is to transform voice waveform into a parametric representation. Mel-Frequency Cepstral Coefficients (MFCC), Relative Spectra Perceptual Linear Prediction (RASTA-PLP), Discrete Wavelet Transforms (DWT), Linear Predictive Coding (LPC), and PCA this technique used for feature extraction of speech recognition, most commonly using the MFCC technique so we represent the MFCC and Delta MFCC, and Double Delta MFCC for Feature Extraction Technique.

B. *MFCC TECHNIQUE*

Extraction of MFCC features from audio signals. These works take the input of disorder speech it is the 39 MFCC features parameters are 12 Cepstrum Coefficients plus the energy term. Then we've 2 more sets like the delta and therefore the double delta values. Extract Mel Frequency Cepstral Coefficients from an audio signal. These features are defined from the 13 MFCC features vectors, which are extracted from the ".wav" file of the input speech signal. After adding energy and then delta and double delta features to the 12 Cepstral features, 39 MFCC features are derived. 12 Cepstral coefficients, 12 Deltas Cepstral Coefficients, 12 double delta Cepstral coefficients, one MFCC energy, one Delta energy, and one Double Delta Coefficient Energy. Extraction of MFCC features from audio signals. These works take the peak loudness across all channels and will be used to calculate the MFCC. The outcome can be different from the individual MFCC calculation for each channel. The Diagram shows the flow of the MFCC Technique.

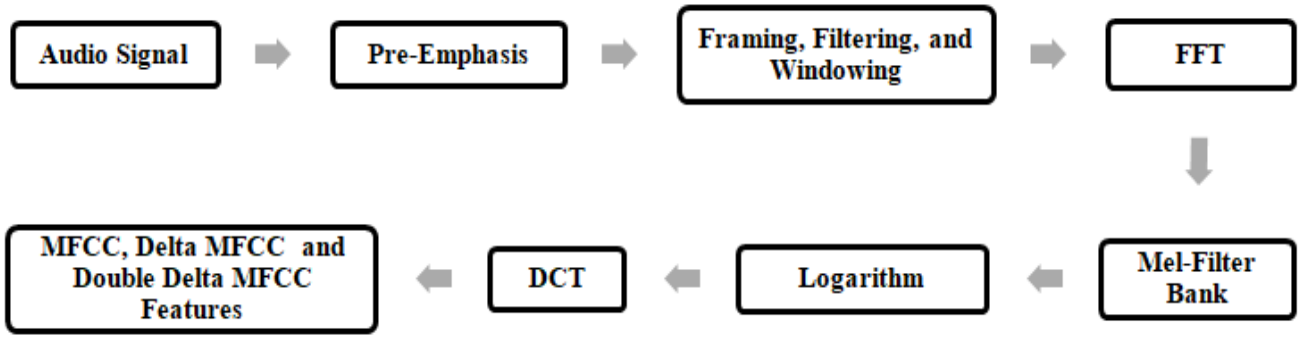


Fig. 2: Block diagram of MFCC computation
(https://www.google.com/search?q=BLOCK+DIAGRAM+OF+MFCC+IN+PYTHON)

1) *Pre-Emphasis:* The Pre-Emphasis step isolated digits data case is moving over purify which emphasizes higher frequencies. It will increase the ability to the signal at a higher frequency. A form of filter called pre-emphasis filtering kept high frequencies in a spectrum. This procedure's goal was to make the input sound less noisy so that sound extraction would be more accurate. [12]

2) *Framing:* Signal. In this frame, the length should not be too short or too long. If the frame length is short we not getting enough bits and for the long length the signal changes. [1] The inconsistency of the sound effect from the vocal production organs made processing signals in a short segment became necessary. The typical frame duration was between 10 and 30 milliseconds. This framing process was carried out continuously until all signals could be processed and generally done overlapping for each frame. [12]
3) *Windowing:* Windowing of an individual is done to eliminate the discontinuities t the edges o the frames. It is used to reduce the spectral effects, and smooth the signal for conditional of the Fast Fourier Transmission (FFT). The frame-blocking procedure had the effect of discontinuing the signal. The signal became discontinuous as a result of the frame-blocking operation. A windowing technique was required in order to lessen the discontinuous impact on the signal caused by the frame-blocking operation. By increasing every nth frame by the value of the nth window, depending on the kind of window being utilized, this windowing method lessened the impact of the frame-blocking process. [12]
4) *Mel-Filter Bank:* The frequency range in the FFT spectrum is extensive, and the voice signal does not follow the linear scale.
5) *DCT*: DCT converts the log Mel spectrums from frequency back to time demine. Transform with DCT is required because FFT has been performed. DCT was the final step in the basic process of MFCC feature extraction. The fundamental idea behind DCT was to embellish the Mel spectrum to create an accurate picture of the local spectral characteristics. To provide an accurate representation of the sound spectral, DCT was used to determine the spectrum. . The result was called the Mel-Frequency Cepstrum Coefficient (MFCC). [13]

C. READ WAVE FILE

The following figure no 3 shows the wave file. The graphical representation of fig 3.1 shows the normal person audio file and fig. 3.2 is showing the Speech disorder person audio file. The time is shown on the X-axis, while the amplitude is shown on the Y-axis of the audio signal digit 'one' (EK) using the Python 3.10.1 version interface.

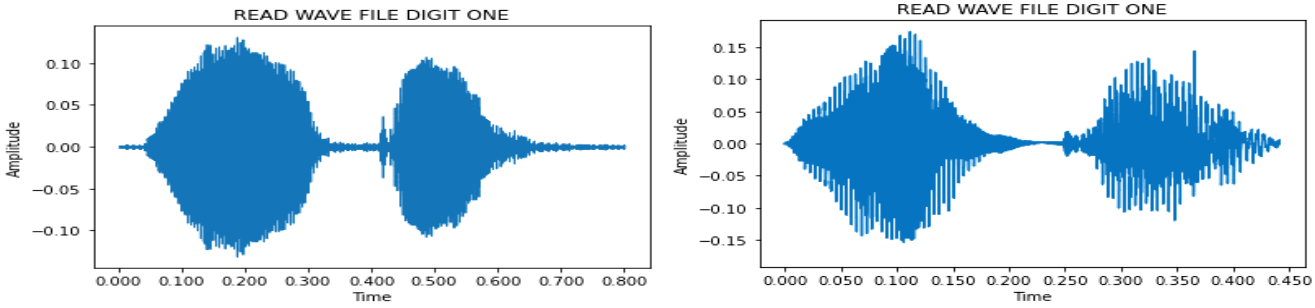


Fig. 3.1:(A) Speech Being Natural

Fig. 3.2: (B) Speech Disorder Person

Fig. 3: Read Wave File of Digit One (EK) in Fig.3.1 and Fig. 3.2

D. MFCC FOR FEATURES

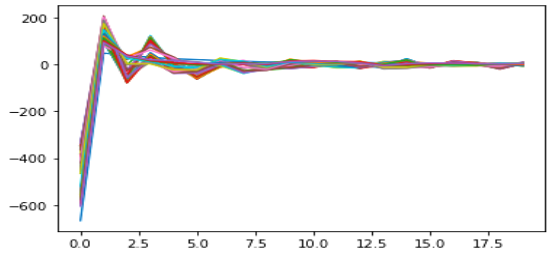
The below table shows the MFCC technique applied to the wave file to extract 13 Cepstrum coefficient features using the LIBROSA package. The following Table no III shows the normal person MFCC features and Table no VI shows the Speech disorder person MFCC features such represent the 10 frames and MFCC features from the given word 'EK'. Additionally, we compute these characteristics' means, medians, and standard deviations. Figure no 4 shows more understanding of the graphical representation of the MFCC features and frames. Every frame of Delta MFCC is represented by a different color in figure no 6. [14] The x-axis is used for each of the MFC Coefficients (from 0 to 13 in this Figure). The y-axis is used for the values of the coefficients (ranging approx. from -600 to 200 in this figure no 5).

TABLE III: MFCC FEATURES WITH 10 FRAME OF NORMAL PERSON FOR DIGITS ONE (EK)

Features/Coefficient	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6	Frame 7	Frame 8	Frame 9	Frame 10
1	8.58996	-16.3123	15.9674	0.971308	13.309	9.3186	17.9343	-0.76546	10.4793	10.1435
2	8.59727	-17.6965	14.5694	4.63684	18.4173	9.48403	21.9203	6.12514	4.78996	4.46653
3	10.2122	0.81918	-0.49578	6.5045	-6.71317	-25.098	9.05064	-7.33703	-4.60293	-12.0686
4	13.1112	-0.71193	-7.78367	27.8805	-34.174	-50.4714	-1.06578	-10.242	-10.5399	-11.178
5	14.1676	0.56538	-7.08866	34.0206	-42.6873	-53.6773	-3.21753	-9.14183	-9.50688	-13.0532
6	14.7932	2.27808	-4.33718	36.833	-39.4988	-53.4421	-1.94362	-9.89292	-9.97664	-18.5427
7	15.5647	2.38319	-3.27176	37.7013	-42.4073	-46.1441	-0.24709	-15.6473	-11.4524	-15.0112
8	16.1299	5.42285	-6.28039	38.7154	-37.3185	-49.1709	0.000919	-17.8232	-4.34494	-20.628
9	16.7671	4.9456	-8.10744	39.277	-33.4478	-52.3782	-3.66273	-10.3427	-9.99502	-16.3988
10	17.2342	5.6528	-10.8829	35.3419	-29.7854	-55.2037	1.17121	-11.2894	-15.4251	-10.6287
11	17.555	2.80836	-10.0373	36.3181	-35.3727	-47.9981	-2.83904	-11.3153	-15.3184	-10.1304
12	17.8338	3.61433	-16.1013	37.2198	-36.9096	-43.23	-3.14819	-18.3234	-10.2682	-12.2424
13	18.3195	1.5141	-16.448	42.1269	-42.3877	-42.2543	-1.85541	-19.3287	-7.19019	-14.8543
Mean	-5.26567									
Median	-4.34106									
Standard Deviation	21.88655									

TABLE IV: MFCC FEATURES WITH FRAME OF SPEECH DISORDER PERSON DIGITS ONE (EK)

Features/Coefficient	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6	Frame 7	Frame 8	Frame 9	Frame 10
1	8.03639	8.11253	9.76278	11.0826	14.1354	13.1307	9.70733	8.67508	7.94389	8.02064
2	7.26704	9.02701	11.181	10.9588	14.9079	13.9107	10.8017	9.95684	9.70622	10.2632
3	10.5851	11.3172	12.3292	12.0489	15.3329	14.462	10.8035	9.46987	9.31972	11.0496
4	10.6178	10.7736	13.2508	12.3254	15.6961	15.011	11.2862	10.4183	10.22	12.3139
5	11.0765	10.4261	13.5461	12.9759	15.8949	15.321	11.6597	10.8796	10.8628	12.9706
6	8.3901	10.5599	13.1616	12.5757	15.9501	15.6419	12.4761	11.5672	11.0607	12.7938
7	9.47058	11.2516	12.7838	12.4245	16.1224	15.9375	12.6485	11.6636	11.122	12.1454
8	8.41341	10.1042	13.0821	12.9834	16.4893	16.3505	12.8096	11.9471	11.7776	12.4213
9	9.90587	9.88652	13.3253	13.3226	16.8421	16.6706	12.8874	12.2212	11.5356	12.3671
10	11.5821	12.0809	13.323	13.5031	17.0067	16.8232	13.1482	12.1664	11.5465	12.3229
11	10.8826	11.7497	12.7508	13.2589	16.9559	16.8085	13.0919	12.293	11.5232	12.1596
12	10.1632	10.1826	13.4487	13.1626	16.6623	16.5498	12.6892	11.8167	11.4409	11.5398
13	10.5032	11.7942	13.7392	13.0269	16.1028	16.0429	12.4249	11.1593	10.4212	10.6237
Mean	12.264562									
Median	12.1525									
Standard Deviation	2.2602386									



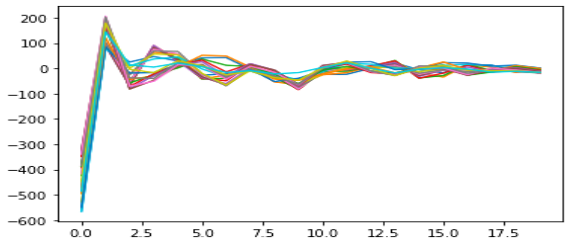


Fig.4.1: (A) Normal Person Fig.4.1: (B) Speech Disorder People

Fig 4: MFCC Features with Frames of Digits One (EK) in Wave Format in Fig. 4.1 and Fig. 4.2

E. DELTA MFCC RESULT

The below table shows the Delta MFCC technique applied to the wave file to extract 26 Cepstrum coefficient features. The following Table no V shows the normal person DeltaMFCC features and Table no VI shows the Speech disorder person Delta MFCC features such as representing the 10 frames and MFCC features from the given word 'EK'. Figure no 5 shows more understanding of the graphical representation of the Delta MFCC features and frames. Every frame of Delta MFCC is represented by a different color in figure no 6. [14] The x-axis is used for each of the MFC coefficients (from 0 to 26 in this Figure). The y-axis is used for the values of the coefficients (ranging approx. from -1 to 45 in this figure no 5).

26 Delta MFCC Formula: Δ k =fk− fk− 1 (One is Energy Coefficient) [9]

TABLE V: DELTA MFCC FEATURES WITH FRAMES OF NORMAL PERSON DIGITS ONE (EK)

Features/Coefficient	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6	Frame 7	Frame 8	Frame 9	Frame 10
1	0.261975	0.439781	-0.799966	0.969755	-1.56554	-1.58128	-0.605259	-0.216514	-0.448528	-0.503359
2	0.263816	0.443556	-0.804156	0.987699	-1.58153	-1.57897	-0.612183	-0.177893	-0.447974	-0.488469
3	0.263902	0.443065	-0.822865	0.978998	-1.58382	-1.55953	-0.630818	-0.166005	-0.465114	-0.472456
4	0.261464	0.43871	-0.818358	0.977918	-1.56854	-1.54472	-0.646289	-0.1501	-0.476275	-0.460086
5	0.25521	0.438851	-0.816287	0.974709	-1.54123	-1.50879	-0.661928	-0.136502	-0.506658	-0.46347
6	0.24519	0.434906	-0.802085	0.962746	-1.48261	-1.47605	-0.670076	-0.134759	-0.529698	-0.463289
7	0.231918	0.42807	-0.755871	0.955059	-1.41815	-1.46897	-0.669256	-0.0988709	-0.50989	-0.469433
8	0.216869	0.420688	-0.714103	0.915715	-1.35315	-1.43371	-0.65585	-0.0794163	-0.490212	-0.456073
9	0.200668	0.406515	-0.65841	0.881409	-1.28178	-1.39529	-0.618989	-0.0600029	-0.478801	-0.428904
10	0.183962	0.384027	-0.607627	0.842881	-1.19759	-1.35518	-0.570502	-0.0233139	-0.459308	-0.406715
11	0.166576	0.357329	-0.553875	0.800798	-1.10673	-1.30719	-0.518831	-0.00186087	-0.468201	-0.401394
12	0.148417	0.33278	-0.487695	0.754778	-1.01546	-1.25432	-0.477137	0.0141643	-0.467542	-0.371072
13	0.130362	0.30929	-0.420879	0.704151	-0.922165	-1.21018	-0.447589	0.0230378	-0.466809	-0.320203
Mean	-0.3573293									
Median	-0.4633795									
Standard Deviation	0.71187									

TABLE VI: DELTA MFCC FEATURES WITH FRAMES OF SPEECH DISORDER PERSON DIGITS ONE (EK)

Features/Coefficient	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6	Frame 7	Frame 8	Frame 9	Frame 10
1	-0.0012381	0.009781	0.0030072	-0.00578327	0.004507	-0.00298776	0.0145048	-0.0180116	0.0082417	0.00197978
2	-0.0012724	0.010019	0.0032549	-0.00525211	0.0044341	-0.00392545	0.0146802	-0.018375	0.0088549	0.0022939
3	-0.0013027	0.010244	0.003498	-0.00468233	0.0043544	-0.00493326	0.0147589	-0.0186601	0.0094407	0.0025811
4	-0.0013307	0.010457	0.0037567	-0.00406181	0.0042625	-0.00598104	0.0147577	-0.0188684	0.0100422	0.00284161
5	-0.001357	0.010654	0.004007	-0.00344937	0.0041679	-0.00700207	0.0147413	-0.0190538	0.0106233	0.00309408
6	-0.0013815	0.010836	0.0042487	-0.00284503	0.0040706	-0.00799634	0.0147098	-0.0192162	0.011184	0.00333851
7	-0.0014043	0.011002	0.0044819	-0.00224877	0.0039705	-0.00896385	0.0146632	-0.0193557	0.0117243	0.00357489
8	-0.0014254	0.011153	0.0047066	-0.00166061	0.0038676	-0.00990461	0.0146014	-0.0194723	0.0122443	0.00380323
9	-0.0014447	0.011289	0.0049229	-0.00108053	0.0037621	-0.0108186	0.0145245	-0.0195659	0.0127438	0.00402352
10	-0.0014623	0.011409	0.0051306	-0.00050855	0.0036537	-0.0117059	0.0144325	-0.0196366	0.013223	0.00423577
11	-0.0014781	0.011514	0.0053299	5.53E-05	0.0035426	-0.0125664	0.0143253	-0.0196843	0.0136818	0.00443997
12	-0.0014922	0.011603	0.0055206	0.00061115	0.0034288	-0.0134001	0.014203	-0.0197091	0.0141203	0.00463614
13	-0.0015046	0.011677	0.0057029	0.00115886	0.0033122	-0.0142071	0.0140655	-0.019711	0.0145383	0.00482425
Mean	0.0017253									
Median	0.0036143									
Standard Deviation	0.0098259									

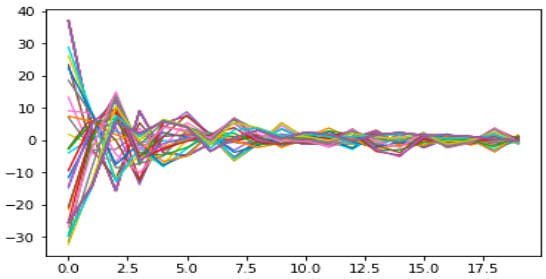


Fig.5.1: (A) Normal Person

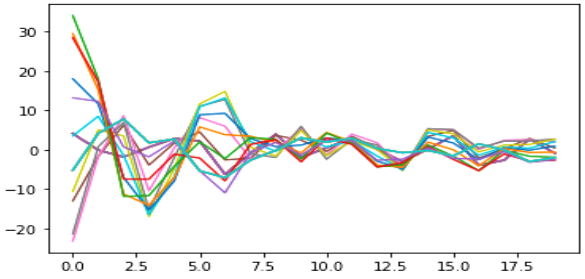


Fig.5.1: (B) Speech Disorder People

Fig. 5: Delta MFCC Features with Frames of Digits One (EK) in Wave Format in Fig. 5.1 and Fig. 5.2

F. DOUBLE DELTA MFCC RESULT

The below table shows the Double Delta MFCC technique applied to the Audio file to extract 39 Cepstrum Coefficient features. The following Table no VII shows the normal person Delta MFCC features and Table no VIII shows the Speech disorder person Delta MFCC features such as representing the 10 frames and MFCC features from the given word 'EK'. Figure no 6 shows more and more understanding of the graphical representation of the Double Delta MFCC features and frames. Every frame of Double Delta MFCC is represented by a different color in figure no 6. [14] The x-axis is used for each of the MFC coefficients (from 0 to 39 in this Figure). The y-axis is used for the values of the coefficients (ranging approx from -1 to 25 in this figure no 6).

39 Double Delta MFCC Formula: $\Delta\Delta k = fk - fk-1$ (one is Energy Coefficient) [9]

TABLE VII: DOUBLE DELTA MFCC FEATURES WITH FRAMES OF NORMAL PERSON DIGITS ONE (EK)

Features/Coefficient	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6	Frame 7	Frame 8	Frame 9	Frame 10
1	-0.0047264	-0.002993	0.0085682	-0.027621	0.0254553	0.0372717	0.00853897	6.40E-05	0.0116304	0.0139464
2	-0.0048747	-0.002946	0.0089154	-0.02864	0.0260861	0.0384498	0.00922959	-0.000234493	0.0119459	0.0141053
3	-0.0050191	-0.00291	0.0092773	-0.0295987	0.0267043	0.0395474	0.00991968	-0.000528513	0.0122462	0.0142252
4	-0.0051592	-0.00289	0.0096549	-0.0304916	0.0272993	0.0405485	0.0106084	-0.000819809	0.0125464	0.0143097
5	-0.0052945	-0.002885	0.0100457	-0.0313165	0.0278756	0.0414615	0.0112901	-0.00110564	0.0128503	0.0143646
6	-0.0054248	-0.002892	0.010461	-0.0320593	0.0284435	0.0422845	0.011962	-0.00136892	0.0131648	0.0143938
7	-0.0055492	-0.002924	0.0108871	-0.0327424	0.0289924	0.0430202	0.0126159	-0.00159074	0.0134846	0.0143965
8	-0.0056669	-0.002979	0.0113236	-0.0333543	0.0295159	0.0436664	0.0132553	-0.00175839	0.0138064	0.0143766
9	-0.0057772	-0.003056	0.0117551	-0.0339092	0.0300143	0.0442291	0.0138857	-0.00187688	0.0141388	0.0143528
10	-0.0058799	-0.003156	0.0122	-0.0344034	0.030487	0.0446912	0.0145172	-0.00193804	0.014472	0.0143229
11	-0.0059748	-0.003275	0.0126638	-0.0348326	0.0309268	0.045044	0.0151509	-0.00195611	0.0147921	0.0142956
12	-0.006062	-0.003415	0.0131339	-0.0351973	0.0313307	0.0452874	0.0157449	-0.00193751	0.0150965	0.0142618
13	-0.0061413	-0.003572	0.0136076	-0.0354726	0.0317052	0.0454391	0.0162958	-0.00187703	0.0154039	0.0142125
Mean	0.008034									
Median	0.011477									
Standard Deviation	0.0194564									

TABLE VIII: DOUBLE DELTA MFCC FEATURES WITH FRAMES OF SPEECH DISORDER PERSON DIGITS ONE (EK)

Features/Coefficient	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6	Frame 7	Frame 8	Frame 9	Frame 10
1	-0.0012381	0.009781	0.0030072	-0.00578327	0.004507	-0.00298776	0.0145048	-0.0180116	0.0082417	0.00197978
2	-0.0012724	0.010019	0.0032549	-0.00525211	0.0044341	-0.00392545	0.0146802	-0.018375	0.0088549	0.0022939
3	-0.0013027	0.010244	0.003498	-0.00468233	0.0043544	-0.00493326	0.0147589	-0.0186601	0.0094407	0.0025811
4	-0.0013307	0.010457	0.0037567	-0.00406181	0.0042625	-0.00598104	0.0147577	-0.0188684	0.0100422	0.00284161
5	-0.001357	0.010654	0.004007	-0.00344937	0.0041679	-0.00700207	0.0147413	-0.0190538	0.0106233	0.00309408
6	-0.0013815	0.010836	0.0042487	-0.00284503	0.0040706	-0.00799634	0.0147098	-0.0192162	0.011184	0.00333851
7	-0.0014043	0.011002	0.0044819	-0.00224877	0.0039705	-0.00896385	0.0146632	-0.0193557	0.0117243	0.00357489
8	-0.0014254	0.011153	0.0047066	-0.00166061	0.0038676	-0.00990461	0.0146014	-0.0194723	0.0122443	0.00380323
9	-0.0014447	0.011289	0.0049229	-0.00108053	0.0037621	-0.0108186	0.0145245	-0.0195659	0.0127438	0.00402352
10	-0.0014623	0.011409	0.0051306	-0.00050855	0.0036537	-0.0117059	0.0144325	-0.0196366	0.013223	0.00423577
11	-0.0014781	0.011514	0.0053299	5.53E-05	0.0035426	-0.0125664	0.0143253	-0.0196843	0.0136818	0.00443997
12	-0.0014922	0.011603	0.0055206	0.00061115	0.0034288	-0.0134001	0.014203	-0.0197091	0.0141203	0.00463614
13	-0.0015046	0.011677	0.0057029	0.00115886	0.0033122	-0.0142071	0.0140655	-0.019711	0.0145383	0.00482425
Mean	0.0017253									
Median	0.0036143									
Standard Deviation	0.0098259									

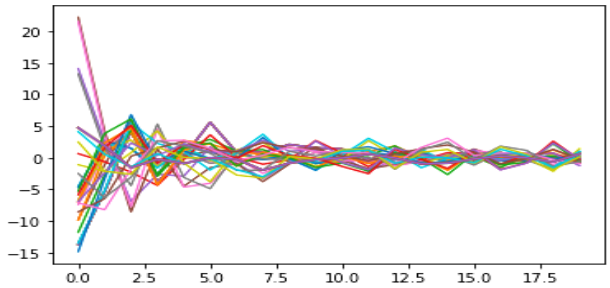


Fig.6.1:(A) Normal Person

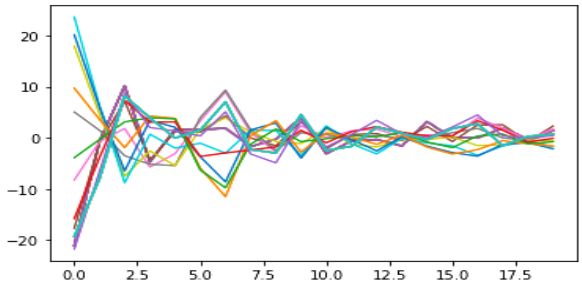


Fig.6.2: (B) Speech Disorder People

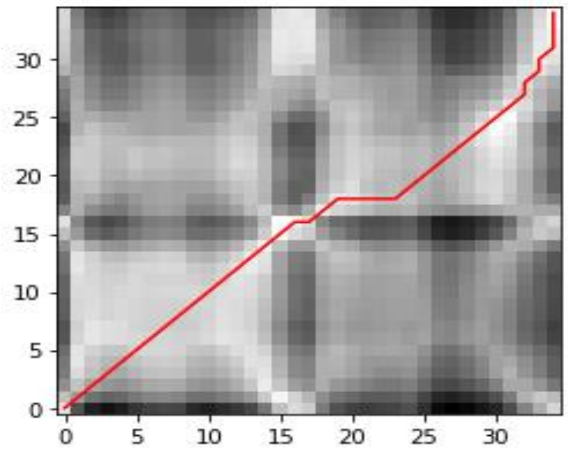
Fig. 6: Double Delta MFCC Features with Frames of Digits One (EK) in Wave Format in Fig. 6.1 and Fig. 6.2

G. PATTERN MATCHING USING DTW TECHNIQUE

Once feature vectors are generated using MFCC, the next step is to find the optimal match. The simplest way to recognize sentence samples is to compare them to a number of stored templates and determine the best match. DTW is a technique for evaluating how closely two patterns match up across time zones. The DTW is used for normal people's speech data and Speech disordered speech data these two audio waves file DTW is an instance of the general class of algorithms known as dynamic programming. With the help of the Dynamic Time Warping method, it is possible to align the reference and test patterns and determine the average distance between the two sequences A and B. The below table no XI represents the DTW table of the digit one (EK). The difference must be calculated between frames in the time domain, not between samples in a "single feature domain. This can be more understood by the following figure no 7.1 representing the normal person's speech data match, we find that the Normalized Distance 12376.31 is two normal person wave files. And figure no 7.2 represent the normal person and Speech disordered person this two-person Marathi digit one (EK) audio data shows how to match the features in the wave file and we find that the Normalized Distance is 15176.01. In this figure red line represent the proper alignment of the match. Every human has different sound characteristics. A unique algorithm called dynamic time warping (DTW) is required to determine whether a sound is compatible. Also the classification by calculating the distance between 12D vectors. It is a Dynamic Time Warping, so the difference must be calculated between frames in the time domain, not between samples in a "Single Feature Domain".

TABLE IX: DTW FEATURES MATCHING TABLE OF SAME WORD IS DIGIT ONE (EK)

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0	447.76	1135.04	1897.32	2665.11	3471.9	4358.1	5253.83	6115.64	6977.24	7827.13	8672.31	9474.88
1	447.76	0	247.11	602.486	974.566	1398.1	1911.41	2441.76	2948.13	3446.33	3942.76	4423.4	4874.26
2	1135.04	247.11	0	133.904	320.873	586.566	924.145	1280.03	1627.71	1967.82	2292.63	2593.63	2866.32
3	1897.32	602.486	133.904	0	88.6881	245.865	476.223	728.063	970.971	1213.25	1445.46	1660.26	1854.94
4	2665.11	974.566	320.873	88.6881	0	92.3651	269.443	462.372	652.992	852.44	1050.79	1234.91	1397.55
5	3471.9	1398.1	586.566	245.865	92.3651	0	101.489	219.021	336.43	470.667	604.929	726.714	846.154
6	4358.1	1911.41	924.145	476.223	269.443	101.489	0	35.2643	100.801	193.199	314.599	451.64	629.856
7	5253.83	2441.76	1280.03	728.063	462.372	219.021	35.2643	0	57.5758	148.418	271.02	412.41	590.855
8	6115.64	2948.13	1627.71	970.971	652.992	336.43	100.801	57.5758	0	37.3422	121.904	228.028	379.611
9	6977.24	3446.33	1967.82	1213.25	852.44	470.667	193.199	148.418	37.3422	0	54.839	141.816	281.917
10	7827.13	3942.76	2292.63	1445.46	1050.79	604.929	314.599	271.02	121.904	54.839	0	43.9193	153.295
11	8672.31	4423.4	2593.63	1660.26	1234.91	726.714	451.64	412.41	228.028	141.816	43.9193	0	71.9797
12	9474.88	4874.26	2866.32	1854.94	1397.55	846.154	629.856	590.855	379.611	281.917	153.295	71.9797	0
13	10265.7	5316.87	3113.8	2009.58	1523.54	1004.87	866.115	838.119	613.62	509.092	363.026	247.985	109.428
14	10931.2	5639.5	3295.73	2193.85	1729.37	1289.91	1234.28	1221.85	985.664	877.023	719.023	580.517	385.745
15	11352.8	5772.27	3600.65	2580.48	2111.72	1735.09	1772.99	1776.89	1512.28	1402.09	1244.32	1091.22	856.321
16	11650.1	5968.54	4020.9	3084.28	2619.41	2282.67	2367.96	2425.8	2135.14	2011.37	1844.34	1681.78	1397.98
17	12228.6	6496.91	4422.42	3481.83	3070.51	2787.52	2840.44	2941.95	2683.61	2556.11	2364.73	2189.46	1866.57
18	12957.1	7128.05	4925.99	3925.75	3517.57	3235.34	3287.4	3357.4	3175.76	3031.35	2819.39	2629.84	2264.62
19	13751.2	7766.42	5440.33	4356.35	3939.17	3646.92	3642.15	3715.25	3582.73	3423.99	3199.39	3086.9	2627.7



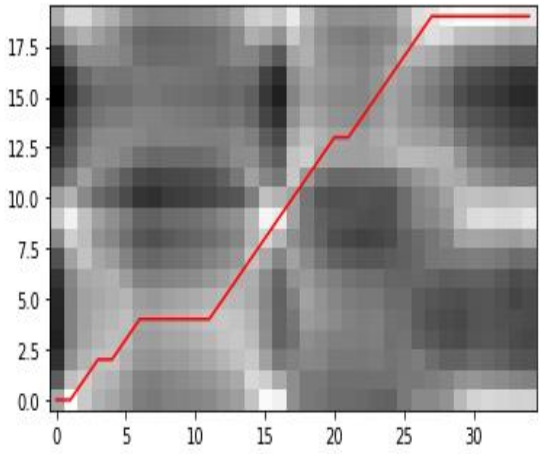


Fig. 7.1: (A) Two Normal Persons (Normalized Distance: 12376.31)

Fig. 7.2: (B) One Normal and Speech Disorder (Normalized Distance: 15176.01)

Fig. 7: DTW Technique for Pattern Matching of Digit One (EK) in Fig. 7.1 and Fig. 7.2

H. PATTERN MATCHING REPRESENTATION IN WAVE FILE

An audio signal creates a time series. Two audio signals, or two-time series, of two separate persons, are seen above. Time series have been matched using the conventional technique of Euclidian matching. The two don't have the same times. [10] Any two-time series can be compared one-to-one on the time axis using the Euclidean distance or another comparable distance. The first time series amplitude at a time (T) will be contrasted with the second time series' amplitude at the same time. Even if the two-time series have extremely similar shapes but are out of phase with respect to time, the comparison and similarity score will suffer as a result. [11] Following figure no 8. Represent how to accurately match the features of two normal persons and figure no 9. Represent the normal person and Speech disordered person's features. Red lines represent the match of each word frame and sound in figures.

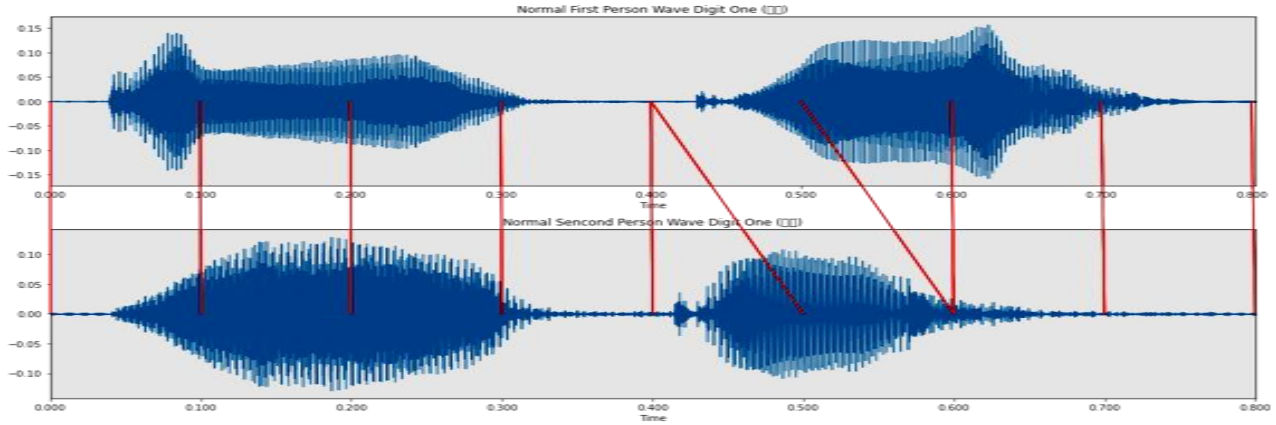


Fig. 8: Match the Features of Two Normal Persons Speech of Digit One (EK)

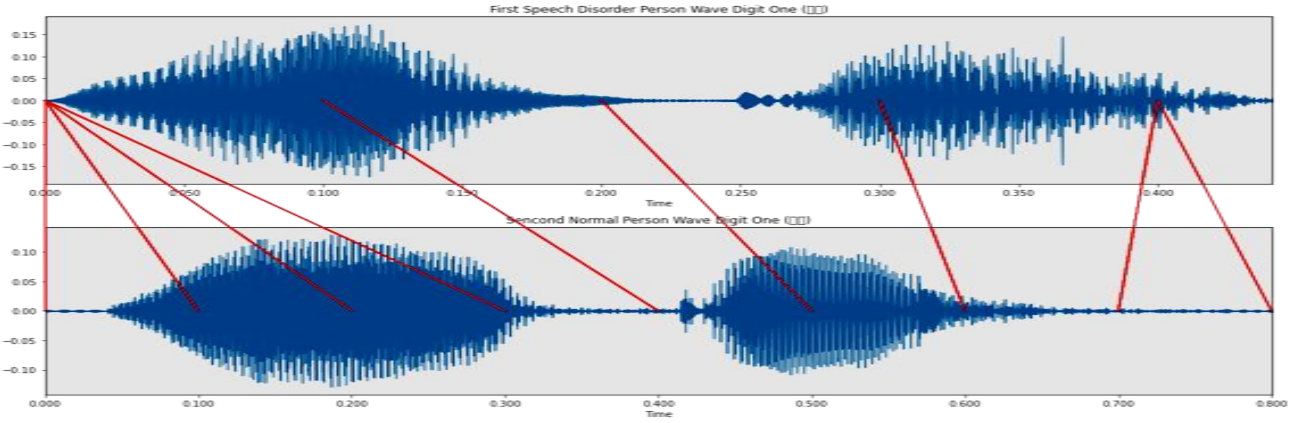


Fig. 9: Match the Features of First Wave is a speech disorder Person and second wave of Speech Normal Person of Digit One (EK)

V. FUTURE SCOPE

To develop Marathi speech recognition is carried out for small vocabulary in Marathi. The same system can be developed with a large vocabulary. This system can be used to develop many applications like small, useful applications/platforms to address various requirements, such as hands-free applications, IVRS, specific devices for small vocabulary, teaching aid, etc. Also, technique-wise, more than one Feature Extraction technique can be applied to enhance the recognition rate.

VI. CONCLUSION

In this work, we presented a Double Delta MFCC for Feature Extraction and DTW for Pattern matching for the Marathi using Automatic Speech Recognition. The speech recognition system is based on Python version 3.10.3 using LIBROSA Package techniques. [7] This paper aims to which technique to better the recognition rate for SVMS recognition and also can recognize a large dataset of the Marathi language. The choice of feature extraction techniques is an essential step in the speech recognition process since the more wisely we choose the extraction technique, the more accurate results we get. This comparison revealed that each extraction method has reliability and performance issues. Also, the results showed the importance of applying mixture feature extraction techniques since each of the presented extraction techniques complements the work. The main objective of best accuracy for the feature extraction method is according to the criteria that matter most for Marathi speech recognition. In the application, using speech-based services will be used to find the recognition rates of the computer. A small vocabulary system in Marathi is developed using simple but effective techniques like MFCC and DTW. As this work has applied delta and double delta MFCC features, the recognition rate is increased, which will help in pattern matching. The dataset considered for this work is small but has samples of daily spoken words in communication, which will help to develop interfaces for many uses.

ACKNOWLEDGEMENT

I would like to acknowledge the help and support of every people for speech data collection and guidance for my research work. Our sincere gratitude to my research mentor Dr. Dr. Manasi R. Baheti of Dr. B. A. M. University for their support and suggestions. Also thanks to the Ministry of Tribal Affairs for giving me this opportunity, to select for the National Fellowship of ST (NFST 2019-2020) to provide me finical needs and Support. And also thank my Department of CS & IT for providing me lab facility.

REFERENCES

[1] Maheshwari A. Ambewadikar, Manasi R. Baheti, "Automatic Speech Recognition system in Marathi for Cerebral Palsy Disabled", CSI Journal of Computing, Vol. No. 3, No. 3, (2020)

[2] Jurhini hoeliltr, Nelqon Morgan. Hynek Hermanskyt,t, H Guenter Hirsch, Grace Tmg, INTEGRATING RASTA-PLP INTO SPEECH RECOGNITION, 0-7803-1775-0/94 $3.00 0, IEEE (1994)

[3] Maria Labied, Abdessamad Belangour, "Automatic Speech Recognition Features Extraction Techniques: A Multi-criteria Comparison", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 8, (2021)

[4] P. Prithvi1, Dr. T. Kishore Kumar2, "Comparative Analysis of MFCC, LFCC, RASTA –PLP", National Institute of Technology, Warangal, Telangana – 506 004, India www.ijser.in ISSN (Online): 2347-3878, Impact Factor: 3.791 Volume.(2015)

[5] Abdulloh Salahul Haq, abdullohsh@student.telkomuniversity.ac.id, Muhammad Nasrun, abdullohsh@student.telkomuniversity.ac.id, Casi Setianingsih, setiacasie@telkomuniversity.ac.id, Muhammad Ary Murti, arymurti@telkomuniversity.ac.id, School of Electrical Engineering, Telkom University Bandung, Indonesia, "Speech Recognition Implementation using MFCC and DTW Algorithm for Home Automation", Article in Proceeding of the Electrical Engineering Computer Science and Informatics, DOI: 10.11591/eecsi.v7.2041 (OCT- 2020)

[6] Leena R Mehta 1, S.P.Mahajan 2, Amol S Dabhade 3,ISSN (Print) : 2320, 3765ISSN (Online): 2278 8875,International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering,Vol. 2, Issue 6, June 2013,Copyright to IJAREEIE, www.ijareeie.com 2133, "COMPARATIVE STUDY OF MFCC ANDLPC FOR MARATHI ISOLATED WORDRECOGNITION SYSTEM", (June- 2013)

[7] D Anggraeni1,2 , W S M Sanjaya1,2, M Y S Nurasyidiek1,2 and M Munawwaroh1,2, 1Department of Physics, Faculty of Science and Technology, Universitas Islam Negeri Sunan Gunung Djati Bandung, Indonesia 2Bolabot Techno Robotic Institute, CV. Sanjaya Star Group, Bandung, Indonesia, 'The Implementation of Speech Recognition using Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machine (SVM) method based on Python to Control Robot Arm' IOP Conf. 012042 doi:10.1088/1757-899X/288/1/012042, Series: Materials Science and Engineering 288 (2017)

[8] Kanke, Rajashri G., Maheshwari A. Ambewadikar, and Manasi R. Baheti. "REVIEW ON SMALL VOCABULARY AUTOMATIC SPEECH RECOGNITION SYSTEM (ASR) FOR MARATHI."(2021)

[9] Tiwari, Sonal A., Rajashri G. Kanke, and A. Maheshwari. "Marathi Speech Database Standardization: A Review and Work." International Journal of Computer Science and Information Security (IJCSIS) 19, no. 7 (2021)

[10] https://betterprogramming.pub/how-to-do-speech-recognition-with-a-dynamic-time-warping-algorithm-159c2a1bb83c.

[11] https://www.theaidream.com/post/dynamic-time-warping-dtw-algorithm-in-time-series.

[12] I D G Y A Wibawa1, and I D M B A Darmawan1 "Implementation of audio recognition using mel frequency cepstrum coefficient and dynamic time warping in wirama praharsini", ICW-HDDA-X 2020, Journal of Physics: Conference Series 1722, IOP Publishing, doi:10.1088/1742-6596/1722/1/01/2014, (2021)

[13] Setiawan A, Hidayatno A and Isnanto R. R. 2011 Aplikasi Pengenalan Ucapan dengan Ekstraksi Mel-Frequency Cepstrum Coefficients (MFCC) Melalui Jaringan Syaraf Tiruan (JST) Learning Vector Quantization (LVQ) untuk Mengoperasikan Kursor Komputer Apl. Pengenalan Ucapan dengan Ekstraksi Mel-Frequency Cepstrum Coefficients Melalui Jar. Syaraf Tiruan Learn. Vector Quantization untuk Mengoperasikan Kursor Komput. 13 82–6.

[14] Rajeev Ranjan, Abhishek Thakur, "Analysis of Feature Extraction Techniques for Speech Recognition System", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-7C2, (May 2019)

[15] https://www.google.com/search?client=firefox-b-d&q=librosa+-feature+extraction.

[16] https://wiki.aalto.fi/display/ITSP/Deltas+and+Delta-deltas.