PalArch's Journal of Archaeology of Egypt / Egyptology

A SPEECH –MAIL ASSISTANCE FOR VISIONLESS

Dr.R.Kumuthaveni[1], Dr.N.Kamalraj[2], Dr.K.Sumangala[3], S.Jisnu[4]
[1]Assistant Professor, Dept of Computer Science(PG),
Kongunadu Arts and Science College, CBE-029,
[2]Assistant Professor, Dept of Computer Science,
Dr.SNS.Rajalakshmi College of Arts & Science, CBE-049,
[3]Assosiate Professor, Dept of Computer Science(PG),
Kongunadu Arts and Science College, CBE-029,
[4]Student, I-year BE(Computer Science Engineering),
PSG Institiute of Technology and Applied Research, CBE-062.

Dr.R.Kumuthaveni, Dr.N.Kamalraj, Dr.K.Sumangala, S.Jisnu, A SPEECH –MAIL ASSISTANCE FOR VISIONLESS-Palarch's Journal Of Archaeology Of Egypt/Egyptology 17(12), ISSN 1567-214x

Abstract – The technology innovations are successful with being accessible to the entire society. On observation of the problems faced by visionless, this paper presents speech supportive e-mailing system, such that the emails can be sent/read without any hassles. Although there are several commercial products for mailing purpose, they undergo accent ambiguity and thus lesser recognition accuracy rates. Besides this visionless people cannot verify the data and also the concept of homophones further reduces the accuracy rate of the system. On consideration of the aforesaid points, this paper conquers all the barriers by the inclusion of a customized dataset. The customer forms a customized dataset and it prompts the customer to utter alphabets one-by-one. Accuracy is the main factor for visionless, rather than time consumption. The proposed work extracts features from the customized dataset by means of Combined Feature Extraction Algorithm (CFEA) algorithm, which is a combination of MEL Frequency Cepstral Coefficient (MFCC), Linear Prediction Cepstral Coefficients (LPCC) and RelAtive Spectral TrAnsform – Perceptual Linear Prediction (RASTA-PLP). Finally, k-NN classifier is employed to distinguish between the alphabets. The experimental results of the proposed work are satisfactory in terms of recognition accuracy rate.

Keywords – Speech recognition, feature extraction, MFCC, LPCC, RASTA-PLP, k-NN classifier.

1. Introduction

Communication is the lifeblood of mankind. Due to the advent of information technology, the means of communication are improvised. Today's technology takes the means of communication to the next grade, which makes sense that all the communications are digitized.

Communication plays a vital role in all the domains such as business, education, inter-personal communication and so on. Communication can be classified into two categories and they are oral and written. Oral communication is to share the ideas and experiences and thereby imparting knowledge to the opposite end. On the other hand, written communication is considered as formal, owing to its reliability and permanence.

The value of written communication is realized even in the historical Bronze Age. The survey of Times of India claims that the Indian user base of mobile internet is the second largest base in the world, as on November, 2015 [1]. Due to the swift drive of the technology, most of the official bodies rely on written communication, which usually happens through the e-mail. The usage of email shows drastic improvement, since 2012. In 2015, the count of worldwide email accounts has reached 4087 millions [2]. Some of the advantages of electronic communication are faster delivery of messages, hassle-free file transfer, 24 :7 based services and environment friendly.

On the other side of the coin, the disabled people cannot enjoy the fruit of the technological progression, which is the disgrace of technological innovations. The main objective of the advancement of technology is to improve the degree of accessibility. Every new sapling of science must be exploited by the society and the same can be made possible by science.

Understanding the complications of visually impaired people, this work proposes a speech recognition system designed exclusively for them. The main intention of the proposed work is to ease the life of visually impaired people. The central theme of this work is the proposal of the customized dataset. The beauty of this dataset is its focus on the individual user, in order to avoid the ambiguity of the accent. The customized dataset takes the alphabets alone into account, which brings in numerous advantages to the system. The advantages of such dataset are reduced training time and enhanced accuracy. The training time is reduced because of the very limited training data. The accuracy of the system is improved by overthrowing issues such as accent ambiguity and homophones.

The proposed work is composed of four phases and they are training, feature extraction, testing and classification. The training phase concerns to train the system with individual alphabets from A to Z. This is followed by the process of feature extraction, which is achieved by CFEA. CFEA is employed such that new features can be extracted from the speech signal. The extracted features of the training set are stored in the database. In the testing phase, the user is allowed to utter an alphabet, which is converted to the text by the process of recognition of that particular alphabet. The process of classification is achieved by k-NN classifier, which is known for its accuracy. The performance of this work is found to be satisfactory in terms of accuracy rate.

2. Existing works

After exhaustive study, we come to a conclusion that the accuracy of the work will get affected, as there are 1,025,109 words in English. It is mandatory to train the system with all these words, so that accurate results can be expected. Besides this, issue such as homophones hit the scene, which is a word being pronounced the same as the other word. For instance, are-or, to-two-too, as-ass, not-knot etc.,

As we focus on visually impaired people, the text must be hundred percent accurate, which cannot be accomplished in this case. Additionally, the meaning of the entire text will get deteriorated, if the system cannot catch the correct word. Some of the major drawbacks associated with such system are prolonged training phase, lesser accuracy, slowest execution, huge memory consumption and so on. These kinds of datasets are available in the market. However, the accuracy rate is affected when these datasets are incorporated, due to the accent issue.

Motivated by these kinds of systems, this work prompts the users to tailor their own dataset, such that the accent issue is overthrown. Besides this, the proposed work focus to reduce the size of the training dataset and to increase the accuracy rate. Both these targets are achieved by the alphabet-by-alphabet utterance.

The reasons for incorporating alphabet-by-alphabet utterance are the error-free text, lesser training time, high accuracy rate. The trade-off is experienced between the time and accuracy rate. However, on the perception of a visually impaired person, accuracy rate is given more importance than the time consumption.

3. Proposed approach

3.1 Overall flow of the work

The proposed work relies on four different phases namely training, feature extraction, testing and classification. The user is prompted to feed the system with the customized dataset. The time required to train the system is very minimal and can be done in a streak. As the dataset is tailored by the user, high rate of accuracy is achieved. The accent and pronunciation of alphabets are diversified with respect to the location. Mostly, the training set of commercial datasets follows the language of native speakers of English. The accent and slang of native speakers of English are completely different from others. This significantly reduces the accuracy of the system, by wrong classification. The overall flow of the proposed work is presented in figure 1.

Accuracy is given more priority, as the visually impaired people cannot correct the text. Though our system takes more time to complete the mail writing process, the accuracy rate is satisfactory. The accuracy rate is affected; if at all the user doesn’t know the correct spell of any word. On the second phase, the features from the training set are extracted by means of CFEA algorithm which clubs MEL Frequency Cepstral Coefficient (MFCC), Linear Predictive Cepstral Coefficients (LPCC), RelAtive Spectral TrAnsform – Perceptual Linear Prediction (RASTA-PLP) into a single system.

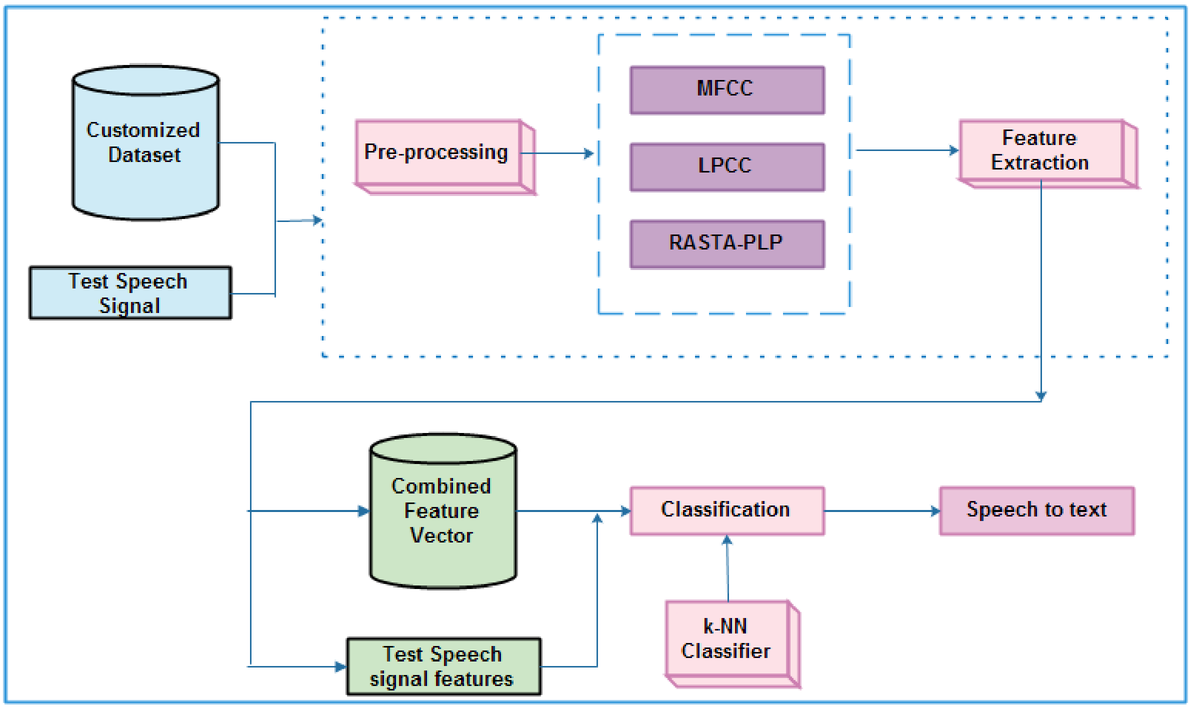

Figure 1: Overall flow of the system

This paves way for detection of many features, which further improves the accuracy rate of the system. In the testing phase, the user can proceed to mail by uttering alphabet by alphabet. The features of the uttered alphabet are extracted and the classifier differentiates between the alphabets to figure out the uttered alphabet.

3.2 Training phase

In this phase, the user has to record the alphabets in order to make the training set ready. This could be done by exploiting sound recorder of the computer and stored. This sort of training process saves much time. Due to the incorporation of this customized dataset, the accuracy of the system is improved. After feeding the training set, the features of the (alphabet) speech signal are extracted and the process of feature extraction is explained in the forthcoming section.

3.3 Pre-processing of a speech signal

The normal sampling frequency of the speech signal is 44.1 kHz. In order to process the speech signals, the samples are sampled to 16 kHz. Initially, the speech signals are pre-emphasized with the help of a filter and the steps involved in pre-processing are depicted in figure 2.

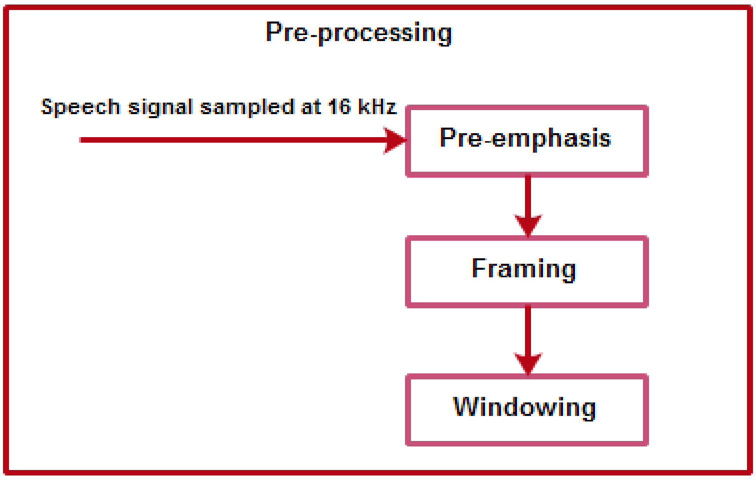

Figure 2: Pre-processing

Usually, the employable filter for pre-emphasis is the first order high pass filter. The main objective of pre-emphasis is to remove the DC part of the signal and also to smoothen the spectral energy. The obtained pre-emphasized signal is divided into several short frames by means of hamming window. The discontinuity being observed in edges are discarded by the hamming window. The hamming window encloses each and every time frame [3, 4].

3.4 Feature extraction

This phase aims at extracting useful features from the train set of data, which plays an important role in the process of classification. This work employs CFEA algorithm, so as to extract features. The CFEA algorithm clubs MEL Frequency Cepstral Coefficient (MFCC), Linear Predictive Cepstral Coefficients (LPCC) and RASTA-PLP. This idea introduces several unique features to the system, such that the speech is recognized accurately. The features extracted by the CFEA algorithm are normalized, in order to maintain a standard range. The working principle of MFCC, LPCC and RASTA-PLP are explained below and are depicted in figure 3.

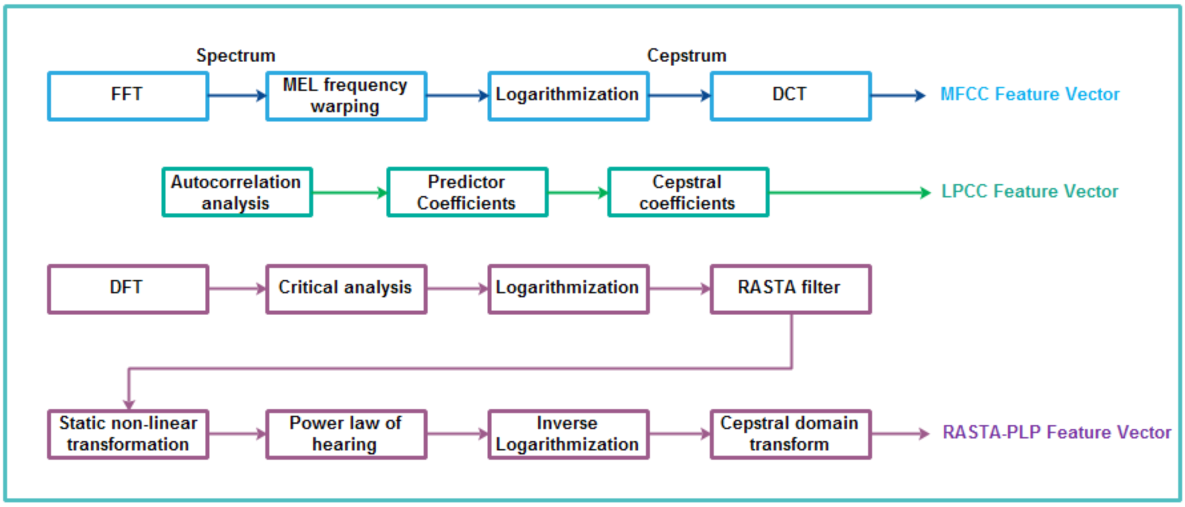

Figure 3: Working principle of MFCC, LPCC and RASTA-PLP

3.4.1 MFCC

MEL Frequency Cepstral Coefficient is one of the most powerful feature extracting techniques for speech recognition. MFCC is known for its accuracy, owing to the logarithmic arrangement of the frequency bands. The frequency bands are equally located in the Mel scale by the Mel frequency spectrum. MFCC performs short-term analysis and so the MFCC vector is computed for each and every frame [5]. MFCC takes the frequency domain features rather than time domain features, which improves the accuracy rate [6,7]. The MEL-scale frequency mapping is computed by

$$Mel_f = 2595 * \log_{10}(1 + \frac{lnr_f}{700}$$

(1)

Mel is the Mel frequency and ln is the linear frequency of the speech signal. The overall pattern of MFCC is presented in figure 3. Initially, the speech signal is categorised into several time frames. The pre-processing step is followed by the application of Fast Fourier Transformation (FFT) over each frame. This is to obtain the frequency components of the speech signal and to expedite the entire process. The next step applies the Mel scale over the transformed frame. The Mel scale follows linear structure until 1 kHz and logarithmic structure at higher frequencies. At last, Discrete Cosine Transformation (DCT) is computed for all the outcomes being present in the filter bank. The task of DCT is to order the coefficients with respect to the degree of significance and the coefficient with least significance, say 0 is eliminated. The outcome of this process is the Mel Frequency Cepstrum Coefficients (MFCC).

Every speech frame is passed through MFCC to arrive at forty two parameters. Out of these, the twelve parameters are original, twelve first order derivatives, twelve second order derivatives, three log energy and another three 0 parameters. These parameters are computed for every speech frame and the feature vector is formed. The feature vector is composed of significant features, which possesses the acoustic features. This work considers twelve coefficient parameters of MFCC.

3.4.2 LPCC

Linear Prediction Cepstral Coefficients is presented in [8,9]. LPCC is based on Linear Prediction Coefficients (LPC) with recursive technique. The computational complexity of this technique is minimal, as it doesn't involve any transformation over the speech signal [10]. LPCC takes over all the merits of LPC[11]. Autocorrelation technique is the most important part of LPCC and is done by the following equation.

$$ac(i) = \sum_{k=1}^{K_w-1} seg_w(k) seg_w(n-p)\, ; 0 \leq p \leq$$

(2)

where K is the length of the window and seg is the windowed segment. The $ac(i - k$ coefficients present an autocorrelation matrix. The cepstral coefficients are obtained by recursive method and by keeping eqn.2 as the base.

Let auto-correlated vector and the cepstral coefficient vector be represented by [a0,a1,a2,..ap] and [c0,c1,c2,…cp] respectively. The process of recursion is given in the equations from 3 through 5.

$$c_0 = ln\gamma$$

(3)

$c_i = a_{i+} \sum_{k=1}^{i-1} \left(\frac{k}{i}\right) c_k a_{i-}; 1 \ \leq i \leq$

(4)

$c_i = \sum_{k=1}^{i-1} \left(\frac{k}{i}\right) c_k a_{i-k}; i >$ (5)

γ is the gain term in the autocorrelation matrix.

ι are the cepstral coefficients

$a$are the predictor coefficients

3.4.3 RASTA-PLP

RASTA-PLP is an enhancement of PLP, in which an exclusive band-pass filter is included into every frequency sub-band of the PLP [12]. The band-pass filter inclusion is to smoothen the short-term noise variation. Initially, the critical band power spectrum is computed and the spectral amplitude is transformed by means of a static non-linear transformation, which is then followed by the filtration of time trajectory by the band pass filter. The filtered signal is then transformed by the expanding static nonlinear transformation.

The power law of hearing is simulated by multiplication of equal loudness curve and the power is increased by 0.33. Finally, an all-pole model of the resulting spectrum is computed.

3.4.4 Feature extraction by CFEA algorithm

This work sandwiches MFCC, LPCC and RASTA-PLP and all these feature extraction techniques produce twelve coefficient parameters. Thus, this work considers a total of thirty six coefficient parameters and is represented in figure 4.

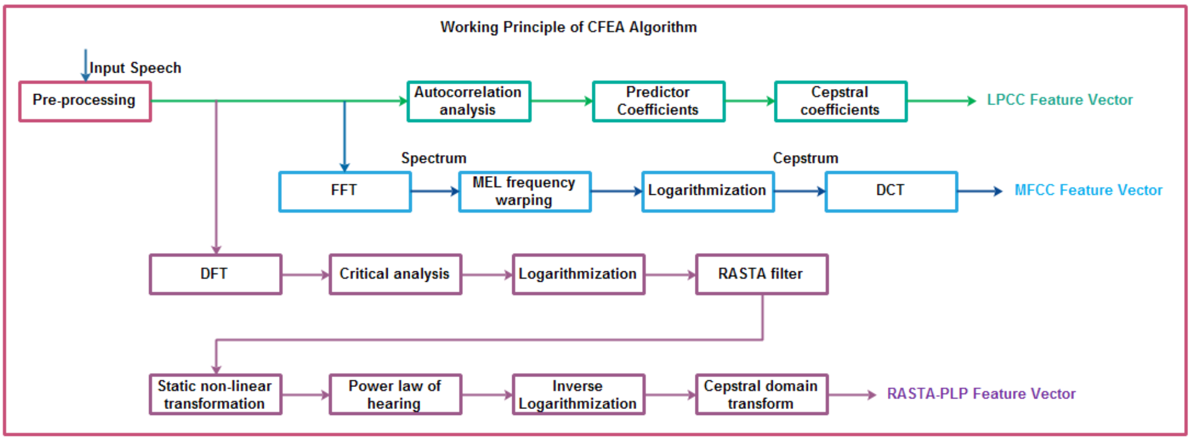

Figure 4: Working principle of CFEA algorithm

The purpose of sandwiching all the aforementioned feature extraction techniques is to arrive at efficient set of features, which can boost up the recognition accuracy of the system.

3.5 Testing phase

The testing phase actually intends to allow a visionless user to exploit the mailing system. Initially, the user can either login to the existing account or a new mail account can be created. The text to speech synthesis is done by the inbuilt function of Matlab. The text to speech synthesis is also necessary, as the visionless people cannot read the requested information from the service provider. The user can pass through each and every field by means of the enter key. All the fields are filled up one by one by the user’s utterance of alphabet by alphabet.

After signing into the mail account, the user is notified through voice about the unread mails. This is followed by the composition of mail and finally the mail is sent to the user. In case of a mail undelivered notice, the system notifies the user.

3.6 Classification by k-NN classifier

It is the task of a classifier to differentiate between the uttered alphabets. The uttered alphabet must be converted to text and the misclassification rate is not tolerable. This work employs k-Nearest Neighbour (kNN) as the classifier, which is a supervised learning algorithm. The distance between the queried entity and each entity of the training set is computed and clusters are determined. This is followed by the comparison of the queried speech signal with the trained speech signal. kNN distinguishes between the alphabets by means of voting strategy by the clusters [13,14].

This work utilizes the Euclidean distance measure $Euc_{dis}(a, b$ to calculate the distance between the queried (a) and the trained (b) speech signal. Both $a\ and$ have a set of features $a = \{a1, a2, ..an$ and $b = \{b1, b2, ..bn$ and the distance is computed by

$$Euc_{dis} = \sum_{i=1}^{n} \sqrt{a_i^2 - b} \qquad (6)$$

From the clusters, the label of test speech signal is determined by the majority voting by t randomly utilizes one to ten for different set of experiments.

The time it takes for classification is little higher, as it compares the queried speech signal with all the trained speech signals. However, the time consumption for classification is minimal in the proposed work, as the training set is very limited.

3.7 Case study

Initially, the visionless people are prompted to customize the dataset. The time consumption to frame such dataset is considerably lesser, because of the utterance of individual characters. This is followed by the technical process namely feature extraction, which extracts the features from all the alphabets. The so extracted features are stored in the database, for future reference. On the testing phase, the user has to utter alphabet-by-alphabet, which may seem to be tedious. However, the accuracy rate is maximum, which is the expectancy of visually impaired people.

Initially, the user has to utter either ‘y’ or ‘n’ to the query ‘create new account’. If the uttered alphabet is ‘y’, then the page to create a new account is opened. The required fields are read by a voice, in order to assist the user. For instance, if the system prompts to enter the ‘First name’, then the user has to utter his/her name by means of alphabets. Once the utterance of first name is completed, then the user has to press enter key, so that the system proceeds with the next field. By this way, all the required fields are filled up and the new account is created. The diagrammatic representation of the concept is presented in figure 5.

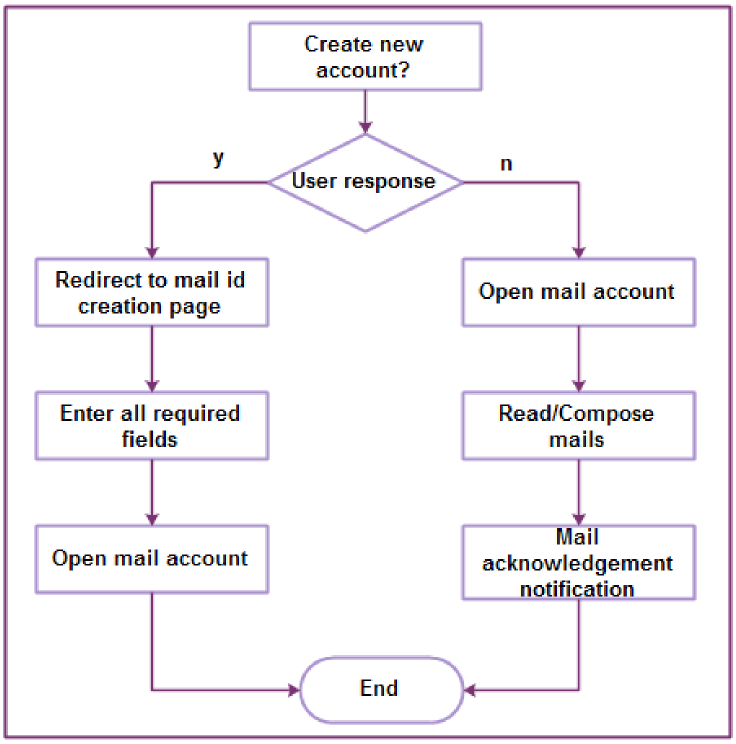
Create new account?
User response
y
n
Redirect to mail id creation page
Enter all required fields
Open mail account
Open mail account
Read/Compose mails
Mail acknowledgement notification
End

Fig 5: Case study

In case, if the user's response is 'n', then the user is redirected to the mail account login page. Once the user has passed successful authentication procedure, then the user can read/send mails. The user is initially notified with the new mails. The user can compose new mail by uttering letter-by-letter. The user is prompted to press the enter key, once the mail creation is done. The mail acknowledgement will be notified to the user, as well.

4. Performance evaluation

The performance of the proposed approach is evaluated in terms of accuracy rate and misclassification rate. The training dataset contains thirty six files, which contains the individual characters along with numerals (26+10=36). By this way, the proposed work is analysed by 10 different speakers with different accent. However, the proposed work proves its efficacy in terms of recognition accuracy.

The experimental results of the proposed work are compared with MFCC, LPCC and RASTA-PLP individually. The proposed work is evaluated with regard to the character count in the mail. As the word count increases, the performance of the analogous techniques deteriorates. However, the proposed work shows stable outcome and the average accuracy rate of the proposed work is 97.31%. The experimental results are presented in figures 6 and 7.

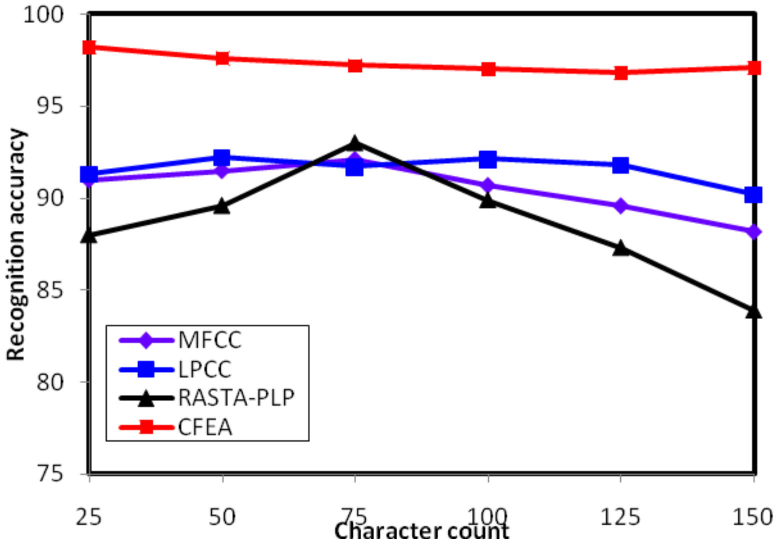

Fig 6: Recognition accuracy wrt character count

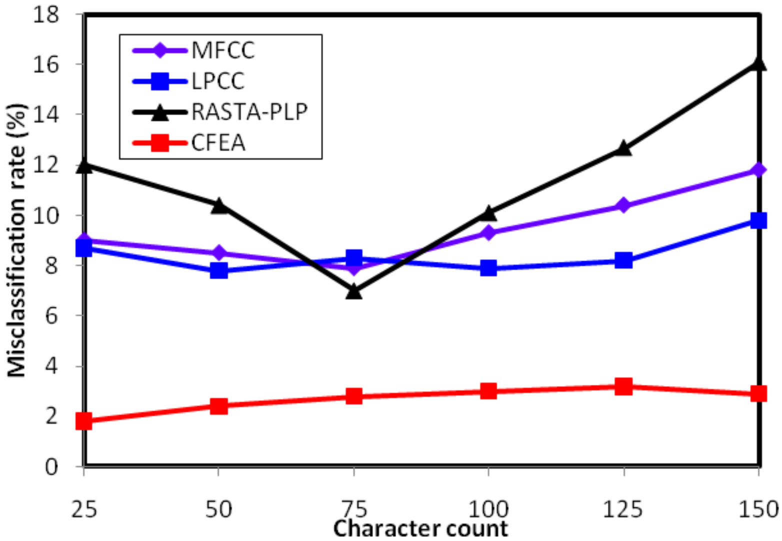

Fig 7: Misclassification wrt character count

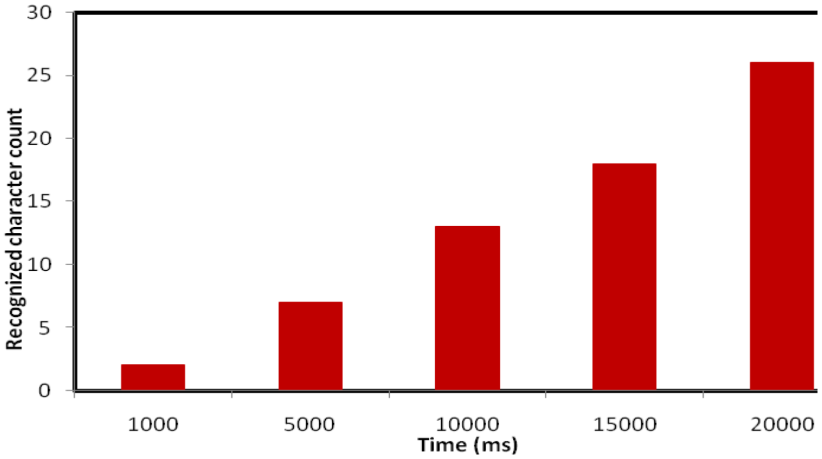

Fig 8: Character recognition rate w.r.t Time

The graph presented in fig.8 proves the capability of the proposed algorithm to recognise individual characters with respect to time. Any speech recognition algorithm must be able to arrive at maximum accuracy with least misclassification rate. This means that the algorithm can recognize the uttered voice effectively with lesser error rates. The proposed work mainly focuses on the welfare of the visually impaired people and thus the accuracy rates must be greater. From the experimental analysis, it is evident that the proposed CFEA algorithm shows greater recognition rates and the least misclassification rate. Thus, the proposed work serves its purpose effectively.

5. Conclusion

This paper presents a novel system for visionless, through which the mail can be sent and read. As the visionless people could not verify the data being spelt, we focus on the accuracy rate rather than the time consumption. To achieve greater accuracy rates, this paper incorporates alphabet-by-alphabet utterance model, CFEA algorithm for feature extraction and k-NN classifier. The CFEA algorithm is the combination of MFCC, LPCC and RASTA-PLP. The performance of the proposed work is tested against MFCC, LPCC and RASTA-PLP individually and the proposed work outperforms the analogous techniques.

References

[1] http://timesofindia.indiatimes.com/tech/tech-news/IAMAI-Indias-internet-user-base-to-hit-402-million-second-highest-in-the-world/articleshow/49816190.cms

[2] http://www.radicati.com/wp/wp-content/uploads/2015/02/Email-Statistics-Report-2015-2019-Executive-Summary.pdf

[3] Goranka Zoric, “Automatic Lip Synchronization by Speech Signal Analysis,” Master Thesis, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Oct-2005.

[4] Lahouti, F., Fazel, A.R., Safavi-Naeini, A.H., Khandani, A.K, “Single and Double Frame Coding of Speech LPC Parameters Using a Lattice-Based Quantization Scheme,” IEEE Transaction on Audio, Speech and Language Processing, Vol. 14, Issue 5, pp. 1624-1632, Sept-2006.

[5] Shrawankar, U. and Thakare, V. M. “Techniques for feature extraction in speech recognition system: A comparative study”, International Journal of Computer Applications in Engineering, Technology and Sciences, pp.412-418, 2010.

[6] Lei Xie, Zhi-Qiang Liu, “A Comparative Study of Audio Features For Audio to Visual Cobversion in MPEG-4 Compliant Facial Animation,” Proc. of ICMLC, Dalian, 13-16 Aug-2006.

[7] Alfie Tan Kok Leong, “A Music Identification System Based on Audio Content Similarity,” Thesis of Bachelor of Engineering, Division of Electrical Engineering, The School of Information Technology and Electrical Engineering, The University of Queensland, Queensland, Oct-2003.

[8] Antoniol, G., Rollo, V. F., & Venturi, G. (2005). Linear predictive coding and cepstrum coefficients for mining time variant information from software repositories. In Proceedings of the 2005 international workshop on mining software repositories.

[9] Rabiner, L., & Juang, B. (1993). Fundamentals of speech recognition. Prentice hall.

[10] Watts, D. M. G. (2006). Speaker identification-prototype development and performance. University of Southern Queensland

[11] Kumuthaveni, R, Chandra, E, "Fuzzy Weight with Voronoi Centroid Vector Quantizer (FW-VCVQ) for Emotional Classifier with Tamil Speech Signals" Special Issue, International Journal of Pure and Applied Mathematics(2018).

[12] Hermansky, H., Morgan, N., Bayya, A. and Kohn, P. (1991) The Challenge of Inverse-E: The RASTA-PLP Method. 1991 Conference Record of the 25th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, 4-6 November 1991, 800-804.

[13] Pallabi, P., & Bhavani, T. (2006). Face Recognition Using Multiple Classifiers. In International conference on 18th IEEE tools with artificial intelligence, 2006. ICTAI '06

[14] Liu, C.-L., Lee, C.-H., & Lin, P.-M. (2010). A fall detection system using k-nearest
neighbor classifier. Expert Systems with Applications, 37(10), 7174–7181.