CSIT (June 2014) 2(2):77–84
DOI 10.1007/s40012-014-0044-6

ORIGINAL ARTICLE

Performance enhancement of syllable based Tamil speech recognition system using time normalization and rate of speech

A. Akila · E. Chandra

Received: 14 May 2014 / Accepted: 20 June 2014 / Published online: 1 August 2014

Abstract The automatic speech recognition (ASR) is an active field of research. The performance of the ASR can be degraded due to various features like environmental noise, channel distortion and speech rate variability. The speech rate variability is one of the important features that affect the accuracy of the speech recognition system (SRS). In this research work, the speech signal is categorized as slow, normal and fast speech using features like the sound intensity level, time duration and root mean square. This paper addresses the enhancement of the performance of a SRS by applying time normalization to the speech signal. The comparison of the proposed Model and baseline syllable based SRS is done.

Keywords Syllable based SRS · Rate of speech · Time normalization · Sound intensity level · Time duration · Root mean square

1 Introduction

In automatic speech recognition (ASR), the input speech is contaminated by a background noise. Noise has two main effects over the speech representation. It introduces a distortion in the representation space and it also causes a loss of information, due to its random nature. The distortion of the representation space due to the noise causes a mismatch between the training (clean) and recognition (noisy) conditions. The acoustic models, trained with speech acquired under clean conditions do not model speech acquired under noisy conditions accurately and this degrades the performance of speech recognizers. The information loss caused by noise introduces degradation even in the case of optimal mismatch compensation. Speech rate is one of the variability influencing ASR [1]. The variations in speaking rate can account for a significant percentage of errors in practical speech recognition task [2]. The timing of speech within a discourse is known to vary both globally and locally due to various factors such as stress, emotion etc. The speaking rate can be measured using some criteria's like syllable or phoneme rate computed from phonetic transcription. This is not suitable for online application. The speaking rate is variable according to speaker and task, always there exists the mismatch between the speaking rates of the training and test data.

Over recent years, efforts have increased to improve ASR by making use of prosodic cues. Speech rate is one of the most important cues because it modifies acoustic cues, phones and even words. The number of speech units like words, syllables or phones per second is used to calculate the speech rate. Speech rate is often referred as speaking rate or articulation rate. Variations in rate of speech (ROS) produce change in both spectral features and word pronunciations that affect ASR system. ROS is an important factor that affects the performance of a transcription system. Possible reasons are that some features commonly used in recognition systems are duration related and clearly influenced by speech rate such as delta and delta–delta features, and that some pronunciation phenomena such as coarticulation and reduction are also speech rate related.

Two methods are typically used to estimate the ROS of an input utterance. One is based on phone durations which

A. Akila (✉)
Department of Computer Science, D.J Academy for Managerial Excellence, Coimbatore, India
e-mail: akila.ganesh.a@gmail.com

E. Chandra
Department of Computer Science, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore 641049, India
e-mail: crcspeech@gmail.com

are often obtained from phone level segmentations by using forced alignments. When the utterance transcription is known, this duration based method can provide robust ROS estimation. When the transcription is unknown [3], we can only use the hypothesis from a prior recognition run whose quality is hard to guarantee. The second method involves estimating ROS directly from the waveform or acoustic features of the input utterance.

Absolute ROS measures such as phones per second and inverse mean duration were often used but are not informative enough since they do not consider the fact that different types of phones have different duration distribution. Speaking too fast or having the speech signal cut off can severely affect speech intelligibility over communicating devices. Speaking too fast can often make the speech incomprehensible and the problem lies in the speaker's unawareness of how quickly they are speaking and whether or not there is need to slow down.

In this paper method for enhancing the performance of a syllable based Tamil speech recognizer is proposed using time normalization and ROS. In Sect. 2, the literature review of the speech rate related works and the parameters of sound are discussed. In Sect. 3, a brief description about the phases of the proposed model is given. In Sect. 3, the measures used for speech rate category identification is discussed. The process for ROS manipulation used to compute the scaling factor for time normalization is discussed. The time normalization of the speech signal depending on the ROS is discussed in Sect. 4. The process of the baseline model syllable based speech recognizer is discussed in Sect. 5. The performance enhancement of the proposed system over the baseline model is discussed in Sect. 6.

2 Rate of speech

Global speech rate [4] is found by dividing the number of segments by the sum of their durations for a complete utterance. Local speech rate is the prosodic feature and allows drawing conclusion on, which constituents in a given sentence are stressed and therefore being articulated with a slower rate or which constituents have a higher speech rate and as a consequence show more assimilation. The local speech rate is estimated by moving average.

2.1 Related work

Falthauser et al. [3], has done a work on speech rate estimation in which the speaking rate is identified using Gaussian mixture model. The GMM is usually used for speaker identification by comparing the test data with the models created for the training data. The same methodology is used in this work to find the speech rate category. The categories considered in this work were slow, medium and fast. The training data was segmented into spurts and for each spurts with their phonetic transcription, the phoneme rate was calculated. The spurts with a speech rate exceeding a threshold value (TH) were marked as fast, the spurts with a speech rate below the symmetric threshold were considered slow and the remaining spurts were marked as medium. The system also allows for gender detection.

The speech rate of a speech uttered by a speaker can vary on every utterance of the same word. The recognition accuracy may be affected to a certain extent because of variation in speech rate. So the speech rate has to be normalized. In the work done by Jin Ah Kang [1], the normalization was done using time scale modification. The speaking rate was controlled by varying a scaling factor of the time scale modification. The best scaling factor was selected for training data based on the maximum likelihood criteria.

The speech recognition system was developed by Sung Min Ban [2] with speaking rate dependent multiple acoustic models. The optimal acoustic model was selected by comparing the speech rate of the training data and test data and then used in recognition. A variable frame shift size was considered instead of flat frame shift size. The frame shift size was controlled by applying the continuous frame rate normalization to each training data.

The model developed by Jing Zheng et al. [5] used parallel rate specific acoustic models where one model for fast speech and another for slow speech. As parallel structure of rate specific models is used, the high quality ROS estimation before recognition was not needed. A new method was also explored to model rate dependent pronunciation variation. They also designed an approach to combine rate specific acoustic models at the pronunciation level [6] and the best matching model is selected by the recognizer based on the maximum likelihood criterion during decoding.

A ROS estimator directly derived from acoustic signals was designed by Nelson Morgan et al. [7]. They have shown that a simple modification of the model transition probabilities based on the estimated ROS can reduce the error rate more than the lexical phone rate method.

Hiroaki Nanjo and Tatsuya Kawahara [8] addressed the problem of speaking rate in large vocabulary spontaneous speech recognition. In spontaneous lecture speech, the speaking rate is generally fast and may vary a lot within a conversation. The work investigated several methods and the application where each method is suitable for is identified to improve the accuracy of the recognition process.

If the speech rate is more than the normal speech rate, the rate of recognition errors will increase in large

vocabulary ASR system. Mathew Siegler and Richard M.Stern [9] proposed a work in which the errors due to fast speech was identified and corrected using the methods based on Baum Welch codebook adaptation, HMM state transition probability and rule based techniques.

2.2 Parameters of sound

(1) *Energy* Sound is the movement of energy through substances in longitudinal waves. Sound is produced when a force causes an object or substance to vibrate. The energy is transferred through the substance in a wave. The energy in sound is far less than other forms of energy. Sound vibrations create sound waves which move through mediums such as air and water before reaching the hearer's ears.
Sound energy is usually measured by its pressure and intensity in units called Pascals and decibels. Figure 1 specifies the high energy and a low energy signal.

(2) *Power* Sound power is a measure of sound energy per time t unit. It is measured in watts. Sound power is neither room dependent nor distance dependent. It belongs strictly to sound source. The threshold of hearing is 10^{-12} w/m^2. In this work, the energy is calculated for each frame using the Eq. 1. The time duration is manipulated using the Eq. 2. The power of the sound signal is calculated using Eq. 3.

$$Energy = \sum_{n=1}^{N} \left| s(n)^2 \right| \tag{1}$$

where s(n) is the sample data of each frame and n is the current frame of the N frames in the signal.

$$Time = \left(len * \left(\frac{1}{fs} \right) \right) - \left(\frac{1}{fs} \right) \tag{2}$$

where len is the length of the sample data of the signal and fs is the sample rate of the signal.

$$Power = \frac{Energy}{Time} \tag{3}$$

(3) *Intensity* Sound intensity is also referred as acoustic intensity. It is defined as the sound power per unit area. Sound intensity is not the same physical quantity as sound pressure. Hearing is directly sensitive to sound pressure which is related to sound intensity. Sound intensity is a specifically defined quantity and cannot be sensed by a simple microphone. The intensity is calculated using Eq. 4 where power is calculated using Eq. 3 and the area has the value 4π.

$$Intensity = \frac{Power}{Area} \tag{4}$$

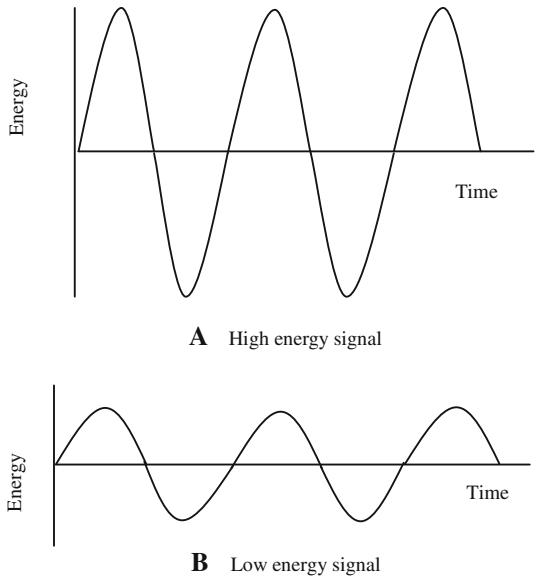


Fig. 1 Representation of sound wave

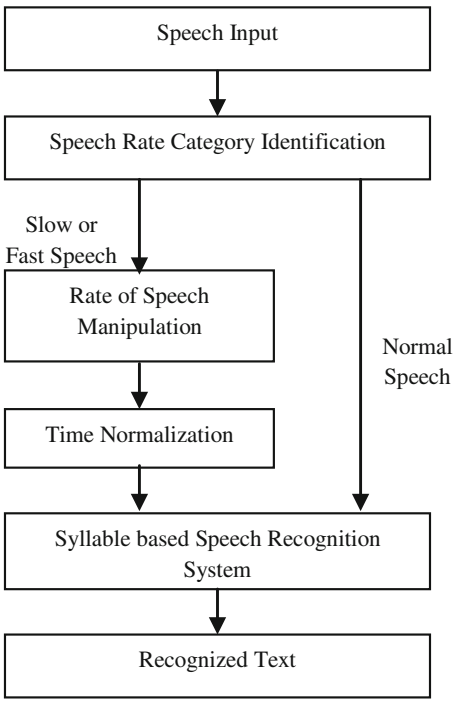


Fig. 2 Phases of the proposed model

3 Proposed model—speech rate dependent syllable based Tamil speech recognition system

Figure 2 shows the phases of the proposed model speech rate dependent syllable based Tamil speech recognition

system. The speech input is categorized in the speech rate category identification phase as slow, normal or fast speech. If the speech is either slow or fast, the manipulation of ROS and time normalization has to be done. If the speech is normal, the process of syllable based speech recognition system [10] is done and the system gives the recognized text as output. Each of the phases is described in the following sections in detail.

4 Speech rate category identification

The speech signal given as input is categorized into slow, normal and fast speech using the parameters sound intensity level (SIL), time duration and root mean square (RMS).

4.1 Sound intensity level

Sound intensity level is the logarithmic measure of sound intensity (measured in watts/meter^2) in comparison to a reference level which is shown in Eq. 5. The SIL is one of the measures used to manipulate the categories the speech signal.

$$L_1 = 10 \log_{10} \frac{I_1}{I_0} \tag{5}$$

where I_1 and I_0 are intensities and L_1 is the SIL measured in decibel. Decibel is a dimensionless quantity. I_1 is calculated using Eq. 4. I_0 is the standard reference sound intensity which has the volume 10^{-12} w/m^2.

4.2 Root mean square

In mathematics, the RMS, also known as the quadratic mean, is a statistical measure of the magnitude of a varying quantity. It is especially useful when variates are positive and negative, e.g., sinusoids. RMS is used in various fields like electrical engineering. It can be calculated for a series of discrete values or for a continuously varying function. Its name comes from its definition as the square root of the mean of the squares of the values. The RMS is one of the measures used in this work to find whether the given signal is a slow, normal or fast signal. The RMS is calculated using the Eq. 6.

$$RMS = \sqrt{\frac{\sum_{i=1}^{n} U_i \cdot conj(U_i)}{n}} \tag{6}$$

where U_i is the sample datum, conj(U_i) is the complex conjugate of the sample datum and n is the length of the sample data of the speech signal.

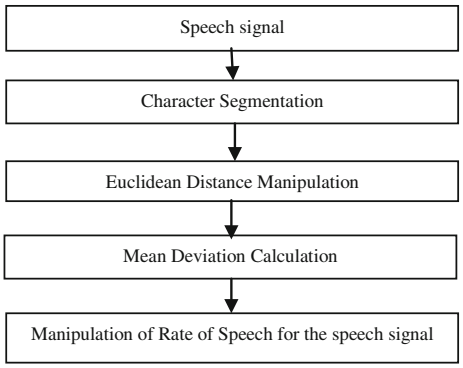


Fig. 3 Process in rate of speech manipulation

4.3 Time duration

The time duration for the speech signal is also a measure used to find the category of the speech. It is measured in unit of micro seconds (ms). The time duration for a slow, normal and fast speech signal will vary. The Time duration of a speech signal can be manipulated using the Eq. 2.

5 Process in ROS manipulation

The ROS manipulation consists of various processes which are shown in Fig. 3. The ROS of the given input utterance is calculated using the mean deviation of each character divided by the number of characters in the given input utterance which is shown in Eq. 7.

$$ROS = \frac{\sum_{i=1}^{n} MD_i}{n} \tag{7}$$

where MD is the mean deviation of each character and n is the number of characters in the input word. The mean deviation of each character is found by finding the mean of the feature vectors of the frames of each character and then finding the difference from the threshold mean. The threshold mean is the found manually by averaging the mean values of different normal speech rate characters.

5.1 Character segmentation

The individual characters are segmented from the given input utterance using the varied length maximum likelihood (VLML) algorithm. The VLML algorithm [11] is an extension of the maximum likelihood algorithm which finds the boundary of each phoneme. The segments with varied length can be segmented using the VLML algorithm.

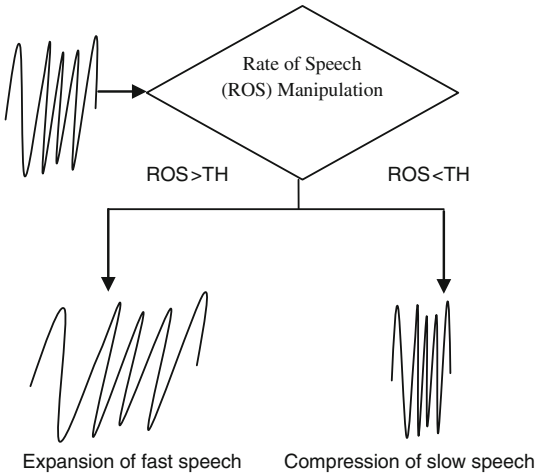


Fig. 4 Time normalization

5.2 Euclidean distance

Euclidean distance is the most widely used distance measure of all available distance measure [12]. In Eq. 8, the data in vector x and y are subtracted directly from each other. The Euclidean distance is manipulated for each frame and its successive adjacent frame and stored as a distance vector.

$$d = \sum_{i=1}^{n} (x_i - y_i)^2 \tag{8}$$

5.3 Mean deviation

The mean value of the Euclidean distance vector is found for each character. The deviation of the mean from the threshold mean value is also manipulated for each individual character. The threshold mean value is calculated using a normal rate signal. The signal with SIL value of range 116–121 dB is considered as the normal rate signal.

6 Time normalization

Time normalization [1] is a method where the speaking rate is controlled by using a scaling factor applied to the speech signal which is shown in Fig. 4. The speech signal is categorized as fast or slow or normal speech. If the speech signal is in the category of fast, then the signal has to be expanded by applying the scaling factor. If the signal is in slow category, then the signal has to be compressed by applying the corresponding scaling factor [13]. If it is a normal category then the signal has no modification. The process of modifying the speech signal either by expansion or compression using the scaling factor is called as time scale modification. If ROS manipulated using Eq. 3 is greater than the threshold value, the signal is treated as a fast signal and if ROS is less than the threshold, the signal is treated as a slow signal. The threshold is found by manipulating the ROS for a normal signal which is characterized by using the PRAAT tool.

6.1 Scaling factor

The scaling factor is manipulated using the ROS calculated in the Sect. 3. The scaling factor is used for applying expansion or compression to the signal. Let S_F denotes the scaling factor and N is the number of sample data in the signal. The expansion or compression is done after each S_F sample data in the signal.

(1) Criteria for fixing the scaling factor: The scaling factor is manipulated distinctly for the slow and fast signals. Let S_{SF} be the scaling factor for slow signal and S_{FF} be the scaling factor for the fast signal. Let ROS be the ROS manipulated using Eq. 8. The scaling factor S_{SF} and S_{FF} are manipulated using the Eqs. 9 and 10 [13].

$$S_{SF} = Ceil\left|\frac{1}{\frac{1}{ROS} - 1}\right| \tag{9}$$

$$S_{FF} = Ceil\left|\frac{1}{ROS - 1}\right| \tag{10}$$

7 Syllable based Tamil speech recognition system

This phase has the process same as the one followed in the baseline syllable based Tamil speech recognition system. The fast and slow speech signals after expansion and compression are given as input to this phase. The normal speech signal is directly given as input without any time normalization.

7.1 Character boundary identification

The segmented isolated word from the input utterance can be used as a model for the recognition system. But the word model has some drawback like the vocabulary size will be large. To form a syllable, initially the individual character boundary should be identified. The formant value can be used to identify the boundary approximately [11]. The formant range of each character of Tamil language is different. The first three formants F1, F2 and F3 are considered for identifying the category of the character of the input word [11]. The appropriate boundary of each character is given as input to the maximum likelihood

Table 1 First three formant limits of character categories and their threshold length in Tamil language

Character category	F1	F2	F3	Threshold length
Short vowel (SV)	<950	<1,450	<3,000	2,000
Long vowel (LV)	<800	<3,000	<3,300	4,000
Diphthong (D)	<800	<1,700	<3,000	3,500
Consonant (C)	<630	<3,000	<3,200	1,900
SV + C	<900	<2,500	<2,900	1,800
LV + C	<850	<2,100	<2,900	2,800
D + C	<800	<2,300	<3,000	3,800

segmentation procedure which finds the absolute boundary of each character.

The utterance of all the characters in Tamil language were recorded with Audacity and each character was uttered multiple times. The length of characters according to their character category was considered to find the mean length of each character category. The mean value found was taken as the threshold length of the particular category. Table 1 specifies the threshold length and the first three formants limits of each category.

The first three formant values F1, F2 and F3 of all characters were found. The maximum formant values for each character category were found by manipulating the mean value. The mean value of F1 formant values for the multiple utterances for each character was found. Similarly mean values were found for the formants F2 and F3. After finding the mean values of the three formants individually, the maximum formant values for each category was found and considered as formant limits.

The threshold length and the formant limit were manipulated for the character set of Tamil language. The number of segments in the given speech could be identified by finding the number of characters in the input speech. The length and boundaries of each character in the input word were approximately computed using the threshold length and formant limit values.

The boundary of the character segments of the given input word was found using VLML algorithm. The VLML is an extension of the maximum likelihood algorithm (ML). The maximum likelihood algorithm is a feature vector based segmentation approach. The spectral property similarity within a segment is checked using ML algorithm. The segment is a collection of frames and the feature vector of each frame was extracted using feature extraction procedures like Mel frequency cepstral coefficient (MFCC) and linear predictive coding (LPC). The segmentation in ML was done based on the spectral distortion within a segment which is the deviation on the spectral properties. The frame with minimum intra segment distortion may correspond to the boundary of the segment. The spectral distortion is measured in terms of intra segment distortion (δ) and generalized centroid (μ). In the traditional maximum likelihood, the segments are assumed to be equal size. The VLML uses a varied length segmentation procedure.

7.2 Character classification

The character segmented in the character boundary identification phase is classified in this phase. The Gaussian mixture model is used to identify the character using the Gaussian mixture model [10]. In the training phase, the model is created for all reference patterns. In the testing phase, the log likelihood is computed using posterior function for the test data with each reference pattern and the reference data with maximum log likelihood is classified as the character of the test data.

7.3 Syllable formation

The sequence of characters classified using the character classification phase is taken as input in this phase and syllable is formed [10]. Initially each character in the input is segregated as short vowel, long vowel, consonants and diphthong. The next step will be combining consonants followed by a short vowel as a single unit, consonants followed by long vowel as a unit and consonants followed by diphthong as a unit. The last step will be to form the syllable using the pattern rules. The segment of sound signal of each character whose boundary was found using VLML are concatenated to form a syllable.

7.4 Recognition system

The syllables are trained using HMM Toolkit. Initially a Grammar is constructed which defines the constraints that the speech recognition engine (SRE) can expect, when an input is given. The list of the syllables used for training the system is listed in the grammar [14]. This grammar is parsed as a word network, which the SRE can understand. The word network is defined using a low level notation called Standard Lattice Format. After constructing the Grammar, the next step will be to build a dictionary. The Dictionary consists of the list of syllables and corresponding output to be given by SRE when the input is recognized. The third step will be creating a transcription file to label the training data. The 39 MFCC feature vectors are extracted after creating the transcription files. The fifth step in training is to create an acoustic model. The acoustic model contains a statistical representation of the distinct sound that make up each syllable in the grammar. The syllables are represented in a statistical format which is used by the SRS to recognize an input in the testing phase.

Table 2 Sound intensity level, time duration and root mean square of some sample data

Speech signal	Speech rate category	Sound intensity level (in decibel)	Time duration (in ms)	Root mean square
S1	Slow	122.0432	3.3599	0.0853
S2	Normal	119.5446	1.3700	0.0601
S3	Fast	101.5424	0.7644	0.0076

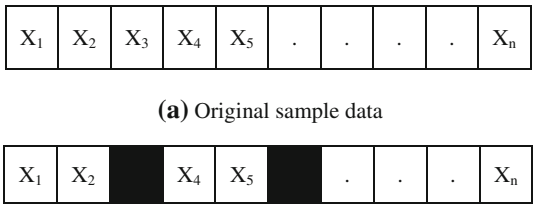
Table 3 Limits of the features of the speech rate category identification for a normal speech signal

Features of speech rate category identification	Lower limit	Upper limit
Sound intensity level	116 dB	121 dB
Time duration	0.9 ms	2.0 ms
Root mean square	0.04	0.07

8 Performance enhancement of speech rate dependent syllable based Tamil speech recognition system

The performance of the proposed model and baseline speech recognition model is compared with a dataset consisting of 200 syllables from agriculture application where the rates of various vegetables, fruits and grains are enquired. The speech data used for training was recorded using Audacity with project rate of 8,000 Hz using the single input channel microphone. The data was recorded in a noiseless room environment. Each input data is uttered 5 times by 4 speakers (2 female and 2 male). One utterance of each input is uttered with ROS less than the ROS threshold value and one utterance is uttered with ROS greater than the threshold value. The character set of Tamil language was recorded with the same setup as specified above with a single speaker.

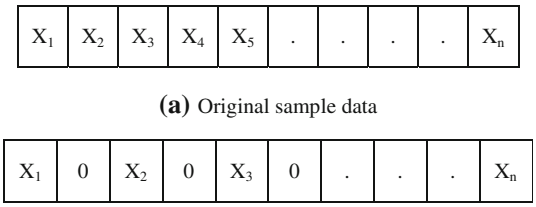
Table 2 shows the SIL, time duration and RMS values of some sample data which are uttered as slow, normal and fast speech. The values shown in the table corresponds to only to these sample signals S1, S2 and S3. The SIL, time duration and RMS are manipulated for all the signals in the dataset and the range of SIL, time duration and RMS for the normal speech signal is set as shown in Table 3. It can be noticed from Table 2 that the SIL of a slow speech signal is above the range specified in Table 3. Similarly the time duration and the RMS is also above the range. So it is clear that when values of the three features for a given input signal are above the upper limit of Table 3, the signal is a slow speech signal. When the values of the three features for a given input signal are below the lower limit of Table 3, the signal is a fast speech signal.



(a) Original sample data

(b) Sample data position where compression is done

Fig. 5 Compression of the speech signal S3



(a) Original sample data

(b) Sample data position where expansion is done

Fig. 6 Expansion of the speech Signal S1

Table 4 Performance comparison of the proposed model and the baseline model

Sample speech signal	Number of segments manipulated		
	Actual number of segments	Proposed model	Baseline model
S1	3	3	5
S2	3	3	3
S3	3	3	1

After identification of the speech rate category, the slow and fast speech signals has to be time normalized. So the ROS is manipulated for the signals using Eq. 8. As specified in Sect. 6, the expansion or compression of the sample data is done using the scaling factors specified in Eqs. 9 or 10. Figure 5 specifies the positions where compression is done for the sample S3 whose scaling factor was 2. As the scaling factor manipulated was having a value 2, so sample datum present after every 2 sample data are removed as shown in Fig. 5b.

Specifies the positions where expansion is done by inserting a zero for the sample S1 whose scaling factor was 1. As the scaling factor manipulated was having a value 1, so after every 1 sample datum, a zero is inserted as shown in Fig. 6b.

After the compression or expansion is done to the input signal, the signal to the baseline Syllable based Tamil

speech recognition system for the training and recognition process.

9 Conclusion

The performance of the proposed model and the baseline model in manipulating the number of character segments is compared. The proposed model after performing the time normalization, manipulates the number of character segments correctly for all the speech rate category signals as shown in Table 4. The baseline system manipulates correctly the number of character segments for the normal speech signal but not for the slow and fast speech signal which is also shown in Table 4 for some sample data of this work. The baseline system gives an accuracy of 70 % and the proposed model gives an accuracy of 74 %. The accuracy of the proposed is slightly higher than the baseline system. The future work is planned to improve the accuracy of the ASR system by considering more factors like channel variability that affect the performance and to apply the proposed system to a continuous speech signal.

References

1. Kang JA, Choi SH (2013) A speaking rate normalization method using time scale modification for speech recognition. ICCA 24:95–97
2. Ban SM, Kim HS (2012) Speaking rate dependent multiple acoustic models using continuous frame rate normalization. Signal and information processing association annual summit and conference (APSIPA ASC), Dec 2012, pp 1–4
3. Falthauser R, Pfau T, Ruske,G (2000) Online speaking rate estimation using Gaussian mixture models. In: IEEE Proceeding of ICASSP 2000, pp 1355–1358
4. Pfitzinger HR (1996) Two approaches to speech rate estimation. In: Proceedings of SST 96, pp 421–426
5. Zheng J, Franco H, Stolcke A (2000) Rate-of-speech modeling for large vocabulary conversational speech recognition. In: Proceedings of 2000 speech transcription workshop, 2000
6. Zheng J, Franco H, Stolcke A (2004) Effective acoustic modeling for rate-of-speech variation in large vocabulary conversational speech recognition. In: Proceeding of ICSLP, September, pp 401–404
7. Morgan N, Fosler E, Mirghafori N (1997) Speech recognition using on-line estimation of speech rate. Eurospeech-97, Rhodes 4:2079–2082
8. Nanjo H, Kawahara T (2002) Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition. In: Proceedings of ICASSP, pp 725–728
9. Siegler MA, Stern RM (1995) On the effect of speech rate in large vocabulary speech recognition systems. In: Proceedings of ICASSP, pp 612–615, May 1995
10. Ganesh AA, Ravichandran C (2013) Grapheme Gaussian model and prosodic syllable based Tamil speech recognition system. In: IEEE Conference Proceedings of ICSC 2013, pp 424–429, December 2013
11. Ganesh AA, Ravichandran C (2013) Syllable based continuous speech recognizer with varied length maximum likelihood character segmentation. In: IEEE conference Proceedings of ICACCI 2013, pp 935–940, August 2013
12. Akila A, Chandra E (2013) Slope finder—a distance measure for DTW based isolated word speech recognition. Int J Eng Comput Sci 2(12):3411–3417
13. Sylvestre B (1991) Time-scale modification of speech: a time frequency approach. ME Thesis, McGill University
14. Akila A, Chandra E (2013) Isolated Tamil word speech recognition system using HTK. Int J Comput Sci Res Appl 3(2):30–38