# Misarticulated /r/ - Speech Corpus and Automatic Recognition Technique

**Suresh Kumar Nagaram  Suman Maloji  KasiprasadMannepalli**

***Abstract:*** ***Technique to recognize the impaired pronunciation of sound /r/ from Telugu speech signals is presented in this paper besides the speech corpus. Rhotacism is called as an inability to pronounce the sound /r/ and is one of the Speech Sound Disorders (SSD) in children. Whose SSD not diagnosed at an early stage may result ina lack of social skills. This demands an efficient automatic speech impairment detection technique, which helps the therapists to treat the patients with impairment specific procedure. Databases for the impaired articulation of /r/ in various languages are explored in this article. The shape of the envelope, timbre, Walsh Hadamard Transform (WHT), Discreet Cosine Transform (DCT) features extracted, from the Mel-Frequency Cepstral Coefficients (MFCC), to discriminate the correct and wrong articulation of /r/ are detailed. Usage of k-Nearest Neighbor (kNN), Support Vector Machine (SVM) and kohonen neural networks in various articles, for classification, are briefed. MFCC features and k-NN algorithm is used to identify the misarticulation in the Telugu language. The 80.1% classification accuracy shows that the proposed method performs good with respect to the methods detailed for other languages. Availability of acoustic databases for the impaired articulation of /r/ and subjects with such impairment restricts the performance validation of the investigated methods. This further demand the more contribution from scholars in the development of automatic techniques and databases for misarticulated /r/ in different languages.***

***Index Terms:*** ***Speech Sound Disorder, Rhotacism, Impaired Articulation, Impaired Speech, Dyslalia.***

## I. INTRODUCTION

In the era of the fully grown and technology driven world, the automatic detection of impaired speech or SSD is not fully addressed. SSD is the inability to pronounce a letter or letters. The sources for this disorder are not fully known yet [1]. There are mainly two categories in SSD, those are Articulation disorder and Phonemic disorder [2, 3]. The former one is due to the difficulty in learning of, how these phonemes are physically produced. This disorder completely deals with the main articulators. Later one is due to the difficulty in learning and understanding the language's sound system. Diagnosis of this speech impairment at the early stages of childhood may help them to recover from this problem. Customary speech impairments are due to the improper articulation of sounds, such as /r/ (Rhotacism), /s/ & /z/ (Sigmatism) etc. Various remediations for SSD, articulation disorder (Rhotacism) in particular, are reported in the literature. Visual feedback of real time speech spectrogram to the patient [4, 5], usage of removable R-appliances [6], Ultra sound [7, 8], hand gesture cues [9] and Electrography [10] modes of therapy for misarticulation of /r/ is reported. Remediation of improper articulation of /r/ using traditional and spectral bio feedback is reckoned in [11]. With the help of speech therapy classes and continuous practice, one can recover from SSD. The practice of speech therapy mainly relies on the methodology of identification followed by the correction of improper sounds, making it a tedious process. During the diagnosis, the therapist has to spend an ample amount of time to identify and categorize the misarticulated sound. An investigation of relation among the acoustical, ultra sound and perceptual routines to categorize the good and bad articulation of /r/ is detailed in [12].

The detection and diagnosis of the SSD vary from language to language. There is very less, or no research is happening on SSD in the dialect of different languages other than English. The works on the impaired speech by the various researchers, the impaired speech database they developed and the techniques they recommended for automatic identification of impaired articulation of sound /r/ is presented in section 2. Section 3 details the proposed method to detect the impaired articulation /r/ sound in Telugu language and section 4 concludes the paper.

## II. INVESTIGATION OF LITERATURE

The automatic detection & databases for misarticulated /r/ sound developed by various researchers are detailed in this section. Ovidiu Grigore et al. [13] presented a technique, which identifies the impaired pronunciation of /r/ in the Romanian language automatically, using variations in the timbre during the evaluated duration, by MFC coefficients and there by performing k-NN classification algorithm on extracted feature i.e., Timbre. The database used in this technique contains the voice recordings of 8 female and 3 male adults, among them 3 female and 1 male have the improper pronunciation of sound /r/. From a group of ten words, each spokesman utters each word three times. All voice recordings are preprocessed initially. During this stage, the initial phoneme /r/ is segmented manually from the word. The words used to create the database contains initial consonant /r/ followed by a vowel. This particular choice in word selection makes manual segmentation easy. The quality of manual segmentation is acceptable, because of the vowel

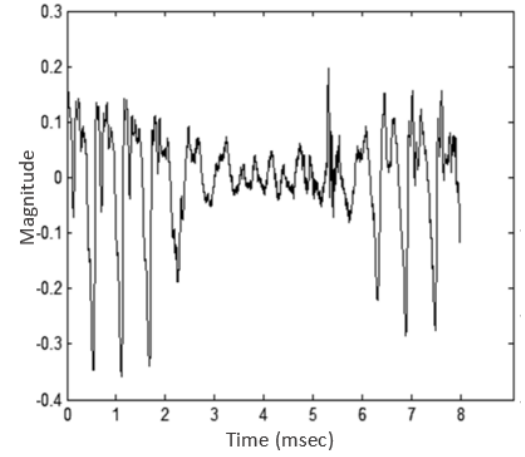
**Revised Manuscript Received on March 20, 2019.**

**Suresh Kumar Nagaram**, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation,Vaddeswaram, Guntur, Andhra Pradesh 522502, India.

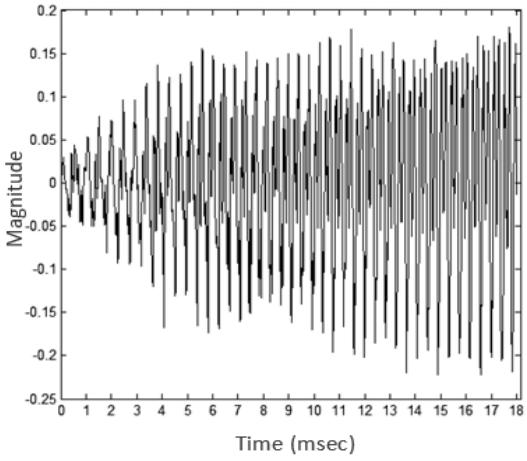
**Suman Maloji**, Professor, Department of Electronics and Computers Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh 522502, India.

**KasiprasadMannepalli**, Associate Professor, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh 522502, India.

after the improper sound /r/. After the preprocessing, the segmented portion of interest is processed to extract the feature. The timbre feature of the speech over the duration of the pronounced phoneme /r/ is extracted from the MFC coefficients. The envelope shape of the speech signal is a feature, that can be used to differentiate the improper with proper pronunciation if, the degree of improper pronunciation is high. The envelope shape of the proper pronunciation (shown in figure 1(a)) resembles the linear modulation (i.e., Amplitude modulation), where as in later case the envelope shape is more constant as shown in the below figure 1(b). This result is due to the replacement of sound /r/ by the other sounds /d/, /t/, /1/ etc. in improper pronunciation and they will have amplitudes varying very slowly. If the degree of improper pronunciation is very less, then the sound /r/ is guttural, making the shape of the envelope similar to the proper pronunciation. This makes the envelope shape feature no longer a good choice.




(a)




(b)

Figure 1. The waveforms for a segmented portion of /r/ from the word "rac" [13]. (a). Proper pronunciation (b). Improper pronunciation Therefore, the author choses Timbre, a spectrum-based feature to characterize the speech. A homologous MFC coefficients standard deviation is used to extract this feature. Later, the feature vector is passed through the kNN classification algorithm with different values of 'k'. It is identified that the performance of the classifier varies drastically based on the vowel after the consonant /r/ in the word. The classifier accuracy (correct classification of proper and improper pronunciation), for the words which contain the vowel 'a' after the /r/, is 90% for k being 3 to 5 and it is 83% for the vowel 'o' after the sound /r/ with 'k' is in-between 7 to 11. In the work of Ovidiu Grigore [14], detection of the extremely effected pronunciation of sound /r/ in Romanian language using kohonen neural network was reckoned. Phoneme /r/ is the most commonly mis-pronounced sound in the children. Enormous vowel detection algorithms are developed, as they are easy to detect, because of their high energy concentration. A little work was done to detect the consonants and its features. Whereas almost no work was done to detect the improper pronounced consonants. Because it is more difficult than detecting properly pronounced consonants. A considerable effort was done by the author to detect the improper pronunciation of consonant /r/ using neural networks. Words with initial /r/ phoneme followed by a vowel 'a' (rac, rana, rama etc.) were uttered by 15 children and 5 adults for database creation. Among them, only twelve children were suffering from rhoticism. The envelope shape of the signal is considered as the feature and is calculated by taking the normalized amplitude of the signal over a small duration. This process is repeated until the duration of interest is covered. Later these normalized amplitudes are interpolated and then its mean is calculated to form the feature vector space with a reduced dimension. This feature vector is processed through the prominent data clustering algorithm called kohonen neural networks or self-organizing maps. With the help of this clustering algorithm, the close relation between the different test samples and the groups of the speech samples relative to their correctness is developed. The size of the input neurons is the length of feature vector space and the output neurons are3*3 array. From the results, it can be observed that more improper pronunciations are accumulated on a single output neuron and proper pronunciations are accumulated on various output neurons. The accuracy of the developed kohonen neural network classifier is 82.5%. Valentin Velican, et al [15] developed an automatic system to detect improperly pronounced initial /r/ consonant in Romanian. The author collected the acoustic information from "Logopedic Interscolar Centre – Bucharestto, Romania" to develop the database. The database contains the voice recordings of the words, where /r/ is initial phoneme in the word such as "rac", "rim" etc. All the words used for recording, contain initial /r/ phoneme followed by a vowel. The choice of initial phoneme followed by a vowel in the word makes the manual segmentation of distorted sound /r/ from the word easy. The database contains recordings of 30 children within the age group of 6 to 8 years, suffering from pronunciation problem with the sound /r/, considered as the first set and the second set contains the recordings of 10 children, of same age group, who pronounce the sound /r/ correctly. The final database contains 44 voice recordings of properly pronounced and 100 voice recordings of improperly pronounced words as set-2 and set-1. Because of the age group, the resemblance of the pitch from child to child is more along with timbre. MFC coefficient analysis was performed on the two sets of databases, to extract the features required for classification. Two classes, proper and improper

pronunciation of the phoneme are the output of the classifier. Out of 256 unique cepstral coefficients of 512-point MFCC analysis, coefficients 5 to 170 (166 coefficients) were chosen for the further step of feature selection. As the phoneme /r/ contains few vocalized excitations, it is required to consider the higher coefficients along with the lower frequency coefficients for better classification. DCT and WHT are applied to the selected coefficients to extract the features. On the spectrum of the selected phoneme, mean was computed on every set. It is observed form the plot of means of two classes given in figure 2, that the coefficients spanning between 5 to 16 clearly discriminates the magnitudes of the mean, between two classes of consideration. The response of mean of all feature vectors of DCT is also similar.

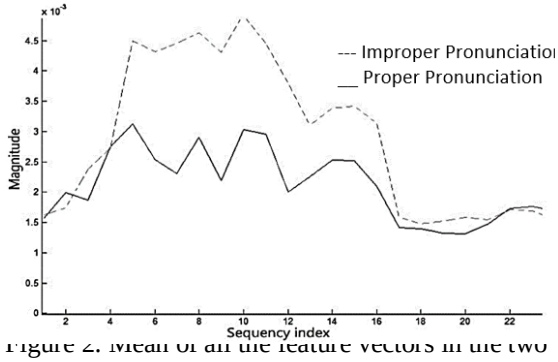



Figure 2. Mean of all the feature vectors in the two classes calculated from intermediate feature space [15].

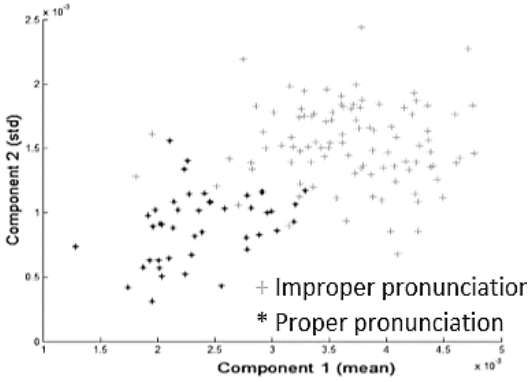



Figure 3. Correct and incorrect pronunciations represented in the final feature space obtained by performing the mean and standard deviation of the selected coefficients out of the WHT [15].

The standard deviation of the WHT & DCT feature vector is determined to make the feature vector space is of two dimensional. The plot of mean Vs standard deviation of WHT feature vector (shown in figure 3) is confirmed that these two features "mean & standard deviation" of spectrum coefficients clearly separates the improper pronunciation from proper pronunciation. The regression line between the two classes is clear without any ambiguities. kNN algorithm is used for classification. Where k varies from 3 to 17 with a step size of 2. 10 out of 44 correct and 25 out of 100 incorrect pronunciations was used for training and remaining recordings of speech were used for testing respectively. The experiment was executed 500 times on each k, because of the small database. Highest correct classification rate for DCT is 87.63% at k=13 and 92.55% at k=3 for WHT. The author concludes that the developed algorithm with the help of WHT and kNN can automatically detect the impairment in child speech and also saves the speech therapy specialist time. Hammami, Nacereddine, et al [16] prepared a database for the automatic diagnosis of children's speech disorders in the Arabic language. The author mainly concentrated on letter /r/, as it is felt difficult to pronounce by most of the children. In disordered pronunciation, the instances of sound /r/ in a word, may be distorted or omitted (completely removed) or substituted (The instance of /r/ sound is substituted with another sound) or added (other sounds are added along with the sound /r/). Children suffering from such a disorder is rare, making it very difficult to find them and register them for speech recordings. So, the author with the help of speech therapy specialists worked on how to pronounce a word with the disorder. Later this disordered pronunciation was simulated with 30 male and 30 female healthy children who do not have any speech disorders. These simulated disordered voices were used to create the database. The database comprises the words uttered by healthy children with the simulated disorder of pronouncing the sound /r/. The selection of words used to prepare the database is chosen such that, the letter /r/ is, at the starting for some words, in the middle for some words and at the end for some words. Based on the position of the letter /r/ in the word, the database was divided in to three sets. Further, each set is divided in to four subsets based on, whether the sound /r/ in the word is distorted or omitted or substituted or added. The example simulated recommendations for the disordered pronunciation of words, in different above-mentioned cases, from the specialist is given in table 1. Each speaker utters the word 5 times. So, the database contains 300 voice files (30 male and 30 female speakers * 5 times).

Table 1 Approved disordered pronunciation of words by specialists.

| Disordered phoneme | /r/ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Disorder place in a word | Beginning | | | | Middle | | | End | | |
| Disorder Type | Correct Pronunciation | Substitution / Distortion | Omission | Addition | Correct Pronunciation | Substitution / Distortion | Addition | Correct Pronunciation | Substitution / Distortion | Omission |
| Specialist point of view about the pronunciation | Rajul | Ghajul Lajul | Jul | Rrrjul | Mariyam | Maghyam Malayam | Marrrriyam | Kabeer | Kabeegh Kabeel | Kabee |

The author only proposes the database for disordered sound /r/. Development of the database for other characters is considered as future work. Automatic detection and classification of voice pathologies using features in different frequency bands were developed by Ahmed Al-nasher et.al. in [17]. The Paralysis, Polyps, and Cysts of vocal folds are the voice pathologies considered. These pathological speech samples are collected from the most familiar voice pathological databases, i.e., Massachusetts Eye and Ear Infirmary (MEEI) database [18], Saarbrücken Voice Database (SVD) [19] and Arabic Voice Pathology Database (AVPD) [20]. The number of speech samples considered to test the method is 71, 20, 10 from MEEI, 212, 45, 6 from SVD and 56, 46, 25 from AVPD respectively for the pathologies mentioned above and 53, 266, 169 number of normal speech samples (no disorder) from each database respectively. These are the sustained vowel /a/ samples. The selection of few samples from a large pool is to maintain the prevalence among the considered databases. To assess and classify the type of voice disorder, maximum peak values, lag values of the peak and entropy features of the pathological and correct speech samples are evaluated. Along with these features, author scrutinized how difference frequency bands effect the detection and classification accuracy of the system. Support vector machine (SVM) algorithm is used for classification of pathology. The accuracy achieved for detection of pathology is 99.69%, 92.79% and 99.79% for the speech samples of MEEI, SVD, and AVPD databases respectively. 99.54%, 99.53%, and 96.02% are the classification accuracy for these databases respectively. The performance of the detector and classifier is less for individual frequency bands in the range of 1KHz to 8KHz. Where the technique achieved the best detection and classification accuracy in collective frequency bands.

## III. MFCC BASED DETECTION TECHNIQUE

Small set of speech database is created with 6 persons among them 3 are suffering from misarticulation of sound /r/. Each person is requested to speak a set of 10 words repeatedly 10 times. The words contain the /r/ in the initial position of word. All persons are native Telugu speakers.

MFCC is the most promising feature in the detection of speaker and speech. The effectiveness of this feature in the detection of pathological speech i.e, impaired articulation of sound /r/ is described here. MFCC is a method where the source and the excitation parameters of the speech can be divided with a clear picture. The MFC coefficients extraction steps are detailed below:

a. Speech is considered as a periodic signal for a short duration of time, say few milli seconds. So, the speech signal is windowed first with Hamming window of 30ms duration with 15msec overlap.

b. Periodogram of the windowed signal is estimated

$$P(k) = \left|\sum_{n=0}^{N-1} w(n)s(n)e^{\frac{-j2\pi nk}{N}}\right| \quad (1)$$

Where *s(n)* is input speech signal, *w(n)*is Hamming window and *N* is the number of points in FFT calculation.

c. Periodogram is mapped on to the overlapped mel scale

$$P(m) = \sum_{k=0}^{N-1} P(k)H(k,m) \quad (2)$$

Where *H(k,m)* is mel filter bank.

d. Take the log of the eq(2)

$$\hat{P}(m) = \log\left(P(m)\right) \quad (3)$$

e. Apply the Discreet Cosine Transform (DCT) on $\hat{P}(m)$ we get the MFC coefficients of length *L*

$$MFCC(l) = \sum_{m=1}^{m} \hat{P}(m)\cos\left(l\frac{\pi}{m}(m-0.5)\right) \quad (4)$$

Where $l$ =1, 2,…,L.

The wave form of misarticulated /r/ (ర) in Telugu language and its corresponding MFC co-efficient are shown in figure.4(a) & (b) respectively. In contrary the wave form and its corresponding MFC coefficients for properly articulated /r/ (ర) in Teluguis plotted in figure.5 (a) & (b) respectively. In the figure.4(b) and figure.5(b) we can observe that there is a distinct behavior in MFC coefficients from the 5$^{th}$ sample to 10$^{th}$ sample. The mean and standard deviation of the MFC coefficients are calculated. These features are applied to the k-NN classifier for classification. The classifier accuracy for various values of k is plotted in the figure 6.From the figure we can observe that the maximum Classification accuracy is 80.1% at k=14. This classification accuracy is good when compared to the other methods proposed to detect this impairment in other languages. The compression between the existing techniques and the proposed is given in table 2.

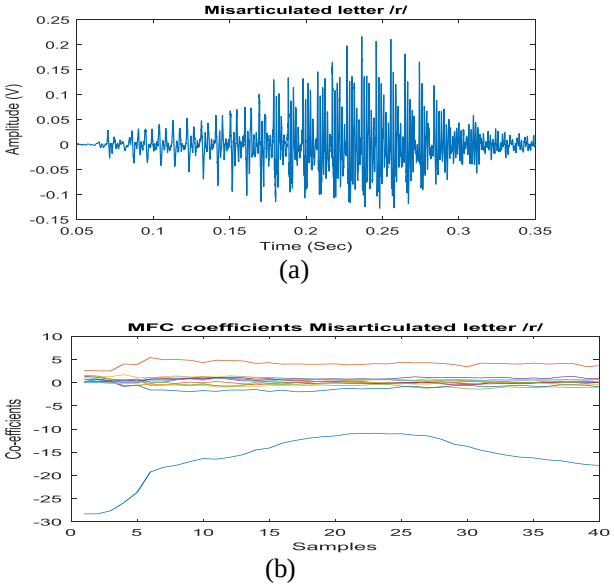



Figure 4. Misarticulated /r/ in Telugu (ఠ) a. Speech signal b. Corresponding MFC coefficients.

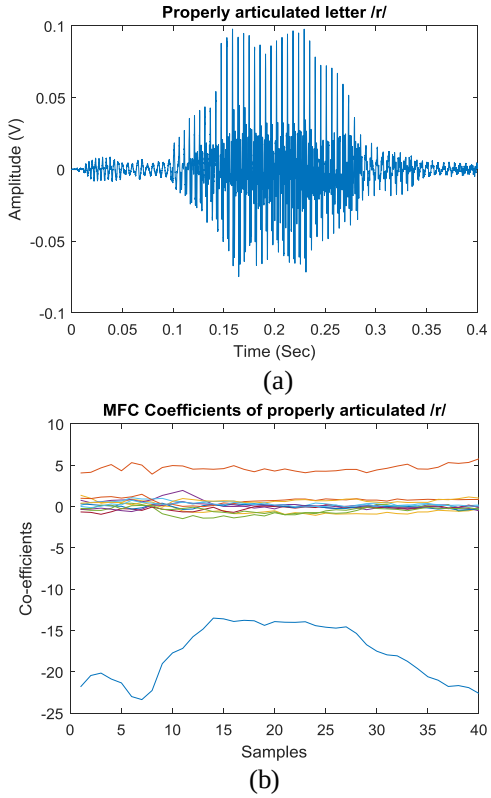



Figure 5. Properly articulated /r/ in Telugu (ఠ) (a). Speech signal (b). Corresponding MFC coefficients.

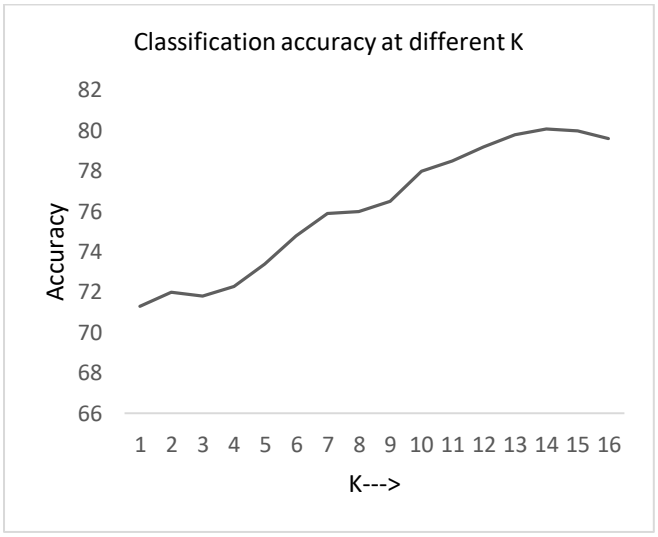



Figure 6. The accuracy of the classifier at different k

Table.2 Classification accuracy of the system at different k

| S. No | Method | Accuracy (%) |
|---|---|---|
| 1 | K-NN [13] | 90 |
| 2 | Kohonen SOM [14] | 82.5 |
| 3 | K-NN [15] | 92.5 |
| 4 | Proposed | 80.1 |

## IV. CONCLUSION

Misarticulation is one of the major speech impairment in children. This impairment is because of the unattended speech and language learning in the early stages of a child. Rhotacism is one most communal speech impairment in children in every language. Earlier detection of this impairment may help the children to recover. The residuals of Rhotacism can also be observed in adults. Unavailability of automatic detection of speech impairment makes it difficult for speech therapists to identify the type of impairment and thereby diagnosis. Automatic detection of misarticulated /r/ form the spoken words in the Telugu language is presented. MFCC features of the speech clearly differentiate the impaired speech with a normal one. The classifier used in this works achieved 80.1% classification accuracy. An automatic segmentation method may improve the accuracy of the developed methods.An improved method may require to perceive the impairment from the continuous and spontaneous speech. There is a necessity for the techniques to be developed to detect impaired /r/ sound form dialects of Indian languages.

## REFERENCES

1. American Speech-Language-Hearing Association. "Speech sound disorders: Articulation and phonological processes." American Speech Language Hearing Association. Retrieved March 17 (2014): 2014.
2. American Psychiatric Association. "Speech Sound Disorder, 315.39 (F80.0)". Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. American Psychiatric Pub, 2013. pp. 44–45.
3. V Saounatsou, "Children's Speech Sound Disorders," International Journal of Language & Communication Disorders, vol. 45, no. 6, pp. 706-706, 2010.

4. Shuster, Linda I., Dennis M. Ruscello, and Kimberly D. Smith. "Evoking [r] using visual feedback." American Journal of Speech-Language Pathology 1.3 (1992): 29-34.
5. Shuster, Linda I., Dennis M. Ruscello, and Amy R. Toth. "The use of visual feedback to elicit correct/r." American Journal of Speech-Language Pathology 4.2 (1995): 37-44.
6. Clark, Charlene E., Ilsa E. Schwarz, and Robert W. Blakeley. "The removable r-appliance as a practice device to facilitate correct production of/r." American Journal of Speech-Language Pathology 2.1 (1993): 84-92.
7. Adler-Bock, Marcy, et al. "The use of ultrasound in remediation of North American English/r/in 2 adolescents." American Journal of Speech-Language Pathology 16.2 (2007): 128-139.
8. Byun, Tara McAllister, Elaine R. Hitchcock, and Michelle T. Swartz. "Retroflex versus bunched in treatment for rhotic misarticulation: Evidence from ultrasound biofeedback intervention." Journal of Speech, Language, and Hearing Research 57.6 (2014): 2116-2130.
9. Rusiewicz, Heather Leavy, and Jessica Lynch Rivera. "The effect of hand gesture cues within the treatment of/r/for a college-aged adult with persisting childhood apraxia of speech." American journal of speech-language pathology 26.4 (2017): 1236-1243.
10. Hitchcock, Elaine R., et al. "Efficacy of electropalatography for treating misarticulation of /r/." American journal of speech-language pathology 26.4 (2017): 1141-1158.
11. Byun, Tara McAllister, and Elaine R. Hitchcock. "Investigating the use of traditional and spectral biofeedback approaches to intervention for/r/misarticulation." American Journal of Speech-Language Pathology 21.3 (2012): 207-221.
12. Klein, Harriet B., et al. "A multidimensional investigation of children's /r/ productions: Perceptual, ultrasound, and acoustic measures." American Journal of Speech-Language Pathology 22.3 (2013): 540-553.
13. Grigore, O., C. Grigore, and V. Velican. "Impaired speech evaluation using mel-cepstrum analysis." International Journal of Circuits, Systems and Signal Processing (1998): 70-77.
14. Grigore, Ovidiu, Valentin Velican, and I. Gavat. "Self-organizing maps for identifying impaired speech." Advances in electrical and computer engineering 11.3 (2011): 41-48.
15. Velican, Valentin, RodicaStrungaru, and Ovidiu Grigore. "Automatic recognition of improperly pronounced initial 'r' consonant in Romanian." Advances in Electrical and Computer Engineering 12.3 (2012): 80-84. s
16. Hammami, Nacereddine, et al. "/r/-Letter disorder diagnosis (/r/-LDD): Arabic speech database development for automatic diagnosis of childhood speech disorders (Case study)." 2015 Intelligent Systems and Computer Vision (ISCV). IEEE, 2015.
17. Al-Nasheri, Ahmed, et al. "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions." IEEE Access 6 (2018): 6961-6974.
18. Kay Elemetrics Corp., Disordered Voice Database, Version 1.03 (CD-ROM), MEEI, Voice and Speech Lab, Boston, MA (October 1994).
19. Barry, W. J., and M. Putzer. "Saarbrucken Voice Database, Institute of Phonetics, University of Saarland." (2007).
20. Mesallam, Tamer A., et al. "Development of the Arabic voice pathology database and its evaluation by using speech features and machine learning algorithms." Journal of healthcare engineering 2017 (2017).
21. Mannepalli, Kasiprasad, PanyamNarahari Sastry, and Maloji Suman. "MFCC-GMM based accent recognition system for Telugu speech signals." International Journal of Speech Technology 19.1 (2016): 87-93.
22. Mannepalli, Kasiprasad, PanyamNarahari Sastry, and Maloji Suman. "A novel adaptive fractional deep belief networks for speaker emotion recognition." Alexandria Engineering Journal (2016)


## AUTHORS PROFILE

**Mr. Suresh Kumar Nagaram** received his Bachelor's degree in Electronics and Communication Engineering from JNTU-Kakinada and Master's degree in Communications and Signal Processing fromAcharya Nagarjuna University. He is currently working towards his Doctoral degree in the area of Speech signal processing. His research interests are Signal, Image and Speech processing.

**Dr. Suman Maloji**is a professor in the Department of Electronics and Computers Engineering. He published many articles with international and national journals. His research interests are Speech coding, Speech compression, speech,and speaker recognition.

**Dr. KasiprasadMannepalli,** Associate Professor of Electronics and Communication Engineering department. He received his Doctoral degree in the field of speech signal processing from Koneru Lakshmaiah Education Foundation. Currently, he is guiding four members for their Doctoral degree. His research interests are Emotional speech recognition, Accent recognition,and pathological speech processing.