Automatic Speaker Identification System for Urdu Speech

Fatima Yousaf[1,2], Muhammad Javaid Iqbal[2], Hassan Raza[2], Agha Ali Raza[1], Muhammad Shahid Iqbal[2]
[1]Department of Computer Science, Information Technology University, Lahore, Punjab 54000, Pakistan
[2]Computer Systems Engineering Department, Mirpur University of Science and Technology, Mirpur 10250, Pakistan

***Abstract*- Speaker recognition is the process of recognizing a speaker from a verbal phrase. Such systems generally operate in two ways: to identify a speaker or to verify speaker's claimed identity. Availability of valuable research material witnessed efforts paid to Automatic Speaker Identification (ASI) in East Asian, English and European languages. But unfortunately languages of South Asia especially "Urdu" have got very less attention. This paper aims to describe a new feature set for ASI in Urdu speech, achieving improved performance than baseline systems. Classifiers like Neural Net, Naïve Bayes and K nearest neighbor (K-NN) have been used for modeling. Results are provided on the dataset of 40 speakers with 82% correct identification. Lastly, improvement in system performance is also reported by changing number of recordings per speaker.**

Keywords: MFCC, K-NN, Formants, Automatic Speaker Identification, Urdu.

I. INTRODUCTION

Automatic Speaker Identification (ASI) has two variants: Text-Independent and Text-dependent speaker identification. Text-independent includes different utterances during enrollment and verification while text-dependent involves same set of utterances during both steps [1-3]. There is normally a trade- off between accuracy and enrollment samples. Greater the duration of the utterances recorded and lesser the number of enrolled models, more will be the accuracy of the system and vice versa [4]. Speaker Identification is further classified into closed-set speaker identification or open-set depending upon its application. Open-set identification is close to speaker verification. In closed-set, test speakers are known to the system. While in open-set, test speaker may not be a part of system and the test user is usually identified by defining a threshold [5, 6]. For instance, if similarity measure of the test user is greater than the threshold, access is guaranteed to the user and vice versa. Speaker Identification has various applications including voice dialing, banking transaction over telephone network, telephone shopping, database access services, information and reservation system, security control for confidential information areas, voice mail and improving customer experience [7]. While designing automatic speaker identification

system (ASI) we have to predefine certain constraints on the speaker data that include language, age, gender, channel and environment on which our system has to be trained [8].

A. Generic Speaker Identification

Speaker Identification broadly involves two phases: Enrollment and verification. Enrollment is the step in which the voice of a speaker is recorded and feature vector (set of features that uniquely identify each speaker) is extracted from each frame which constitutes a template or model e.g. pitch or frequency etc. Enrollment is further divided into two steps: voice recording and feature extraction[9, 10]. In first step, Speech data (we call it S) is acquired using the dedicated hardware like telephone or microphone, and then preprocessed to visualize the acoustic patterns of speech in the form of frequency-time waveform. Speaker dependent features like frequency, pitch and loudness are then extracted from the waveform. Later, this voice recording (S) is divided into cut-out samples of equal length of 10ms, called frames. Hamming window, that remove the discontinuities at the edges is then applied to remove the discontinuities at the edges. Lastly, speaker dependent features, collectively known as a feature vector are extracted from one interval of speech (frame). These feature vectors are then trained for each speaker to form a template or model of that speaker and stored in the database. In verification, the utterance is compared with the multiple existing templates to find the optimal match. Verification is also accomplished in two steps that involves pattern matching and decision. In pattern matching, a test sample (T) is compared with the stored speech model (SM) to calculate match score that defines resemblance of the input feature vector of the newly recorded sample (T) with the known templates stored in a database. Thus, match score counts the number of features of the test sample (T) that have been matched with the recognized speaker model. Different pattern matching algorithms are used for this purpose of which most common are Hidden Markov Model, NN method and Dynamic Time Warping [11, 12]. Lastly based on the matching score, the speaker model (T) is either accepted or rejected (Decision).

B. Motivation

In Pakistan, with the growing trend of e-commerce and online transactions, maintaining and improving the customer experience is one of the core challenges of business and banks. Identifying the speaker over telephone is a challenging area mainly in our telecoms and banking sector. Most of the users, specifically who are semi-literate or non-literate are uncomfortable with remembering

passwords and pins so they often keep their passwords and sensitive information along with them which increases the theft and fraud rate in Pakistan every year.

A lot of work has been done in speaker identification in subsequent languages including English, Tamil, Japanese and Portuguese [13] . But not sufficient work has been done so far in Urdu and work has been done using features like formants only that when used alone will give us very limited information about speaker. However, systems that have been trained for other languages can't be used for Urdu language because each language has its own phonemes and all are designed on their native speakers.

C. Overview

This introductory section has been presented with a goal to describe general framework and need for speaker identification. The endeavor trains different classifiers with the aim to determine such a feature set that cannot only uniquely recognize a speaker but also provide better results as compared to the previous researches which has been conducted within the context of other languages. This has been accomplished by evaluating our dataset on different measures i.e. changing value of K and changing number of utterances etc.

Figure 1 describes the block diagram of ASI. Our research is based on system that is text-dependent and specific for Urdu Language. Section 2 gives an overview of datasets used in existing researches and gives a detailed description on requirements for collecting data set. Section 3 gives an insight on features i.e., MFCCS, Pitch and formants. Section 4 demonstrates training of speaker identification algorithms effectively used for ASI. Section 5 describes the results that comes from evaluation of data set and algorithms and section 6 gives conclusion and future prospects.

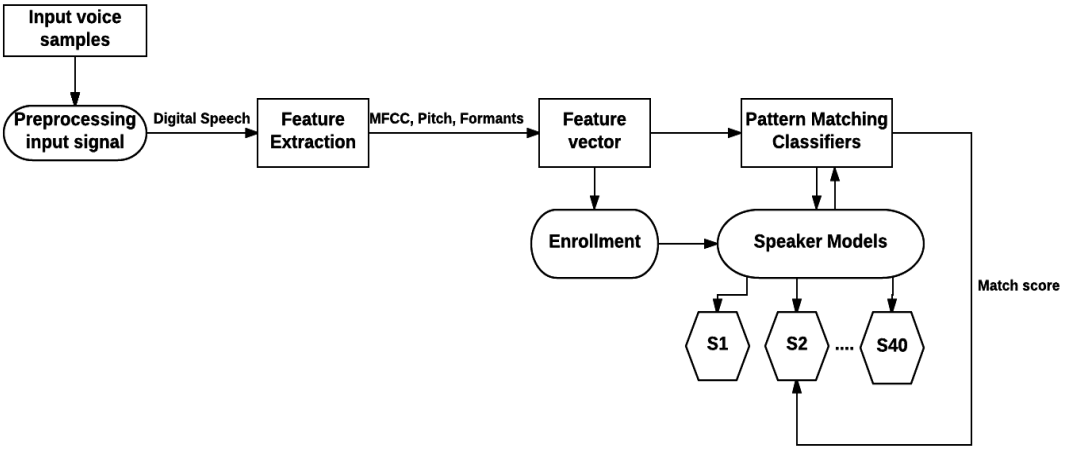

Figure. 1 Block Diagram of Automatic Speaker Identification

II. INPUT DATA REQUIREMENTS

Data collection for speaker identification poses certain constraints like number of speakers, language, medium of recording, size, gender distribution, text selection etc. [14]. Summary of different datasets used in multiple researches for Non-Urdu Languages is shown in the table 1.

TABLE I
SUMMARY OF DATA SETS IN EXISTING RESEARCHES

Number of Speakers	Gender Distribution	Text Spoken(Language medium, sentence, word)	Number of sentence/word(per speaker)	Medium of recording	Accuracy
10	Females only	"May we all learn a yellow lion roar"	6 utterances	High quality microphone	80%
50	Females+ Males	Isolated digits (1-10)	NA	Microphone	71%
40	22 males + 18 females	Telephone speech English corpus	20 utterances	Telephone	68.5%
51	Males only	Telephone quality version of Japanese speech	10 utterances	High quality microphone	96.80%
26	Males only	English Vowels	3 utterances	Recordings done in 5 months	72%

Urdu is a phonetically rich language with a large directory of 44 consonants, 7 long nasal vowels, 7 long oral vowels, 3 short vowels and various diphthongs [15]. All these consonants, vowels are made up of phonemes.

Fundamental distinct unit of language is called phoneme[16]. Phoneme is distinct in a sense that it separates words of a language. Following are major groups of phoneme: Vowel, nasal, fricative, consonants and stops. Our study aims towards vowels and their speaker dependent characteristics. We have recorded five words that cover all vowels as shown in Table 2. Atal et al., 1976 mentioned that larger the number of utterances, larger the accuracy would be[4]. Practically, collecting large number of recordings imposes huge computation but accuracy increases.

TABLE II
WORDS TAKEN FOR RECORDINGS

Words	Pronunciation	Phones (CISAMPA)
A	/e:/	A_Y
E	/i:/	I_I
I	/ɑ:/ /i:/	A_A I_I
O	/o: /	O_O
U	/j/ /u:/	J U_U

Initially we make recordings of 11 speakers. For each speaker, we took 25 recordings so that maximum data against single speaker must be preserved. Later, following this research we found that 11 speakers were not sufficient for making any strong argument and comparing accuracy. So taking all these researches as baseline and for analysis and evaluation purpose, we have taken average of all researches that have been done so far in the domain of speaker recognition and these researches has been conducted on around 40 numbers of speakers (18 females and 22 males).

III. FEATURE EXTRACTION

Once the data collection is completed the next step involved in the simulation environment is to train the system. In order to efficiently train the system, we need to have a feature set. Usually acoustic signals in the waveform contains a lot of parameters that can be either directly extracted or after transforming signal from spectral to frequency domain [17]. For successful speaker identification the most important step is extraction of such parameters from the acoustic signal that represent the maximum user dependent information. Extensive research is available regarding

selection of efficient speech parameters. Ideally, the selective parameters must be efficient enough to represent speaker information and should have some properties like time-stable, easy to compute, occur frequently in acoustic signal, environment independent and least subjected to mimicry [14].

For audio classification, speech features that have been widely used are Mel-frequency cepstral coefficients (MFCC), formants and pitch [18]. The idea behind is to report the accuracy difference by using either MFCC alone or MFCC along with formants and pitch.

A. Mel-frequency Cepstral Coefficients

MFCCs [9, 19] provides the best representation of speech signal and proves more efficient recognition performance. As human ears doesn't follow frequency contents on a linear scale so for each sound with some frequency, f, in Hz, a subjective frequency is calculated on a scale named as 'Mel' scale [9]. The Mel scale has a linear frequency spacing below 1000Hz and logarithmic spacing above 1000Hz. Like a tone with 1 KHz and 40 dBs above the perceptual hearing threshold defined to have a pitch of 1000 Mel's. So, for a given frequency f in Hz, following formula can be used to calculate the Mel's.

$$\text{Mel}(f) = 2595 * \log_{10}(1 + f/700) \quad (1)$$

We have used filter bank to simulate subjective spectrum where there is one filter for every Mel-frequency component. This filter bank consists of spacing and triangular band pass frequency response as well as bandwidth is calculated by a constant Mel-frequency interval. This Mel scale bank has a sequence of triangular band pass filters deigned to perform the band pass filtering supposed to happen in auditory system. This leads to a sequence of band pass filters on Mel-frequency scale having constant bandwidth and spacing. In last step, the log Mel spectrum is converted back to time and in result we get the Mel-frequency Cepstral coefficients (MFCC).

We can do discrete cosine transformation in order to transfer back the Mel coefficients to time domain.

$$C_{\underline{n}} = \sum_{k=1}^{k} (\log S_k) \cos\left\{ n\left(k - \frac{1}{2}\right) * \frac{\pi}{k} \right\}, \quad (2)$$

$$n = 1, 2, \ldots k$$

Whereas S_k, K = 1, 2, … K are the outputs of last step.

B. Pitch and Formants

Pitch is also a user-dependent feature and has been supported in research [18] but pitch is easily to mimic within same gender so using this feature alone doesn't gave fruitful results. Formants are the spectral peaks of the signal and Peterson and Barney have measured first and second formants for a large set of speakers and have documented 60% accuracy. As per above discussion, lesser the number of features, lesser would be the results. Formants are generally calculated as first, second and third only so they gave not much information against single speaker.

IV. TRAINING OF CLASSIFIER

It has been studied that different classification models have been proposed [20-22] to train the system. The classification models are dependent upon their preference/priority which are presented as follows:

- Naïve Bayes
- K-Nearest Neighbor
- Neural Network

First, we have taken the MFCCs features without normalization. By this, we mean that as each recording has different number of frames depending upon the user characteristics like speed of uttering word because different speakers utter the same word with different speed [23]. Taking MFCCs without normalization doesn't leads to useful results.

Therefore, in the second experiment, we take the mean of each Mel cepstral co-efficient for all frames and combine the results of all recordings of a single speaker. This is like let we have a recording of user consisting of 13 * 95 vectors. We first convert it into 13 * 1 vector then do this for all 25 recordings of the single user. Then we make a single file of 13 * 25 vector for a single user. This normalization technique resulted in comparatively better results than without normalization.

In the third experiment, we took normalized MFCCs along with pitch and formants and trained them on different classifiers to compare the accuracy in order to deduce that which set of features gave us the best accuracy. This experiment proved to be the best set of features for speaker identification task.

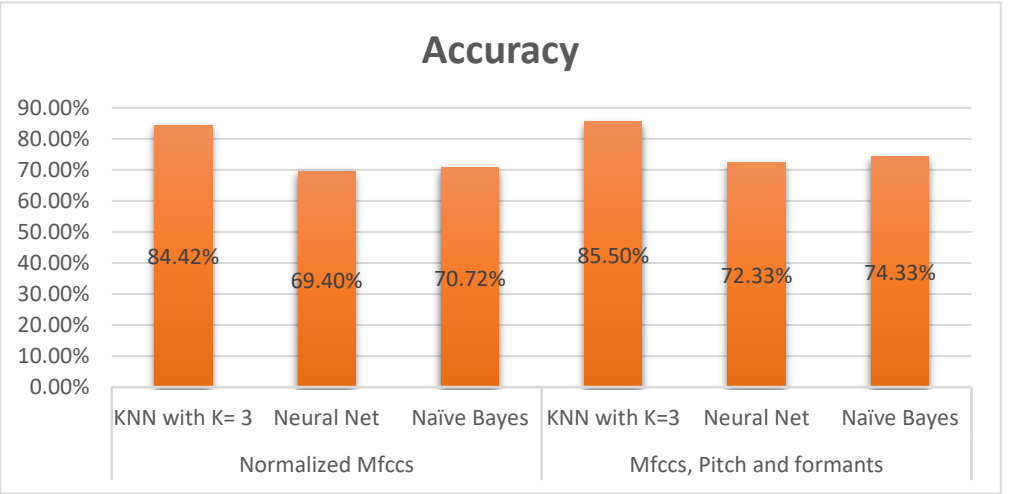

Figure. 2 Accuracy Results of all classifier

As shown in figure 2, it can be seen that KNN with K= 3 was able to produce the better results in both techniques. So there is a high probability to utilize the KNN with K= 3 classification model for training. We have got the best accuracy of 85.50% again by using the K-NN classifier with a combined feature set of MFCC, pitch and formants.

V. Analysis of Results

After training different models, additional experiments were made on data to ensure that the accuracy that have been achieved is sustainable or not. For this, following analysis were made on the dataset:

1. Changing value of K
2. Reducing number of utterances

A. Changing value of K

The NN, or more general k-NN, method can be used to estimate the probability density function (PDF) of the data within a class or itself can be used as a classifier. According to our analysis, the results suggest that the speaker population should directly be proportional to k. This is logical since for large populations, there will be more confusion among the nearest neighbors.

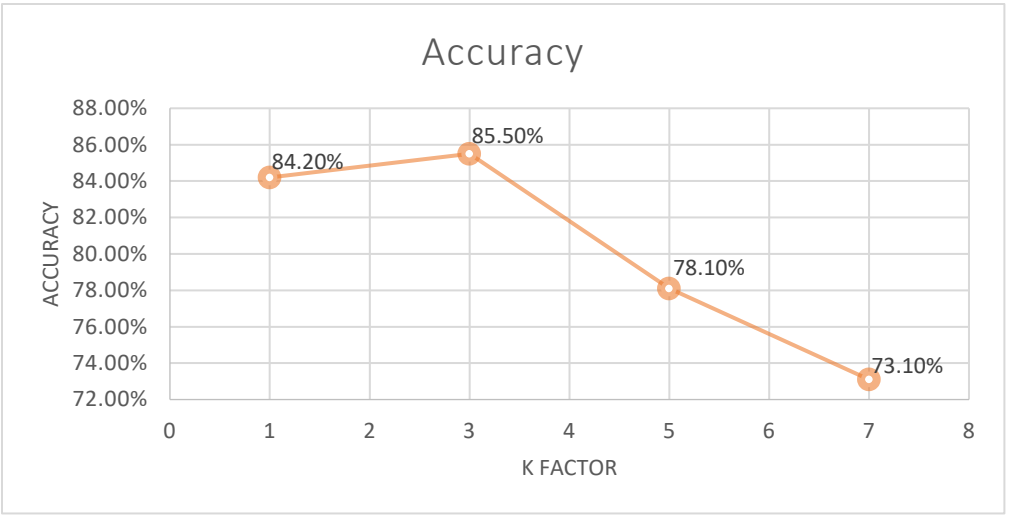

Figure. 3 Effect by changing NN k factor

The experiment has been conducted for identifying the change that might occur in accuracy by changing k factor. Data set has been trained and tested on odd values of k because setting odd value of k supports us in deciding the class label of the test utterance [23, 24]. It has been clear from the simulation results that excessive increase in the nearest neighbor results in drop of accuracy. As shown in Figure 3, accuracy falls as the value of k on the x-axis increases.

B. Reducing number of utterances

Reference [25], experiment has been conducted to measure the effect of reducing number of recordings within the context of classification model. There are two categories of recordings which has been utilized in the experiment such as “18 utterances/speaker and 25 utterances/speaker”.

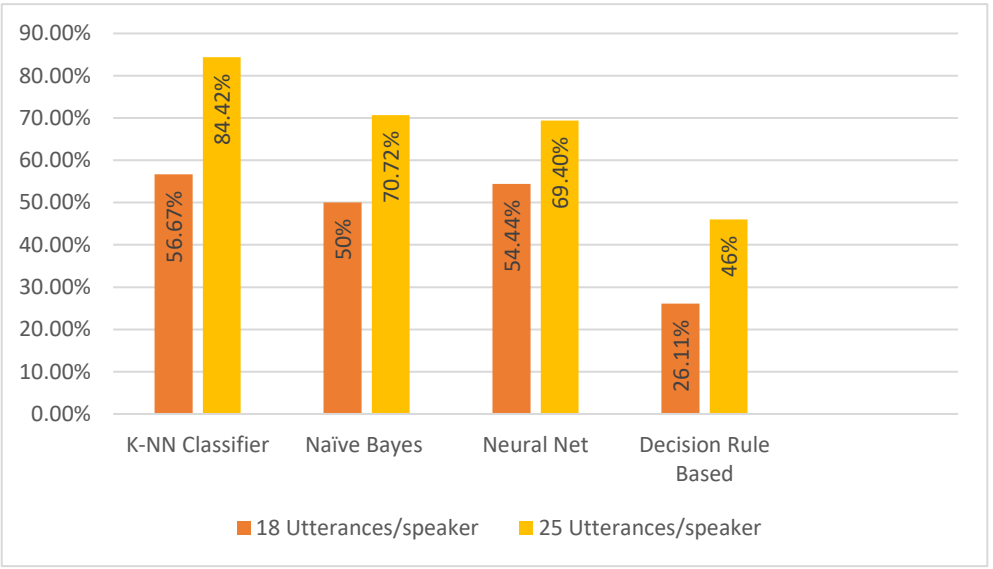

Figure. 4 reducing number of utterances

After conducting the experiment results as shown in Figure 4, it can be visualized that increasing the number of utterances per speaker helps us improving the accuracy measure. The figure clearly demonstrates that all classifiers have resulted in better and improved accuracy against 25 utterances rather than 18 utterances. Additionally, among all of them, k-NN classifier again here has been able to provide the best accuracy as compared to others for both set of utterances. The results obtained by the technique on both set of speakers are 56.67% on 18 utterances/speaker and 84.42% on 25 utterances/speaker.

VI. Conclusion

We came up with a new feature set for ASI that are not simple, robust and easily computable but also gives us high accuracy. We have shown that more competitive system can be made by using MFCC along with pitch and formants rather than using MFCC alone. Using pitch and formants (highly speaker dependent features) clearly demonstrates inter speaker and intra speaker variability. Use of K-NN as a pattern matching classifier has also shown significant improvement. Due to time constraints we have developed a text dependent system so in future we will try to improve the system to text independent system.

As future recommendations, it is highly recommended that porting code to C will also help in efficiency improvement of Automatic Speaker Identification System for Urdu Speech. In addition, Psychological studies have shown that human speech varies over a period of 2-3 years so speaker's data must be updated to maintain the accuracy of the system.

REFERENCES

[1] Reynolds, D.A. and R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech & Audio Processing, 1995. 3(1): p. 72-83.
[2] Kenny, P., et al., A study of interspeaker variability in speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 2008. 16(5): p. 980-988.
[3] Pillay, S.G., et al. Open-Set Speaker Identification under Mismatch Conditions. in Tenth Annual Conference of the International Speech Communication Association. 2009.
[4] Atal, B.S., Automatic recognition of speakers from their voices. Proceedings of the IEEE, 1976. 64(4): p. 460-475.
[5] Reynolds, D.A., Speaker identification and verification using Gaussian mixture speaker models. Speech Commun., 1995. 17(1-2): p. 91-108.
[6] Kalaivani, S. and R.S. Thakur, Modified Hidden Markov Model for Speaker Identification System. International Journal of Advances in Computer and Electronics Engineering, 2017. 2(3): p. 1-7.
[7] Subhashini, P. and T. Pratap, TEXT-INDEPENDENT SPEAKER RECOGNITION USING COMBINED LPC AND MFC COEFFICIENTS.
[8] Raza, A.A., et al. Design and development of phonetically rich Urdu speech corpus. in Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on. 2009. IEEE.
[9] Reynolds, D.A. An overview of automatic speaker recognition technology. in Acoustics, speech, and signal processing (ICASSP), 2002 IEEE international conference on. 2002. IEEE.
[10] Shinde, R. and V. Pawar, Fusion of mfcc & lpc feature sets for accurate speaker identification. International JOurnal of Current Engineering and Technology, 2013. 3: p. 1763-1766.
[11] Zhao, L. and Z. Han. Speech recognition system based on integrating feature and HMM. in Measuring Technology and Mechatronics Automation (ICMTMA), 2010 International Conference on. 2010. IEEE.
[12] Gaikwad, S.K., B.W. Gawali, and P. Yannawar, A review on speech recognition technique. International Journal of Computer Applications, 2010. 10(3): p. 16-24.
[13] Nawaz, O. and T. Habib. Hidden Markov Model (HMM) based speech synthesis for Urdu language. in Conference on Language & Technology (CLT). 2014.
[14] Radha, V. and C. Vimala, A review on speech recognition challenges and approaches. doaj. org, 2012. 2(1): p. 1-7.
[15] Ahmed, Z. and J.P. Cabral. HMM-Based Speech Synthesiser for the Urdu Language. in Spoken Language Technologies for Under-Resourced Languages. 2014.
[16] Huang, J.-T., et al. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. 2013.
[17] Raza, A.A., Design and Development of an Automatic Speech Recognition System for Urdu. 2009, Thesis, FAST-National University of Computer and Emerging Sciences, Lahore Pakistan.
[18] Reynolds, D.A., et al. The 2004 MIT Lincoln laboratory speaker recognition system. in Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on. 2005. IEEE.
[19] Dumitru, C.O. and I. Gavat. A comparative study of feature extraction methods applied to continuous speech recognition in Romanian Language. in Multimedia Signal Processing and Communications, 48th International Symposium ELMAR-2006 focused on. 2006. IEEE.
[20] Marhon, S.A. and D.N.U. Al-Aghar, Speaker Recognition Based On Neural Networks. The Higher Institute For Industry, Misrata, Libya.
[21] Cai, D., et al., Modeling splice sites with Bayes networks. Bioinformatics, 2000. 16(2): p. 152-158.
[22] Katz, M., et al. Sparse kernel logistic regression using incremental feature selection for text-independent speaker identification. in Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The. 2006. IEEE.
[23] Maheswari, N.U., A. Kabilan, and R. Venkatesh, A hybrid model of neural network approach for speaker independent word recognition. International Journal of Computer Theory and Engineering, 2010. 2(6): p. 912.
[24] Sarfjoo, S.S., et al., Using eigenvoices and nearest-neighbors in HMM-based cross-lingual speaker adaptation with limited data. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2017. 25(4): p. 839-851.
[25] Maas, A.L., et al., Building DNN acoustic models for large vocabulary speech recognition. Computer Speech & Language, 2017. 41: p. 195-213.