Automatic Recognition of Correctly Pronounced English Words using Machine Learning

Ronalyn C. Pedronan, Rizaldy A. Manglal-lan Jr.,
Kristine Joy B. Galasinao, Reychell P. Salvador, and James Patrick A. Acang*
Department of Computer Science, Mariano Marcos State University, City of Batac 2906, Ilocos Norte, Philippines

ARTICLE INFORMATION

Article History:
Received: 21 March 2017
Received in revised form: 9 November 2017
Accepted: 11 December 2017

Keywords:
digital signal processing; Hidden Markov Model; Mel Frequency Cepstral Coefficient; Pronunciation Recognition; Speech Recognition.

**Corresponding author: James Patrick Acang (jamespatrickacang@gmail.com)*

ABSTRACT

Speech recognition is a form of human-machine communication where interpreting speech is done by the computer. This research deals with the problem of recognizing correct pronunciation of words in English. In view of using this technology to help in education, the researchers gathered voice samples from middle graders and they labelled them based on ground-truth, English, pronunciation data from Google. The words were based from the current curriculum of the samples. The words were also clustered according to syllables to see how the model performs as the complexity of the words to be recognized is increased. Since there are numerous voice or speech features to consider, the researchers selected three of the known feature extraction techniques subjected for evaluation. Results show that the Mel Frequency Cepstral Coefficient with Linear Predictive Coding model have better performance with high and stable recognition rates compared to the other models. It was also observed that the model only needs four syllables to reach its optimum 100% recognition rate when recognizing English words. To make the model more robust to noise, an automatic signal segmentation approach is needed to detect the significant components of the signal for analysis.

Introduction

Voice Analysis is one of the technological advancements that has been developed nowadays. This has been investigated as a natural form of Human-Machine Communication. It is focused on the understanding of speech generation, coding, transmission and recognition (Anusaya et. al., 2009). Voice analysis decodes the analogue signals to digital signals to be used in computers. Speech recognition is in line with this idea. Speech recognition is the process of machine interpretation or understanding voice commands from spoken words it receives. There are two

subsystems of Speech Recognition, the Automatic Speech Recognition (ASR) and Speech Understanding (SU) (Varshney, 2014). The goal of ASR is to transcribe natural speech while SU is to understand the meaning of the transcription.

Voice Recognition systems, can be classified into two categories speaker-dependent and speaker-independent (Aurora et. al., 2012). Speaker-dependent systems work by comparing a whole word input with a user-supplied pattern while speaker independent systems require no training operations. Voice recognition systems perform two fundamental operations: signal modelling and pattern matching. Signal modelling represents process of converting speech signal into a set of parameters. Pattern matching is the task of finding parameter set from memory which closely matches the parameter set obtained from the input speech signal (Ananthi et. al., 2013).

To ensure high recognition rates, proper selection of features should be considered. Feature enhancement, distribution normalization, and noise robust feature extraction are often used. Feature enhancement tries to remove the noise from the signal, such as in spectral subtraction (SS) (Muzaffar et. al., 2005). Distribution normalization reduces the distribution mismatches between training and test speech, like those presented in cepstral mean subtraction (CMS) (Boll, n.d.) and in Cepstral Mean and Variance Normalization (CMVN) (Furui, 1981). Noise robust features include improved Mel Frequency Cepstral Coefficients (MFCCs), which is similar to root-cepstrum (Viikki, 1998). One feature extraction which has a melodic cepstral analysis is MFCC (Sarikaya, 2001). It represents the dominant features used in speech and speaker recognition domains (Adams, 1990). Other feature extraction method is the FFT (Fast Fourier Transform) which is used to make the spectrum of each windowed sequence be computed after the MFCC feature. To fuse everything together, voice or speech recognition uses various machine learning models to get the best possible accuracy from data. Some of these include the Hidden Markov Model (HMM) (Barbu, 2007) and the Artificial Neural Network (ANN) (Shi,et. al., 2006). Figure 1 presents the HMM where y_i are the observed variables and the s_i are the hidden states.

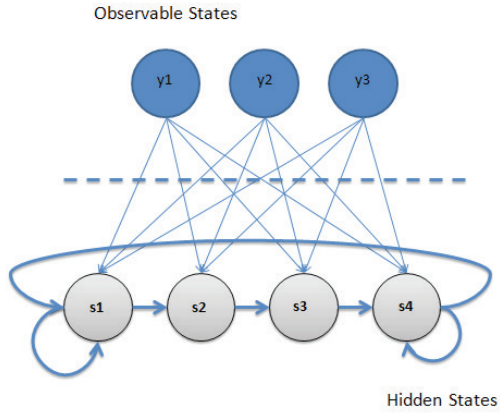

Figure 1. The Hidden Markov Model.

However, there is still a vague space on how pronunciation is represented in the analysis as this is an inherent or hidden entity (feature) in speech. These speech features are captured and described by feature extraction methods but because of the diverse number of feature extraction techniques in voice analysis (Krenker, n.d.), it is difficult to determine the suitable feature for the domain of recognizing correct pronunciation. Hence, this study is aimed to determine the model to recognize correct pronunciation in English words. Specifically, this research is aimed to accomplish the following activities: collect voice samples from middle graders to be used in the training and testing of the recognition models; gather ground-truth word pronunciation data or the basis to be used in the labelling of the said collected middle grader voice samples; build three recognition models from three known voice features; and evaluate the recognition models based on sensitivity, specificity, and accuracy metrics.

In the field of education, proper pronunciation of words, especially English words, is necessary as this is widely used in

communication (Bagge, 2001). In this area, pronunciation recognition can be applied as a bridge in learning languages with the aid of an individual expert or the computer itself. In fact, the Philippines uses various dialects hence an application powered by a pronunciation recognizer can be very useful in learning these languages. Since the platform is pronunciation driven, its application can extend from local to international language learning (Alsulaiman et. al, 2011; Lyu et. al, 2014; Nitta et. al, n.d.). This recognizer can also be implemented to voice operated user interfaces that can be used in instruction providing extraordinary learning environment to students. This simply shows that the recognizer has unending potentials to be effective from simple text-to-speech to complex applications.

Literature Review

The field of Digital Signal Processing (DSP) has drastically and significantly improved since its conception (Anusaya et. al., 2009; Aurora et. al., 2012; Muzaffar, 2005; Furui, 1981; Viikki, 1998; Sarikaya, 2001). A manifestation of which is the development of DSP driven applications like the Google Voice Search (Hispanicallyspeakingnews, 2011), VLingo (Maurice, 2013) and the Siri Assistant (Ludwig, n.d.). The vision of the researchers is to exploit these capabilities and to apply these in education for students.

In speech recognition, speech features are very important. In fact, Thakare emphasized that speech signals carries all auditory information as compared to speech feature extraction methods that can effectively be used in various speech recognition domains. He highlighted in this work that the Mel Frequency Cepstral Coefficient feature reduces the frequency information of the speech signal into small number of coefficients which is relatively fast to compute. He also stressed in the work that the Fast Fourier Transform feature is good because of its linearity in the frequency domain as it does not discard or distort information in any anticipatory manner. Moreover, he gave a highlight on the Linear Predictive Coding feature that it reduces error rates found in difficult conditions.

The fast and effective performance of the Mel Frequency Cepstral Coefficient was exploited in the work of Daphal and Jagtap (2012). They reported that the said feature has significant impact in their classifier. The same strategy was employed in the work of Sapijaszko and Michael (2012) where they tried to experiment with this feature with frame algorithms. In this research Linear Predictive Coding was also considered to enhance the recognition rates.

Noise is also a factor that degrades the performance of the recognizer. In the work of Jarng (2011), noise is included as part of the subject. In this research he asserted that Mel Frequency Cepstral Coefficient is resilient to noise but including new parameters can dramatically improve recognition.

All these work have focused on the recognition of speech itself. The researchers have extended this by considering the characteristic of the speech which is pronunciation.

Methodology

In machine learning, data is required to develop models to capture the behaviour of the subject. Similarly, in pronunciation recognition, voice samples are needed to construct these models for investigation.

The Speech Recognition Process

Speech Recognition is the process where speech signals are used to create recognizers for speech (Mastin, 2011). The speech signals are usually processed in digital representation as it undergo series of processes. A typical speech recognition process include: data gathering, feature

extraction and model construction, and analysis as shown in figure 2.

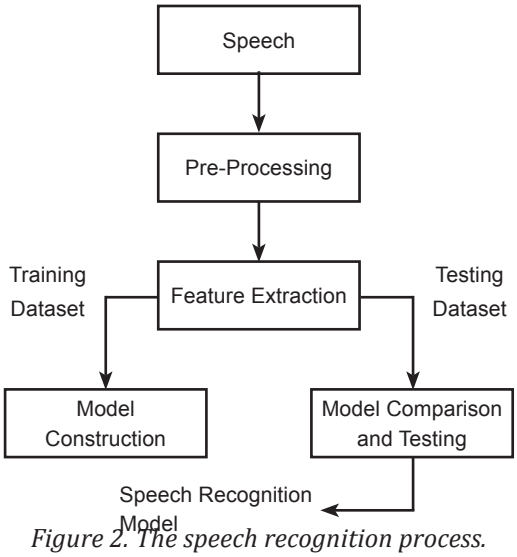

Figure 2. The speech recognition process.

Data gathering includes the process of collecting voice samples. Gathered data are processed and divided into testing and training set. The training set is used to build the recognition models while the test set is used to analyse the models. This process involves extraction of features that are relevant for classification, which are common in both phases.

In the training phase, the parameters of the classification model are estimated using a large number of class examples (training data). During the testing or recognition phase, the feature of the test pattern (test speech data) is matched with the trained model of each and every class. The test pattern is declared to belong to that model who matches the test pattern best (Kale, n.d.). The analysis stage deals with suitable models for further analysis and evaluation. The recognition rates of the different feature extraction methods are shown in the Experimental Results section.

The speech signal serves as input for the speech recognition process. Pre-processing describes any type of processing performed on raw data to prepare it for another processing procedure. In here, transformations of the data into a format that will be more easily and effectively processed are done. When the data input of an algorithm is too large to be processed, then it can be transformed into a reduced set of features. This process is called feature extraction. The extracted features are expected to contain the relevant information from the input data so that the desired task can be performed using this reduced representation instead of the complete initial data. In a dataset, a training set is implemented to build a model, while a test (validation) set is to validate the models built. Data points in the training set are excluded from the test set. To construct a model, accurate labelling and tagging of the dataset are needed. Analysis is the process where the test data together with the model are evaluated to get the recognition rate and to describe how the model behaves. Moreover, the output is the result to be produced by the model out of the inputted data.

Data Acquisition

To enable and determine the correct pronunciation of each word, we used ground truth word recordings from Google (Barett, 2006). This is used to tag or label the gathered voice samples for the model construction. Manual tagging was utilized to ensure that the collected voice samples are labelled correctly. In view of using this technology in education, we selected 300 middle graders from Ilocos Norte, Philippines to gather voice samples. Two-hundred voice samples per word were gathered for correctly pronounced words, half of which are males and the other are females. This is to capture the properties of the voice on both genders. Fifty percent (50%) of which, where each gender is equally represented, will constitute the training set and the other fifty percent (50%) is the testing set. Each student was requested to speak for words, five times each, as they are recorded to capture the pronunciation characteristics of the signal.

Since the researchers are concerned with pronunciation, we also considered mispronounced words. To take the mispronounced words into consideration, the other 100 students were considered to gather mispronounced samples. The same gender arrangement as the correctly pronounced words is followed. This provided 50 males and 50 females who mispronounced voice samples for the testing set collectively.

The researchers also selected and clustered the word samples by difficulty (syllables) based on the current curriculum of the samples. Fifteen (15) words per cluster were considered. The researchers considered five clusters for one syllable to five syllable words to model and capture the complexity of the word being pronounced. That is, easy for the one syllabled words and difficult for the five syllabled words. This is shown in the table below.

A key ingredient to an accurate recognizer of this domain is the proper selection of features that can capture the characteristic of the word pronunciation effectively. In here, the researchers considered three of the commonly known feature extraction techniques in literature (Karpagachelvi et. al., 2010). These features include: Mel Frequency Cepstral Coefficient (MFCCs), Full Fast Fourier Transform (FFT) and Mel Frequency Ceptral Coefficient with Linear Predictive Coding (MFCC + LPC).

MFCC is the dominant feature used in speech recognition systems such as systems that can automatically recognize speech spoken into a computer. They are also common in speaker recognition, which is the task of recognizing people from their voices (Vimala et. al., 2012). Noise sensitivity is one of the considerations for choosing this method. MFCC are not very robust in the presence of additive noise, and

Table 1
List of Words per Cluster

1 Syllable	2 Syllable	3 Syllable	4 Syllable	5 Syllable
ache	compare	abattoir	acquaintance	civilization
aide	complete	accurate	advantageous	economical
aisle	compose	aerospace	anaesthesia	enthusiasm
arm	congress	antimissile	annexation	inconceivable
art	connect	assemblage	beneficial	inexhaustible
ash	conscious	circuitous	caricature	inextricably
awed	consent	clandestine	catastrophic	investigation
badge	corsage	combatant	choreography	itinerary
bait	cottage	credulous	clairvoyance	pronunciation
balm	council	disputant	diminutive	unavoidable
beach	country	fractional	dirigible	unconquerable
beat	couple	icicle	exigency	university
beige	courage	negligee	inclination	vocabulary
bend	cousin	pathetic	interlining	
biped	cover	posthumous	lamentable	

so it is common to normalize their values in speech recognition systems to lessen the influence of noise (Li, J. et. al., 2013). The job of MFCC is to accurately represent the phoneme being produced.

The following formula were used (Mermelstein, 1980):

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n}^{N} = 1\, n^2} \qquad (1)$$

where d_t is a delta coefficient, from frame computed in terms of the static coefficients c_{t+N} to c_{t-N} and N is the analysis window. A typical value for N is 2.

FFT is the traditional technique to analyze frequency spectrum of the signal in speech recognition (Ernawan et. al., 2011). As compared to methods exploiting knowledge about the human auditory system, the full FFT spectrum carries relatively more information about the speech signal. The logarithm of the FFT spectrum is also often used to model loudness perception. The full FFT formula is defined as:

$$X(k) = \sum_{j=1}^{N} x(j)\, e^{\frac{-2\pi i}{N}(j-1)(k-1)} \qquad (2)$$

where $k = 0, \dots, N-1, x(j)$ is the sample time index j and i is the imaginary number $\sqrt{-1}$. $X(k)$ is a vector of N values at frequency index k corresponding to the magnitude of the sine waves resulting from the decomposition of the time index signal.

LPC is one of the most powerful speech analysis techniques and is a useful method for encoding quality speech at a low bit rate (Speech, n.d.). However, LPC cannot stand alone, it only serves as an enhancer to feature extraction. LPC analysis is usually most appropriate for modelling non-nasalized vowels which are periodic.

MFCC is a representation of the speech signal as a linear cosine transform of its log power spectrum on a nonlinear Mel scale of frequency as shown in the following formula (Anand et. al., n.d.).

$$mel(f) = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \qquad (3)$$

In this formula, f signifies the frequency of the speech signal. Discrete cosine transform (DCT) is finally applied to convert log Mel spectrum into time domain. The result of this conversion is called Mel Frequency Cepstral Coefficients.

When the Linear Prediction Analysis was developed, the basis of it is the prediction of current sample as linear combination of past samples, where is the order of prediction

$$\hat{S}(n) = -\sum_{k=1}^{p} a_k * s(n-k) \qquad (4)$$

and $a_k's$ are the linear prediction coefficients and $s(n)$ is the pre-processed signal as shown in the aforementioned formula. Then, the prediction error is defined as:

$$e(n) = s(n) - \hat{s}(n) = \hat{S}(n) + \sum_{k=1}^{p} a_k * s(n-k) \qquad (5)$$

The primary objective of this method is to minimize the total prediction error $\sum_{-\infty}^{\infty} e^2(n)$ and to find the linear prediction coefficients.

Three models were constructed for each of these feature extraction methods from the same training dataset. Each of them was evaluated with metrics using the test dataset. These were built using HMM since this offers faster training compared to the other models (Sak, 2015). In here minimal HMM configuration was applied. Five states, which is commonly used in speech recognition, was utilized (Rabiner, 1989). To estimate the parameters, the researchers used the Baum-Welch algorithm, a variant of the well-known Expectation-Maximization algorithm (Baum, 1970; Dempster, 1977).

Analysis and Performance Comparison

The collected data were analysed together with the methods to determine which of them will be tagged as the best performer. In here, those models are created based on 3 different features. The training data is used to compose these models. Correctly pronounced words were gathered and properly labelled. Five (5) speech samples per word, for each student sample, were used in the training. The labelling is based on ground truth data from Google (Li, n.d.). This is to capture the pronunciation details of the voice.

Since the researchers also gathered and properly labelled mispronounced words, the recognition of mispronounced to properly pronounced words is achievable. Since it is assumed that the words for each test are known, then the classification is binary. That is, if it is correctly pronounced or otherwise.

These models were analysed using the test data and were compared based on accuracy, specificity, and sensitivity. The formula of the metrics are the following:

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (6)$$

$$Sensitivity = \frac{TP}{P} \quad (7)$$

$$Specificity = \frac{TN}{N} \quad (8)$$

TP refers to the number of true positives or the recognized correctly pronounced words while *TN* refers to the number of true negatives or the correctly recognized mispronounced words. *FP* and *FN* refers to false positives and false negatives respectively which defines the number of misclassified word pronunciations. *FP* is the number of misclassified mispronounced words while *FN* is the number of misclassified correctly pronounced words. Sensitivity or true positive rate determines the capacity of the model in recognizing correctly pronounced words while sensitivity or true negative rate works on the mispronounced terms. Moreover, accuracy determines how good the model in recognizing correctly pronounced to mispronounced words.

Results and Discussion

In speech recognition systems, training datasets are used to create the models for analysis. The test dataset is needed to determine the accuracy, sensitivity, and specificity of the different models. In this case, since the researchers used the three feature extraction methods namely: MFCC, FFT, MFCC+LPC, three models were compared with the metrics.

In here researchers used codes to illustrate the model and the metric in each box plot. A prefix (the code before the dash) represents the models that were compared. In here ML represents the MFCC and LPC (MFCC+LPC) model, M represents the MFCC model and F represents the full FFT model. The suffixes (code after the dash) represent the metric done in each plot. In here A, Se, Sp represents accuracy, sensitivity and specificity, respectively. The (+) symbols are the outliers referring to the values that are numerically distant from the rest of the data points.

It was observed from the data that the models have the same behaviour in each of the clusters. Say for example, it was observed that most of the models suffer from the 1-syllable group due to the limited signal to work on. Say for example, MFCC+LPC misclassifies the words *ache, arm,* and *beat*; MFCC misclassifies the words *badge, bait,* and *balm*; and FFT misclassifies the words *ache, aide,* and *biped*. They have increasing accuracy while the number of outliers decreases as the syllables are increased.

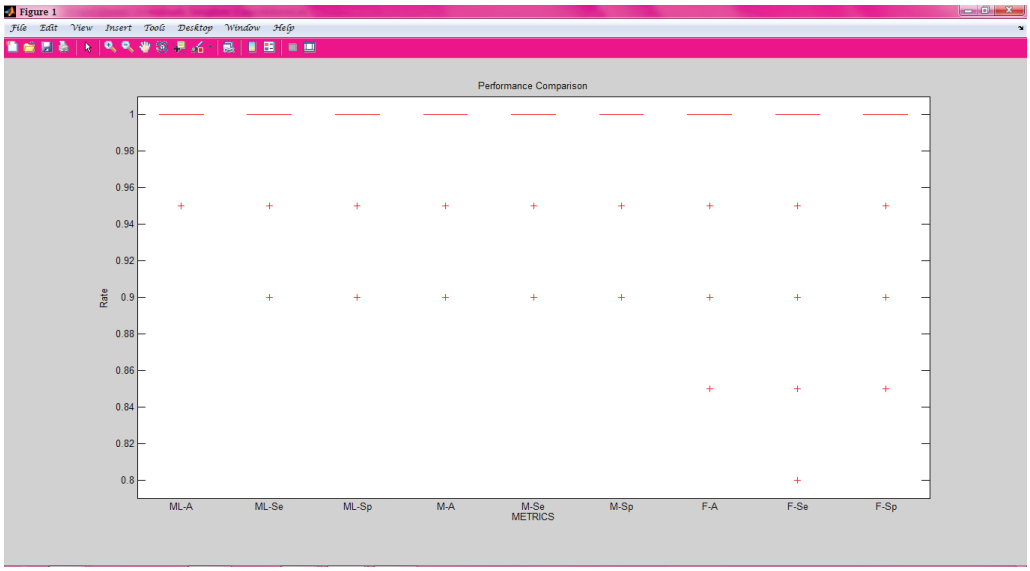

Figure 3. Performance Comparison of the Models.

Hence, the researchers consolidated or merged the data from each word cluster to have a better view of their performance. Consolidating the data from the results of 1 up to 5 syllable words, using the three models, can help us picture the overall performance of each model in each metric as shown in Figure 3.

In the above figure, we can observe that MFCC+LPC performs better than the other models. Although they all have high recognition rates. MFCC+LPC has better accuracy than MFCC. This conforms to the observation of MFCC+LPC in its 4 syllable performance that in only needs 4 syllables to be optimum. Most of the correctly pronounced words were correctly identified

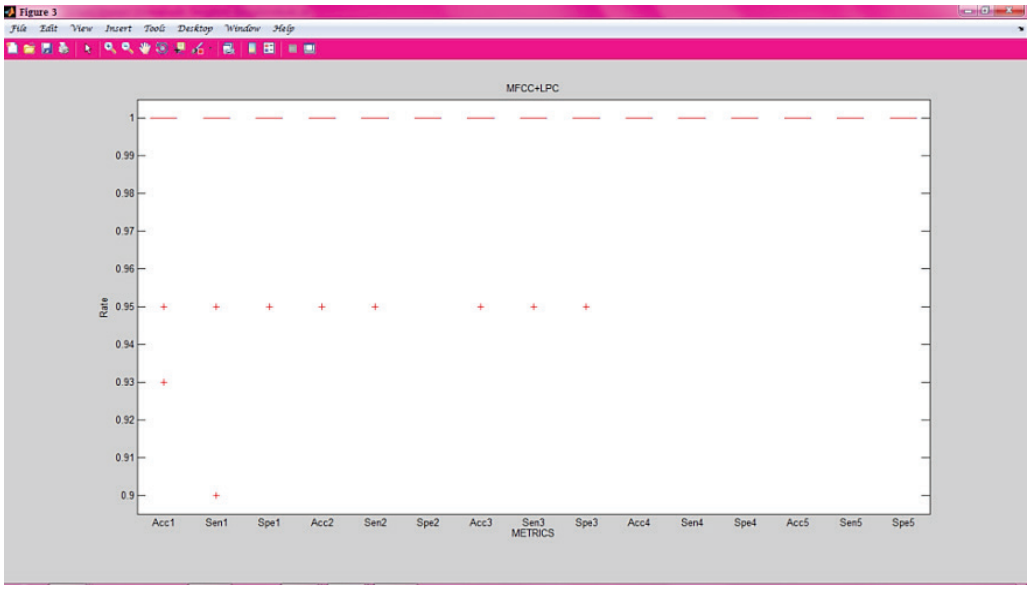

Figure 4. MFCC+LPC Performance.

by the model since most of the data points lie in the 0.9 to 1.0 rates. It can also be observed here that the MFCC+LPC model has the same performance as MFCC in recognizing mispronounced and correctly pronounced words having the recognition rates (specificity and sensitivity respectively) of 0.9 or higher. Though this performance is similar to MFCC, MFCC+LPC is better since it can recognize less syllabled words as its performance is optimum from 4 syllables as shown in figure 4.

The label that the researchers used is the code of the metrics with the number of syllable of the word. *Acc* is for accuracy, *Sen* is for sensitivity and *Spe* is for specificity. It can be observed here that the model improves as the difficulty of the word being recognized increases. This is apparent since MFCC+LPC works best in longer signal samples. Also, it only needs four syllables to be optimum. This behaviour was not observed in the two other models since they need as much as five syllable to be on their optimum performance.

Conclusion and Recommendation

In this research the researchers investigated the problem of recognizing correct pronunciation in English words. Three models were created, with different features namely: MFCC+LPC, MFCC and FFT, using HMM. Results show that the model with the MFCC+LPC feature works best in recognizing correctly pronounced from mispronounced English words. This manifested in the high recognition rates of the model from the three different metrics defined. Also, the model is better in recognizing correct pronunciation on less syllabled words as the model is at its optimum from 4 syllables. This implies that the model has higher recognition rates on less syllabled words thereby making the MFCC+LPC model more stable and more suitable model for pronunciation recognition.

Though the MFCC+LPC has outstanding performance, the model suffers when the input data is continuous. This means that the speech or voice sample may contain one or more words to analyse. In this research, the model was designed to receive single-word voice input hence noise and other sound in the background could be a problem. To make the model more robust, an approach that could extract significant segment of the signal, like extracting only the important words for analysis, is needed to enhance the model. Real-time continuous signal processing may also be a good improvement to make interaction more natural to the users in the educational context.

■ ■ ■

References

Adams, R. E. (1990). Sourcebook of automatic identification and data collection. New York: Van Nostrand Reinhold.

Alsulaiman, M., Muhammad, G., & Ali, Z. (2011). Comparison of voice features for Arabic speech recognition. 2011 Sixth International Conference on Digital Information Management. doi:10.1109/icdim.2011.6093369.

Anand, D., & Meher, P. (n.d.). Combined LPC and MFCC Fetaures based technique for Isolated Speech Recognition. Hyderabad India.

Ananthi, S., & Dhanalakshmi, P. (2013). Speech Recognition System and Isolated Word Recognition based on Hidden Markov Model (HMM) for Hearing Impaired. *International Journal of Computer Applications, 73*.

Anusuya, M., & Katti, S. (2009). Speech Recognition by Machine: A Review. Proceedings of the International Conference on Computer Applications, 6.

Aurora, S., & Singh, R. (2012). Automatic Speech Recognition: A Review.

International Journal of Computer Applications, 60.

Bagge, N., & Donica, C. (2001). Final Project Text Independent Speaker Recognition.", ELEC 301 Signals and Systems Group Projects.

Barbu, T. (2007). A supervised text-independent speaker recognition approach. In Proceedings of the 12th International Conference on Computer, Electrical and Systems Science, and Engineering, CESSE 2007, 22, 444-448.

Barrett, G. (2006). Kale, K., Mehrotra, S., & Manza, R. (n.d.). Computer Vision and Information Technology: Advances and Applications. Retrieved September 15, 2015, from http://www.waywordradio.org/Official_Dictionary_of_Unofficial_English-Grant-Barrett-0071458042.pdf.

Baum, L., Petrie, T., Soules, G., & Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. The Annals of Mathematical Statistics, 41, 164-171.

Boll, S. (n.d.). Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust., Speech, Signal Process, ASSP-27, 113-120.

Daphal, S., & Jagtap, S. (2012). DSP Based Improved Speech Recognition System. International Conference on Communication, Information & Computing Technology (ICCICT).

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*(1), 1-38.

Ernawan, F., Abu, N., & Suryana, N. (2011). Spectrum Analysis of Speech Recognition via discrete TchebichefTransform.Spie.Digital Library. Retrieved September 15, 2015, from http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1197978.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech, and Signal Processing, 29(2), 254-272.

Hispanicallyspeakingnews. (2011). Google's Voice Search Sends Hunter Around the World, Hispanically Speaking News. Retrieved September 4, 2015, from http://www.hispanicallyspeakingnews.com/latino-daily-news/details/googles-voice-search-sends-hunter-around-the-world/9520/

Jarng, S. (2011). HMM Voice Recognition Algorithm Coding. International Conference on Information Science and Applications.

Kale, K., Mehrotra, S., & Manza, R. (n.d.). Computer Vision and Information Technology: Advances and Applications.

Karpagachelvi, S., Arthanari, M., & Sivakumar, M. (2010). ECG Feature Extraction Techniques - A Survey Approach (Vol. 8, Ser. 1).

Krenker, A., Bester, J., & Kos, A. (n.d.). Introduction to the Artificial Neural Networks. Slovenia: University of Ljubljana.

Li, J. et. al. (n.d.) . An Overview of Noise-Robust Automatic Speech Recognition. Retrieved September 15, 2015, from https://www.lsv.uni-saarland.de/fileadmin/publications/non_articles/an_over-

view_of_noise_robust_automatic_speech.pdf.

Ludwig, S. (n.d.). Siri Assistant 1.0 for iPhone. Retrieved September 4, 2015, from http://www.pcmag.com/article2/0,2817,2358823,00.asp.

Lyu, M., Xiong, C., & Zhang, Q. (2014). Electromyography (EMG)-based Chinese voice command recognition. 2014 IEEE International Conference on Information and Automation (ICIA). doi:10.1109/icinfa.2014.6932784.

Mastin, L. (2011). Language Issues: English as a Global Language. Retrieved September 15, 2015, from http://www.thehistoryofenglish.com/issues_global.html.

Maurice. (2013). 5 Ways to get Siri Alternatives for Android Phones. Retrieved September 4, 2015, from http://www.tipsotricks.com/2013/03/5-best-siri-alternatives-for-android-phones.html.

Mermelstein, D. (1980). Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE T. Acoust., Speech Signal P., *28*(4), 357-366.

Muzaffar, F., Mohsin, B., Naz, F., & Jawed, F. (2005). DSP Implementation of Voice Recognition Using Dynamic Time Warping Algorithm. 2005 Student Conference on Engineering Sciences and Technology, 1.

Nitta, T., Murata, T., Tsuboi, H., Takeda, K., Kawada, T., & Watanabe, S. (n.d.). Development of Japanese voice-activated word processor using isolated monosyllable recognition. ICASSP 82. IEEE International Conference on Acoustics, Speech, and Signal Processing. doi:10.1109/icassp.1982.1171875.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition.

Sak, H. et. al. (2015). Google Voice Search: Faster and More Accurate. Retrieved September 15, 2015, from http://googleresearch.blogspot.com/2015/09/google-voice-search-faster-and-more.html.

Sarikaya, R., & Hansen, J. (2001). Analysis of the root-cepstrum for acoustic modeling and fast decoding in speech recognition. Proc. Eurospeech'01. Aalborg, Denmark.

Sapijaszko, V., & Michael, W. (2012). An overview of recent window based feature extraction algorithms for speaker recognition. IEEE 55th International Midwest Symposium on Circuits and Systems (MWSCAS).

Shi, Z., Shimohara, K., & Feng, D. (2006). Intelligent information processing III. IFIP TC12 International Conference on Intelligent Information Processing (IIP 2006).

Speech. (n.d.). Retrieved September 15, 2015, from https://www.uic.edu/classes/ece/ece434/chapter_file/Chapter5_files/Speech.htm.

Thakare, V. (n.d.). Techniques for Feature Extraction In Speech Recognition System : A Comparative Study. Retrieved September 4, 2015, from http://arxiv.org/abs/1305.1145#.

Varshney, N. (2014). Embedded Speech Recognition System. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 3.

Viikki, O., & Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Communication, *25*(1-3), 133-147.

Vimala, C., & Radha, V. (2012). A Review on Speech Recognition Challenges and Approaches. *World of Computer Science and Information Technology Journal (WCSIT), 2*(1), 1-7.