International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 4 Issue 5, May-2015, pp: (200-204), Impact Factor: 1.252, Available online at: www.erpublications.com

Age Verification from Speech Using Excitation features and Gaussian Mixture Models

Pooja Pokhriyal[1], Swati Devliyal[2], Anurag Barthwal[3]

[123]Dept. of Computer & Science Engineering, GEHU, Dehradun, India

Abstract: Detecting the age of the speaker is one of the major applications of speech recognition in Today's world. This work is related to Age Detection of speakers of different age groups. The age group considered in this study are- 10-20, 21-30 and 31-40. Since very little work has been done with natural database so in this work a natural database is used which is recorded in Hindi from male and female speakers of different age group. Excitation Source features is extracted from Speech for characterizing the different age groups of the speaker. This paper explores the Linear Prediction (LP) residual of speech signal for characterizing the age groups as in excitation features the LP residual is a error signal which contains 50-60% noise only .So in this work main aim is to find out whether LP residual contains more age specific information or not. Gaussian Mixture models are used to develop age models.

Keywords: Age Verification, CS-AVSC (Computer Science – Age Verification Speech Corpus), Excitation Source features, GMM (Gaussian Mixture Models), Linear Prediction coefficient.

Introduction

Speech is an ability through which a person expresses thoughts and feelings. Speech is produced by the articulation sounds. Each word that is spoken by human is created using the phonetic combination of limited set of vowel and consonant speech sound unit. Speech is an important and crucial mode of communication as it delivers a lot of information like age of speaker, health of speaker, gender, language, emotions and so on. Among these, age of a speaker is an important characteristic features which finds several applications like- To restrict particular websites and their contents from youth, In sports to confirm the age, In voting system to confirm the age of a voter, to restrict the minors (young children) of below 13 years of age from social networking sites, in online business to increase targeting for marketing and sales through real time age verification. Generally the difference between the estimated age and actual ages based on voice recording are small so this work mainly finds the age of a speaker [10].

For characterizing the age of a speaker one needs to extract features from following levels -excitation source features (sub-segmental) ,spectral features (system features or vocal tract),prosodic features (supra segmental). In this work Excitation source features are used for classifying the age of a speaker which confines its different scopes in futures. Excitation source are obtained after suppressing the vocal tract (VT) characteristics which is achieved by predicting the VT information by the use of filter coefficient known as Linear prediction coefficient (LPCs) .Excitation source signal may contain the age specific information in the form of unique features such as higher order relations among linear prediction (LP) residual samples [2]. The term 'Linear Prediction' means predicting new values or output from the previous values or inputs. The resulting signal LP residual contains the information about the excitations source features from speech samples. From the literature study it is also found that LP residual contains 50-60% noise only, so in this work our motivation is to find whether LP residual contains the age specific information or not.

In order to prepare the system model database collection is one of the first and necessary tasks. Generally age specific database can be classified into three categories- natural, simulated and elicited. Natural database is that which is collected from the real world situations like teacher –student, doctor-patient talk. Its advantage is that it is completely natural and availability of real world age for modeling and analysis in the system .Its disadvantage is of copy rights issues and collection as sometimes it is very difficult to collect. Simulated database is that where the sentence is given to the actor and actor tries to speak it in different age groups. This database is based on speaker. Advantage of this database is that it is mostly used, more reliable and standard known database. Its disadvantages are that in this database only pre decided speech is read which is only read and not expressed. Third is elicited speech database which is collected when the actor is not aware that he/she is been recorded. In this anchor prompts the speaker to produce voice in different age groups in order to get different age groups samples .The advantage of this database is that it is easily available and it is also somewhat like natural database. Disadvantage is that the quality of the database degrades if the

actor comes to know about his/her recording .As a classifier Gaussian Mixture Model have been used in this study for developing age recognition models.

The rest of the paper is organized as- Section II describes about the database details .Section III provides the details about the feature extraction .Section IV gives the overview of GMM classifier .Development of age approximation models and results is provided in Section V .Section VI contains the summary and conclusion. Section VII concludes the paper, containing references.

Database: CS-AVSC

Database collection is one of the first and necessary perquisite tasks as the whole performance of the system models depend on the quality of the database. According to literature survey, a lot of work has been done with simulated and elicited database so in this work we are working with natural database which is collected from the speakers of the different age groups. Age groups considered in this work is- 10-20, 21-30, 31-40. Database named in this work is Computer Science -Age Verification Speech Corpus (CS-AVSC) where five sentences are recorded in sequence by 10 speakers (5male and 5 female) of Hindi language which is spoken in flat pitch for each age group. The database contains 250 utterances (5 sentence* 5 sessions* 10 speakers) and sums to 750 utterances (250*3 age groups).Recording has been done in anechoic room where the voice is recorded with the frequency of 16 kHz and each sample is stored as a 16 bit number in mono channel. Out of this 70% of the database is used in training the GMM models and the rest 30% is used in testing the trained models.

While collecting the database, there are few steps. Initially the sampling rate is set to 16 kHz and mono channel with a bit of 16 resolutions are chosen. The reason for taking 16 kHz sampling rate is that human voice vary between 20-32 kHz if it is taken as 32 kHz than it is very difficult to capture it and if it is taken below 10 kHz than very less number of information will be collected from that .So in order to get good quality database this much sampling rate is taken. While pre-processing, longer silences are noticed which are removed with the help of tool wave surfer.

Features Extraction

Choosing suitable features for developing any of the speech system is an important decision .Here in this work excitation source features known as sub-segmental features are used for analyzing the age present in the speech. Speech features derived from excitation source are called as source features which are obtained after suppressing the vocal tract (VT) characteristics. It is achieved by predicting the VT information using filter coefficient (linear prediction coefficient (LPCs)) from speech and then separating it by filter formulation. The resulting signal is linear residual which contains excitation source information. Features extraction from speech is done with small speech segments of length 20-30 msec. It is a block processing approach where the entire speech signal is processed block by block and the block size is taken as 20 msec. Blocks are also known as frames. LP residual is obtained by inverse filtering of speech signal which can be shown by the equation –

$$S(n) = 1 + \sum_{k=1}^{p}((a_k (S(n-k)) \qquad (1)$$

Where S (n) is a current speech sample, p is a order of prediction, a_k is the filter coefficient and S (n-k) is the $(n-k)^{th}$ sample of speech [3].

Due to variations in speech signals the features extraction is done with short frames of 30-50 per second. The sampling rate considered in this work is 16 kHz, value of LP order is 13 and frame size is taken as 320 ms with a shift of 160 ms every time. Preempflag and plot flag are taken as 0 and 1.Here is a LP residual format where the full code runs in MATLAB-

LPresidual_v3 (speech, frame size, frame shift, lporder, preempflag, plot flag)

Function [residual, LP Coefficient] = LPresidual_v3 (speech1, 320, 160, 13, 0, 1]

Classifier

Different classifiers can be used for developing speech system like age recognition. But the important task is in choosing the one classifier among all classifiers. Generally pattern classifiers are classified into-

i. Linear classifiers and

ii. Non- Linear classifiers

Linear classifier performs the classification decision based on the value of linear combination of the object characteristic. Non Linear is based on weighted combination of object characteristics [2]. In this work GMM (Gaussian Mixture Models) is used to develop the age recognition model from speech samples.

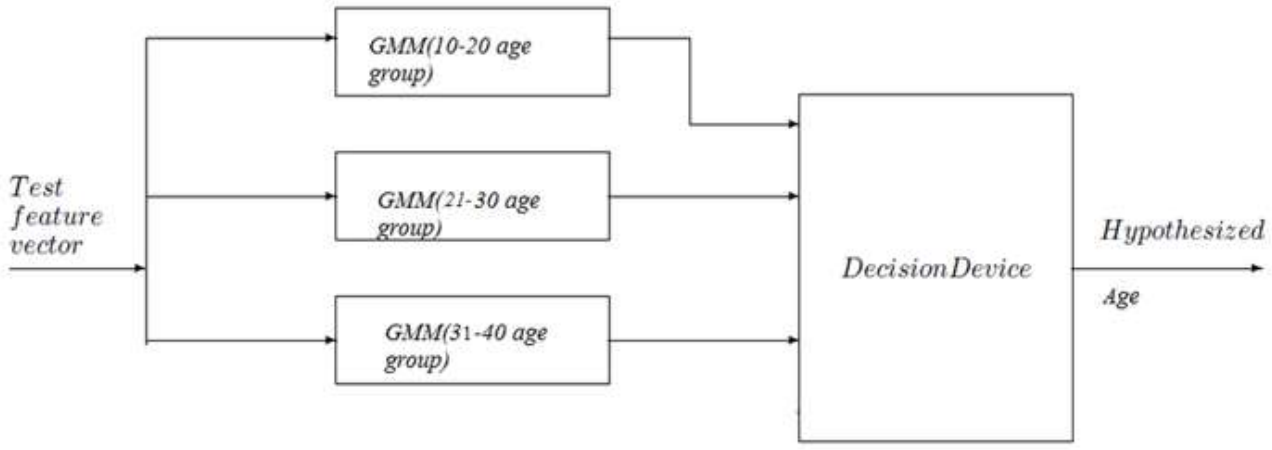

Figure 1: GMM Model

GMM is most matured method used for clustering and density estimation as it captures high order relations among samples. Since the values generated after the feature extraction are large so this classifier is easier to develop the models as GMM is based on cluster making. In this work, one GMM is developed to capture the information about one age group. The GMM components capture finer level details among the features vectors. Number of Gausses in the mixture model is the number of components which indicates the number of clusters in which data points are to be classified. There are two algorithm used in this models –

1. K-Mean Clustering algorithm –which is used to make the clusters of features vectors .

2. EM (Expectation Maximization) algorithm – which refines the weights of each distribution given by the set of inputs.

Development of Age Verification Models (AVM's)

Age verification models are developed from the test utterances which are given inputs to all GMM models. Test utterances are of male and female speaker where 70 % data is used in training the age recognition models and 30% is used in testing or validation. The output of each model is given to the decision device. Decision device determine or detects the age based on the highest score among the evidence provided by the age models.

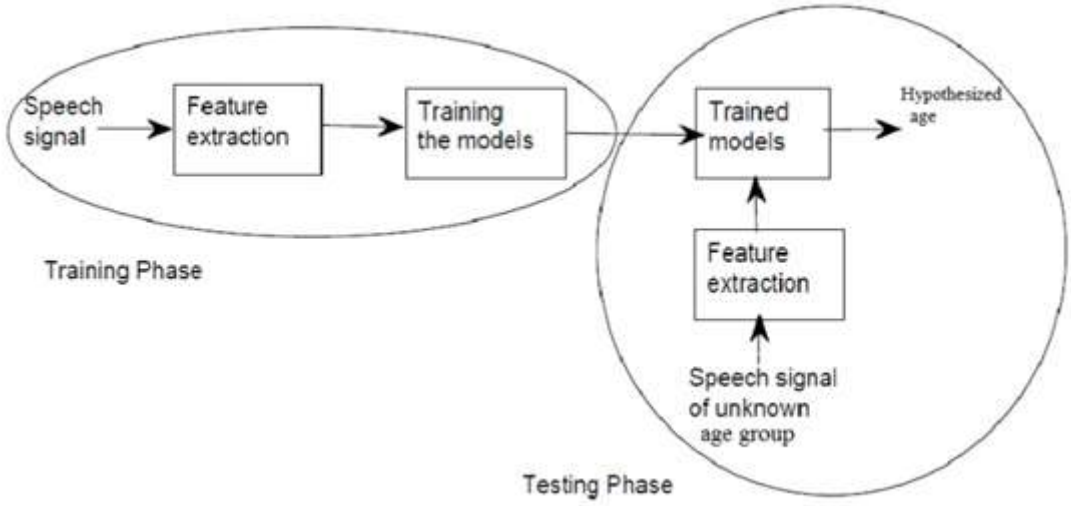

Figure: 2- Development of Age Recognition System

The overall methodology of age recognition is divided into two phase-In the first stage age recognition models are developed by training the models by features vectors which are extracted from the known age group. This type of learning is called supervised learning. From this phase we get three GMM model for each age group 10-20, 21-30 and 31-40 where the age group GMM model 10-20 recognizes only age groups between 10-20 .GMM model 21-30 recognizes age group between 21-30 and similarly for the last age group model 31-40. In the second phase, evaluation of the trained model is done in which the feature vector of unknown age group is given to the entire trained model.

Here the Euclidian distance between the models is calculated. The model that gives the least distance gives the hypothesized age.

Results

Table I: Age Verification performance for Male Speaker for Database (CS-AVSC)

Ages Verification Models	Approximation Performance (in %)		
	10-20	21-30	31-40
10-20	**90**	10	0
21-30	0	**100**	0
31-40	0	8	**92**

First this experiment is conducted for Male speakers where 4 sentences of male speakers are considered for training and 6 for testing. From Table 1it is analyzed that 10-20 age group is recognized 90% as 10-20 age models,10-20 age group is recognized 10% as 21-30 age group and 10-20 age group is recognized 31-40 as 0 % . Here in this only first column shows the correct classification and rest column shows the misclassification. From second row 21-30 age group is recognized 21-30 age group as 100 %, 21-30 age group is recognized 10-20age group as 0%, 21-30 age group is recognized 31-40 age group as 0 %. Only second column shows correct classification rest is misclassification. Similarly for last age group, 31-40 age models recognized 31-40 age group as 92%, 31-40 age group is recognized 10-20 as 0%, and 31-40 age group is recognized 21-30 as 8%. Column third shows the correct classification rest shows misclassification.

Table II: Age Verification performance for Female Speaker for Database (CS-AVSC)

Age Verification Models	Approximation Performance (in %)		
	10-20	21-30	31-40
10-20	**92**	8	0
21-30	3	**97**	0
31-40	0	6	**94**

Secondly this experiment is conducted for female speaker where 6 sentences are used for training and 4 are used for testing. From Table II it is analyzed that in first row 10-20 age group is recognized 92 % as 10-20, 10-20 age group is recognized 21-30 age group as 8%, 10-20 age group is recognized 31-40 age group as 0 %.Only the first column shows correct classification and rest shows misclassification. For second row 21-30 age group is recognized 97 % as 21-30, 21-30 age groups recognized 10-20 as 3 %, 21-30 age group recognized 31-40 age group as 0%. Only the second column shows the correct classification rest shows the misclassification. For last row 31-40 age group is recognized 94 % as 31-40 age group, 31-40 age group is recognized 10-20 as 0 % and 31-40 age group is recognized 21-30 as 6%. Only the third column shows the correct classification rest shows the misclassification.

Table III: Age Verification performance for (male +Female) Speaker for Database (CSAVSC)

Ages Verification Models	Approximation Performance (in %)		
	10-20	21-30	31-40
10-20	**93**	7	0
21-30	5	**95**	0
31-40	0	10	**90**

Third one this experiment is conducted for male and female speaker (male + female) where both utterances are used for training and testing simultaneously, where 5male and 4 female are taken for training and for testing 4 male and 6 female are considered for testing. Similarly from Table III it is analyzed that in first row 10-20 age group is recognized 10-20 as 93 %, 10-20 age group is recognized 21-30 as 7%, 10-20 age group recognized 31-40 as 0%.Only the first column shows correct classification and rest shows misclassification. For second row 21-30 age group recognized 21-30 age group as 95 %, 21-30 age group recognized 30-20 as 18 %, 21-30 age group recognized 31-40 as 0%. Only the second column shows the correct classification rest shows the misclassification. For last row 31-40 age group recognized 30-40 age group as 90%, 31-40 age group recognized as 10-20 is 0 % and 31-40 age group recognized as 21-30 is 10%. Only the third column shows the correct classification rest shows the misclassification.

Table IV: Average Age Verification performance (in %) for Database (CS-AVSC)

Average Age Approximation	Approximation Performance (in %)
Male	93.3
Female	94.3
Male + Female	92.6

Table IV display the average age performance of male, female and male +female only .The results obtained using CS-AVSC as a database is acceptably good in all cases in performance. Average Performance in case of Male is 93.3 %, for female it is 94..3% and for both is 92.6%.In Table I of male speakers the age group 21-30 is showing the best results for excitation features due to the fact that voice of male speaker varies in pitch in small ages but the male voice tends to stabilizes in the age of 20.In Table IV the average age performance comparison is done where the female performance is finds slightly better than male, due to les energy loss at the glottal closure.

Conclusion and Future work

From the results it can be concluded that excitation source features can be used for detecting the age of speakers which satisfy the motivation successfully that excitation features contain age specific information. Average age performances of male and female speakers are found to be good.

For future work Prosodic and spectral features can be used as the combination to enhance the performance, database can be taken in multiple languages and for classifiers hybrid models can be used in developing the age models. Thus the performance percentage can be increased in age models.

References

[1]. S. Prasanna, C. Gupta and B. Yehnanarayana, "Extraction of speaker specific information from linear prediction residual of speech, J. Acoust" Soc., Amer., Speech Communication,vol.48,pp. 1243-1261, Oct.2006.
[2]. Shashidhar .G. Koolagudi ,Swati Devliyal, Bhavna Chawla, Anurag Barthwal and K. Sreenivasa Rao," Recognition of Emotion From Speech Using Excitation Source Features," in Proc of International Conference Modeling ,Optimization and Computing June 2012
[3]. Arun Chauhan, S.G.Koolagudi, Sabin Kafley and K. Sreenivasa Rao," Emotion Recognition Using LP Residual, Proceedings of the 2010 IEEE Students Technology Symposium, 3-4 April 2010,IIT Kharagpur.
[4]. A Brief Introduction to MATLAB September 5, 1999, Jeffrey A.Fessler.
[5]. Tanushri Mittal,Anurag Barthwal, S. G. Koolagudi , "Age Approximation From Speech Using Gaussian Mixture Models," June 2012.
[6]. B.Atal, "Automatic Speaker Recognition Based On Pitch Contours," J. Acoust Soc. Amer, vol 52,pp.1972.
[7]. S.R. Mahadeva Prasanna , Cheedella S. Gupta , B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech ," Springer, 19 June 2006.
[8]. D., Neiberg, K., Elenius, and K., Laskowski, "Emotion recognition in spontaneous speech using GMM," in INTERSPEECH 2006, ICSLP , (Pittsburgh, Pennsylvania), pp. 809-812, 17-19, September 2006.
[9]. L., R., Rabiner, and B., H., Juang, Fundamentals of Speech Recognition. Englewood Cliffs, New Jersy: Prentice-Hall, 1993.
[10]. Lass, N.J., Justice, L.A., George, B.D., Baldwin, L.M., Scherbick, K.A. and, Wright, D.L.(1982). Effect of vocal disguise on estimations of speaker's ages. Perceptual and Motor Skills, 45 (3), 1311-1315.
[11]. Gupta, C.S., 2003., " Significance of source features for speaker recognition ," MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai600 036, India.
[12]. Reddy, K.S., 2004. "Source and System features for speaker Recognition," MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India.