Certain Investigation on Phoneme Segmentation Techniques for Speech Signal

Bagavathi S[1]* and Padma SI[2]

Department of ECE, PET Engineering College, Vallioor, Tamil Nadu, India

Review Article

Received date: 24/10/2016

Accepted date: 14/11/2016

Published date: 20/11/2016

***For Correspondence**

Bagavathi S, Department of ECE, PET Engineering College, Vallioor, Tamil Nadu, India, Tel: 04637 220 999

E-mail: Sbagavathi1@gmail.com

Keywords: Phoneme segmentation, Speech signal, Speech synthesis, Support vector machine, Hidden Markov model (HMM)

ABSTRACT

The healthcare networks have grown up to the very large scale in the recent years and still occupying the healthcare sector in the variety of the areas in order to propagate the health related information to the centralized servers. The healthcare networks are being utilized in the post-treatment patient's health analysis application or the telemedicine applications for the assessment of the person's health in the remote areas for the correct medicine prescription. The proposed model has been designed for the prioritization of the healthcare data according to the criticality level of the information. The critical data handling and the primary categorization method has been designed in the proposed model for the handling of the critical data in the healthcare applications. The proposed model has been designed for the prioritization of the healthcare data according to the multi-level criticality assessment. The proposed model has been undergone the performance evaluation on the basis of the throughput and end-to-end delay parameters. The proposed model has been found efficient based upon the parameters evaluated from the proposed model simulation.

INTRODUCTION

The capacity to express thoughts and emotions by articulate sounds is called speech. Speech is the vocalized type of correspondence based upon the syntactic mix of lexicals and names that are drawn from extensive (for the most part around 1,000 one of a kind words) vocabularies. Each talked word is made out of the phonetic mix of a confined course of action of vowel and consonant speech sound units. These vocabularies, the linguistic structure which structures them and their arrangement of speech sound units change, making the presence of various a considerable number of different sorts of ordinarily jumbled human languages. Most human speakers can pass on in two or a more amount of sound measured, along these lines being polyglots [1]. The vocal capacities that enable people for individuals to convey speech furthermore outfit individuals with the ability to signal. The speech signal is made at the Vocal ropes, goes through the Vocal tract and Produced at speakers mouth. The gets to the audience members ear as a pressure wave. Non-Stationary, but can be isolated to sound segments. Two Major classes: Vowels and Consonants. The speech Production is a sound source energizes a (vocal tract) channel Voiced and Unvoiced. [Voiced: Periodic source, made by vocal ropes and Unvoiced: An occasional and loud source]. The Pitch is the fundamental frequency of the vocal lines vibration [2]. The fundamental sound of a language (e.g. “an” in “father”) is called phonemes. Phoneme segmentation is the capacity to separate words into individual sounds [3].

LITERATURE SURVEY

Chen et al. [4] depicts the IBM way to deal with Broadcast News (BN) translation. Regular issues in the BN interpretation undertaking are segmentation, bunching, acoustic displaying, clustering, demonstrating and acoustic model adjustment. This paper shows new calculations for each of these center issues. Some key thoughts incorporate Bayesian data rule (BIC) and speaker/group adjusted preparing.

Toledano et al. [5] presents a way to deal with programmed division of speech corpora. The accessibility of adequately exact marked sentences can evade the requirement for a division by human specialists. The objective of this procedure is to get ready speech corpora both for preparing acoustic models and for concatenative content to discourse union. This framework just needs the speech signal and the phonetic sequence for every sentence of a corpus [6]. It gauges a GMM by utilizing all sentences, where each Gaussian distribution speaks to an acoustic class. A DTW calculation settles the phonetic limits utilizing the known phonetic arrangement. This DTW is a stage inside an iterative procedure which plans to portion the corpus and re-estimate the conditional probabilities.

Amit and Carol [7] propose a technique that joins acoustic-phonetic knowledge with support vector machines for segmentation of nonstop speech into five classes - vowel, sonorant consonant, fricative, stop and quiet. This algorithm utilized a probabilistic phonetic component feature hierarchy and four classifiers are required to perceive the five classes. The hierarchical approach permits the utilization of tantamount measure of training data of two classes that every classifier is intended to segregate [8]. The segmentation with 13 learning based parameters performs extensively superior to a setting free Hidden Markov Model (HMM) based methodology that utilizations 39 mel-cepstrum based parameters. The probabilistic nature of the calculation permits the strategy to be expanded to phoneme and word acknowledgment with a little number of classifiers.

Adell and Belafonte [9] presents a way to deal with take the phone segmentation and the methodology in view of a Regression Tree to perform boundary specific correction of the HMM segmentation and distinctive evaluation techniques were discussed and the algorithm framework depends on HMM.

Prahallad et al. [10] address the pronounciation demonstrating for conversational speech amalgamation and different things with two distinctive HMM topologies for sub-phonetic demonstrating to catch the erasure and inclusion of sub-phonetic states during speech creation process and demonstrate that the tested Gee topologies have higher log probability than the customary 5-state successive model.

Hoffmann and Pfizer [11] provides phonetic segmentation of speech. This techniques extremely time consuming and slows down porting of speech system to new languages. In the setting of prosody corpora for text-to-speech (TTS) system, we explored strategies for completely automatic phoneme segmentation utilizing just the corpora to be segmentation and a naturally produced interpretation and exhibit another technique that enhances the execution of HMM-based segmentation by adjusting the boundaries between the preparation phases of the phoneme models with high accuracy [12].

Khanagha et al. [13] proposed a novel phonetic segmentation strategy in view of speech examination under the Microcanonical Multiscale Formalism (MMF) and depends on the calculation of nearby geometrical parameters, singularity exponent (SE). We demonstrated that SE convey on significant data about the nearby flow of speech that can promptly and basically used to recognize phoneme boundaries. In the initial step, this algorithm recognizes the boundaries of the original signal and a low-pass filtred form. The second step utilizes a theory test over the nearby SE distribution of the original signal to choose the last boundaries.

Chen et al. [14] present a novel methodology to combine acoustic data and emotional point data for a robust automatic reorganization of a speaker's feeling. Six discrete emotional states are perceived in the work. Firstly, a multi-level model for feeling acknowledgment by acoustic components is introduced. The determined elements are chosen by fisher rate to recognize diverse sorts of feelings. Besides, a novel emotional point model for Mandarin is set up by Support Vector Machine and Hidden Markov Model [15]. This model contains 28 emotional syllables which reflect rich emotional data. At last the acoustic data and emotional point data are coordinated by a soft decision technique and demonstrate that the use of emotional point data in speech feeling acknowledgment is successful.

Qiao et al. [16] proposed unsupervised phoneme segmentation without utilizing earlier data on etymological substance and acoustic models of an input sequence and develop the unsupervised segmentation by method for greatest probability, and demonstrate that the ideal segmentation relates to minimizing the coding length of the input sequence [17]. Under different presumptions, five distinctive target capacities are produced namely, specifically log determinant, rate distortion (RD), Bayesian log determinant, Mahalanob is separation and Euclidean separation goals and demonstrate that the ideal segmentation have the change invariant properties, present a time-constrained agglomerative clustering algorithm to discover the ideal segmentation, and propose a productive execution of the calculation by utilizing incorporation capacities [18]. The outcomes demonstrate that RD accomplishes the best execution, and the proposed strategy beats the past unsupervised segmentation techniques.

Khanagha et al. [19] displays the use of a profoundly novel methodology, called the Microcanonical Multi scale Formalism (MMF) depends on local scaling parameters that depict the inter-scale relationships at every point in the signal space and gives productive intends to consider local non-straight progression of complex signals and present an efficient route for estimation of these parameters.

RESULTS AND DISCUSSION

In the survey paper [11] were discussed from the speech system SVOX, the prosodic segment for the capacity to utilize the system. The segmentation procedure must not depend on the accessibility of any physically segmentation information for the language.

In the survey paper [7] shows the classes that are prepared against each other for building these four SVMs. since all the decisions are binary, the method used to good multi-class SVMs. In spite of the non-probabilistic chain can be utilized to restrict the quantity of phonetic feature, methodology for probabilistic segmentation errors at phonetic component level-will not are corrected by language and more length limitations **(Table 1)** [20].

Table 1. Training of phonetic feature svms.

Branch in hierarchy	Class +1	Class -1
P1	silence	speech
P2	sonorant	non-sonorant
P3	sonorant consonant	vowel
P4	stop burst	frication noise

In the survey paper [9] were observed the results are bad in DTW. They were performed with all the more physically segmented sentences, and these sentences were picked and utilizing a greedy calculation from the speak language variability. The exactnesses are appeared in **Table 2**.

Table 2. Dtw physically segmented sentences.

Sentences	<5	<10	<15	<20	<25
40	30%	50%	62%	69%	73%
200	37%	61%	72%	80%	85%
300	39%	59%	72%	80%	84%
400	40%	62%	77%	85%	88%

In the survey paper [21] were discussed the frame of going before silence are appropriately grouped by the HMM states. At the phonetic unit speaking to quiets is viewed a special case, for utilized the topology. Sub-sampling rate is 200 Hz and HMM with 8 radiating states drives minimum phone duration of 40 ms, which is longer than some phonetic units.

In this survey paper [16] were observed by the normal log probability scores of utterance from Mod1 and Mod2 are better than Mod0 subsequently showing a better fit for the speech information **(Table 3)**.

Table 3. Average log probability scores of utterances.

Model	Avg. Log probability
Mod0	24217
Mod1	23522
Mod2	23978

In this survey paper [4] were Compared to thresholding for the BIC method tends to support more Gaussians for complex sounds (vowels)and support less Gaussians for basic sounds(fricatives).

In this survey paper [22] were observed for the segmentation quality can be utilizing three different coding schemes: the Hit Rate (HR) which is the rate of effectively recognized coding and the False Alarm Rate (FA) which is the rate of incorrectly recognized coding **(Table 4)** [23].

Table 4. For three different coding schemes.

Coding scheme	HR	FA
8-Mel-bank	86	30.69
MFCC	76	31.33
Log Area Ratio	70	34.16

The comparative analysis in the survey paper [19] shows that the response was poor for the minimum Test dataset is reported only for 20 ms tolerance.

ACKNOWLEDGMENT

The author would like to thank my guide SI Padma, Assistant Professor, Department of ECE, PET Engineering College.

CONCLUSION

In this paper, we have discussed about the voice signal and unvoiced signal with the speech on emotion recognition rate from continuous speech were the various parameters such as signal to noise ratio and hit rate, tolerance were compared based on emotional point.

REFERENCES

1. Adell J and Bonafonte A. Towards phone segmentation for concatenative speech synthesis in Proc. 5th ISCA Workshop

Speech Synthesis. 2004;139-144.

2. Almpanidis G and Kotropoulos C. Phonemic segmentation using the generalised Gamma distribution and small sample Bayesian information criterion. Speech Communication. 2008;50:38-55.

3. Amit J and Carol EW. Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines in Proc Int Joint Conf. Neural Newts. 2003;1:675-679.

4. Aversano G, et al. A new text-independent method for phoneme segmentation in Proc. 44th. IEEE Midwest Symp. Circuits Syst. (MWSCAS). 2001;516-519.

5. Burckhardt F, et al. A database of German emotional speech. Proc Inter speech. 2005;1517-1520.

6. Chen S, et al. Automatic transcription of broadcast news. Speech Communication. 2002;37:69-87.

7. Chen L, et al. Speech emotional features extraction based on electroglottograph. Neural Compute. 2013;25:3294-3317.

8. Forney GD. The Viterbi algorithm. Proc IEEE. 1973;61:268-278.

9. Granqvist S, et al. Simultaneous analysis of vocal fold vibration and transglottal airflow: Exploring a new experimental setup. J Voice. 2003;17:319-330.

10. Hartem D. Multifractals: Theory and applications. CRC Press, Boca Raton, FL, USA; 2001.

11. Hoffmann S and Pfister B. Fully automatic segmentation for prosodic speech corpora. Proc Interspeech 2010;1389-1392.

12. Huang NE, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc Roy Soc 1998;903-995.

13. Kania RE, et al. Fundamental frequency histograms measured by electroglottography during speech: A pilot study for standardization. J Voice 2006;20:18-24.

14. Khanagha V, et al. Improving text-independent phonetic segmentation based on the micro canonical multiscale formalism. In Proc Acoustic Speech Signal Process (ICASSP). 2011;4484-4487.

15. Khanagha V, et al. Phonetic segmentation of speech signal using local singularity analysis. Digital Signal Process. 2014;35;86-94.

16. Nakagawa S, et al. Speaker recognition by combining MFCC and phase information. Spectrum. 2007;60:76-84.

17. Prahallad K, et al. Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. Proc IEEE Int Conf Acoustic Speech Signal Process. 2006;I1-CI4.

18. Qiao Y, et al. Unsupervised optimal phoneme segmentation: Theory and experimental evaluation. IET Signal Process. 2013;7:577-586.

19. Qiao Y, et al. Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons. Proc Acoustic Speech Signal Process. 2008;3989-3992.

20. Romsdorfer H and Pfister B. Phonetic labeling and segmentation of mixed-lingual prosody databases. Proc Interspeech. 2005;3281-3284.

21. Toledano DT, et al. Automatic phonetic segmentation. IEEE Trans Speech Audio Process. 2003;11:617-625.

22. Yuan J, et al. Automatic phonetic segmentation using boundary models. Proc Interspeech. 2013;2306-2310.

23. Chen LJ, et al. Mandarin emotion recognition combining acoustic and emotional point information. Appl Intell 2012;37:602-612.