ISSN:0975-9646

Identification of Age And Gender Using HMM

Bhavana R. Jawale.
(PG student)
G.H.R.I.E.M.,Jalgaon.
Jalgaon.

Swati patil.
(Asst.Prof.)
G.H.R.I.E.M.,Jalgaon.
Jalgaon.

Mayur Agrawal.
(Asst.Prof.)
G.H.R.I.E.M.,Jalgaon.
Jalgaon.

Abstract—**Deep Neural Network Hidden Markov Models, or DNN-HMMs, are recently very promising acoustic models achieving good speech recognition results over Gaussian mixture model based HMMs (GMM-HMMs).The accurate methods to profile different characteristics of a speaker from recorded voice patterns, which facilitate to identify him/her or at least narrow down the number of suspects. Here they propose a new gender and age group recognition approach based on Hidden Markov Model (HMM). First, an acoustic model is trained for all speakers in a training database including male and female speakers of different age. Finally, Supervised HMM is applied to detect the gender and age group of unseen test speakers. Designing a new HMM-based approach for speaker, gender and age estimation, which improves the accuracy of the state-of-the-art speaker age estimation methods with statistical significance. Analyzing the effect of major factors influencing the automatic gender and age estimation systems.**

1.INTRODUCTION

Recently there has been a growing interest to improve human-computer interaction. It is well-known that, to achieve effective Human-Computer Intelligent Interaction (HCII), computers should be able to interact naturally with the users, i.e. the mentioned interaction should mimic human-human interactions. HCII is becoming really relevant in applications such as smart home, smart office and virtual reality, and it may acquire importance in all aspects of future peoples life[6].

Speech age & gender recognition aims at recognizing the underlying state of the speaker from his or her speech signal[2]. This is mainly motivated by intelligent Human Machine Interaction required for different kinds of applications. In the field of speech age & gender recognition, a number of classification approaches have already been explored. According to the used acoustic features, the recognition models can be classified into two types: 1) for supra segmental prosodic features, such as the mean, median, standard deviation, range, or percentile of short time pitch (energy), estimated over the whole utterance, global models such as Gaussian mixture model(GMM), support vector machine(SVM), artificial neutral networks (ANN) and k-NN have been adopted. 2) for frame based dynamic spectral features like Mel Filter bank Cestrum Coefficient (MFCC), dynamical models such as Hidden Markov Model (HMM) are considered [1], [2]. Compared to the global models, dynamic modelling approaches provide a better consideration of the temporal dynamics of speech[6].

Understanding human states is indispensable for human-human interaction and social contact. Human age & gender states affect perception and rational decision making during human-human interactions. Hence, automatic speech recognition is important in harmonious interactions or communication between computers and human beings. The challenging research field, “affective computing,” introduced by Picard [1] aims at enabling computers to recognize, express, and have age & gender.

The speech recognition focused only on considering vocal modality. Based on psychological analysis [8], [9] it was found that human states were mainly transferred through multiple channels such as face, voice, body gesture, and speech content. For this reason, exploring data fusion strategies can achieve better recognition performance[12]. Many data fusion approaches have been developed in recent years. Fusion operation can be conducted at the feature level, decision level, and model level for audio-visual speech recognition [9]. In feature-level fusion [10], facial and vocal features are concatenated to construct joint feature vectors and then modeled by a single classifier for speech recognition. However, fusion at the feature level will increase the dimensionality and may suffer from the problem of data sparseness. In terms of decision-level fusion [14], multiple signals can be modeled by the corresponding classifier first and then the recognitions from each classifier are fused in the end.

Although fusion at the decision level enables us to interpret the performance of different classifiers and to gain insights into the role of multiple modalities, the assumption of conditional independence does not consider mutual correlation among multiple modalities. Contrary to the decision level, model-level fusion focuses on the mutual correlation among the multiple signal streams, but it is difficult to explore the contributions of multiple modalities.

A peculiar and very important developing area concerns the remote monitoring of elderly or ill people. Indeed, due to the increasing aged population, HCII systems able to help live independently are regarded as useful tools. Despite the significant advances aimed at supporting elderly citizens, many issues have to be addressed in order to help aged ill people to live independently. In this context recognizing people age & gender state and giving a suitable feedback may play a crucial role. As a consequence, age & gender recognition represents a hot research area in both industry and academically. There is much research in this area and there have been some successful products [10].

Usually, age & gender recognition systems are based on facial or voice features. This is a solution, designed to be employed in a Smart Environment, able to capture the age & gender state of a person starting from a registration of the speech signals in the surrounding obtained by mobile devices such as smart phones.

Traditional recommender systems then use these profiles, together with meta-data and ratings from other users in the network, to provide personalization. One of the issues however, in the context of broadcast TV, is the lack of an uplink channel, through which information such as ratings can be exchanged with the remaining users. It is therefore highly desirable that feedback from users be collected locally, in the set-top box or smart TV if possible, and as unobtrusively as possible, e.g. such as through unobtrusive relevance feedback [3].

By means of local recommendation and implicit user feedback, these systems can work quite effectively, but it is important to consider the preferences of a group of users as well as a single user. This is a particular issue when multiple consumers share a single device, such as a home television, but each has their own user profile and tastes [4]. In the Socially Aware TV Program Recommender for example [5], groups of users who want simultaneous access to the TV are taken into account, where individual profiles that have a common interest are combined.

2.History And Analysis

2.1. Literature Review:

The several works have been dedicated to DNNHMMs based large vocabulary continuous speech recognition. However, to knowledge only few works on the application of DNN-HMMs in age & gender recognition, have been reported. In [8], a Generalized Discriminant Analysis (GerDA) based on DNNs, is to learn the discriminative features for classifying high or low of arousal and positive or negative valence.

Recently, most researchers have seen increased attention being given to decision level and model-level fusion in data fusion approaches. Accordingly, two popular data fusion approaches at decision and model levels: error weighted classifier combination and the coupled hidden Markov model (C-HMM). The former used an empirical weighting scheme for recognition decision, and the latter modeled the asynchronous (e.g., audio and visual) nature of the multi-stream features for different applications. These models were successfully used in different fields such as age & gender recognition, interest detection, human identification, hand gesture recognition, 3-D surface inspection, speech prosody recognition, audio-visual speech recognition, and speech animation.

Visual information has been shown to be useful for improving the accuracy of speech recognition in both humans and machines [4]. These improvements are the complementary nature of the audio and visual modalities. For example, many sounds that are confusable by ear are easily distinguishable by eye. The improvements from adding the visual modality are often more pronounced in noisy conditions where the audio signal-to-noise ratio (SNR) is reduced [5].

When developing a speech recognition system that incorporates both the audio and visual modalities, a principled method for integrating the two streams of information must be designed. Because of the success of hidden Markov model (HMMs) in audio speech recognition, most audio-visual speech recognition (AVSR) systems extend HMM techniques to incorporate both modalities. This is describe efforts in developing an AVSR system which is built upon existing segment-based speech recognizer [7]. This AVSR system incorporates information collected from visual measurements of the speaker’s lip region using an audio-visual integration mechanism that we call a segment-constrained HMM [8].They are a new unified training algorithm for both the feature extractor and HMM classifiers. We interpret the feature extractor as a multilayer perceptron (MLP) with four layers, i.e., one for the filter banks, one for the feature transformation, and two for the delta and acceleration calculations. It enables us to derive efficient expressions of weight update formulas systematically by back propagation for all of the feature extractor modules. The back propagation starts with the output of HMM classifiers through an efficient inversion algorithm.

Determining both the age and gender of speakers is a complicated task and has received considerable attention in recent years. The achieved are encouraging and are beginning to make it feasible to use this technology as a viable alternative to existing methods of providing user demographics. Age and gender classification systems are generally implemented as a fusion of several subsystems [6], with each subsystem operating using a form of Gaussian mixture model, multilayer perceptron, hidden Markov models and/or support vector machines [8].

If the phone is aware of its owner mood can offer more personal interaction and services. Mobile sensing, in recent years, has gone beyond the mere measure of physically observable events. Scientist studying affective computing [2], have published techniques able to detect the age & gender state of the user [2], allowing the development of age & gender-aware mobile applications [7]. Existing work focused on detecting age & genders rely on the use of invasive means such as microphones and cameras [5], and body sensors worn by the user [7]. There is method based on the employment of audio signals represents an efficient alternative to the mentioned approaches.

The general influence of speaker age on voice characteristics is being studied since the late 1950s [1] and sustained continuous attention since then (see e.g. [2]), the first actual systems estimating the age and the gender of the speaker were developed only recently [6]. The quality of these systems is difficult to compare, as they vary considerably regarding the number and distribution of speaker age as well as the types of speech material.

The variability of IVR system use patterns across age and gender is investigated in [7], indicating that dialog strategies tailored to specific age and gender groups can be very useful in improving overall service quality. In this context recognizing people age & gender state and giving a suitable feedback may play a crucial role. As a consequence, age & gender recognition represents a hot research area in both industry and academic field. There is much research in this area and there have been some successful products [11].

Usually, age & gender recognition systems are based on facial or voice features. This is a solution, designed to be employed in a Smart Environment, able to capture the age & gender state of a person starting from a registration of the speech signals in the surrounding obtained by mobile devices such as smart phones.

Main problems to be faced concern: the concept of age & gender, which is not precisely defined for the context of this is the lack of a widely accepted taxonomy of age & genders and age & gender states; the strong age & gender manifestation dependency of the speaker. Age & gender recognition is an extremely difficult task. This paper presents the implementation of a voice-based age & gender detection system suitable to be used over smart phone platforms and able to recognize as widely used for age & gender recognition. Particular attentions also reserved to the evaluation of the system capability to recognize a single age & gender versus all the others. For these purposes, a deep analysis of the literature is provided and state-of-the-art approaches and age & gender related features are evaluated. In more detail, to capture age & gender information, 182 different features related to speech signals' prosody and spectrum shape are used; the classification task is performed by adopting the Support Vector Machine (SVM) approach.

2.2.Features Extraction:

Many different speech feature extraction methods over the years. Methods are distinguished by the ability to use information about human auditory processing and perception, by the robustness to distortions, and by the length of the observation window. Due to the physiology of the human vocal tract, human speech is highly redundant and has several speaker-dependent features, such as pitch, speaking rate and accent. An important issue in the design of a speech age & gender recognition system is the extraction of suitable features that efficiently characterize different age & genders. Although there are many interesting works about automatic speech age & gender detection [9], there is not a silver bullet feature for this aim. Since speech signal is not stationary, it is very common to divide the signal in short segments called frames, within which speech signal can be considered as stationary. Human voice can be considered as a stationary process for intervals of 20_40 [ms]. If a feature is computed at each frame is called local, otherwise, if it is calculated on the entire speech is named global. There is not agreement in the scientist community on which between local and global features are more suitable for speech age & gender recognition.

2.2.1. Gender Recognition Features:

Together with the Mel Frequency Spectral Coefficients (MFCC) [10], pitch is the most frequently used feature [11] since it is a physiologically distinctive trait of a speaker's gender. Other employed features are formant frequencies and bandwidths, open quotient and source spectral tilt correlates, energy between adjacent formants, fractal dimension and fractal dimension complexity [13], jitter and shimmer (pitch and amplitude micro-variations, respectively), harmonics-to-noise ratio, distance between signal spectrum and formants .

2.3. Database:

The database, also called dataset, is a very important part of a speech age & gender recognizer. The role of databases is to assemble instances of episodic age & genders. It is used both to train and to test the classifier and it is composed of a collection of sentences with different age & gender content.

The most used are:

- Reading-Leeds Database: project begun in 1994 to meet the need for a large, well-annotated set of natural or near-natural speeches orderly stored on computers. The essential aim of the project was to collect speeches that were genuinely age & gender rather than acted or simulated.
- Belfast Database: it was developed as part of a project called Principled Hybrid Systems and Their Application (PHYSTA), whose aim was to develop a system capable of recognizing age & gender from facial and vocal signs.
- CREST-ESP (Expressive Speech Database): database built within the ESP [12]. Research goal was to collect a database of spontaneous, expressive speeches.
- Berlin Speech (BES): this is the database employed in it.

3. MODELS

3.1. Hidden Markov Models:

In the Markov model each state corresponds to one observable event. But this model is too restrictive, for a large number of observations the size of the model explodes, and the case where the range of observations is continuous is not covered at all. The Hidden Markov concept extends the model by decoupling the observation sequence and the state sequence. For each state a probability distribution is defined that specifies how likely every observation symbol is to be generated in that particular state. As each state can now in principle generate each observation symbol it is no longer possible to see which state sequence generated a observation sequence as was the case for Markov models, the states are now hidden, hence the name of the model. A Hidden Markov model can be defined by the following parameters:

- The number of distinct observation symbols M.
- An output alphabet ={ } 1 2 M V v ,v ,...v
- The number of states N.
- A state space Q = {1, 2,...N}

States will usually be indicated by i, j a state that 'the model is in' at a particular point in time t will be indicated by qt. Thus, qt = i means that the model is in state i at time t. A probability distribution of transitions between states { } ij

A = a , where

$$a_{ij} = P(q_{t+1} = j | q_t = i) \quad 1 \le i, j \le N$$

3.2. Deep Neural Network:

DNN is a feed-forward artificial neural network that has more than one hidden layers. Each hidden unit uses a nonlinear function to map the feature input from the layer below to the current unit. They use the traditional logistic function as the mapping function.

y = 1/(1 + e−(b+xw))

where x denotes the input feature, w denotes the weights between connections, b denotes the bias and y denotes the output unit. DNN is capable of modeling very complex and highly nonlinear relationships between inputs and outputs, due to its flexible structure with multiple hidden layers and multiple hidden units.

DNN can be discriminatively trained by back-propagating (BP) derivatives of a cost function that measures the discrepancy between the target outputs and the actual outputs produced for each training case [6]. There are two methods to pre-train a Deep Neural Network, the unsupervised pre-training method [7], and the so called discriminative pre-training method, being a supervised pre-training approach [6].

3.3. Unsupervised pre-training:

Unsupervised pre-training method uses stacked Restricted Boltzmann Machine (RBM) to initialize the Deep Neural Network. RBM is a type of undirected graphical model constructed from a layer of binary stochastic hidden units and a layer of stochastic visible units, which will either be Bernoulli or Gaussian distributed conditional on the hidden or visible units depicts the structure of a RBM. Since in this work, the input of the network are real values, this is use a RBM in which the hidden units are Bernoulli distributed, and the visible units are linear real-valued variables with Gaussian noise [11].

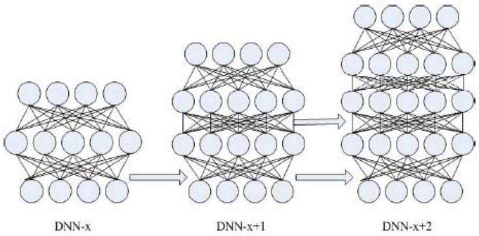

Fig.3.1. Supervised pre-training method - using a network with less hidden layers to initialize a deeper network reconstructed data.

3.4. Discriminative pre-training:

To remedy the modelling inaccuracies in unsupervised pre-training, They follow the alternative by [5] referred to as discriminative pre-training (DPT). The general architecture is shown in Fig. 2. It works as follows, in a first step, a layer wise Back Propagation (BP) is used to train a one-hidden layer DNN to full convergence using every frame's state label, then the softmax layer [5] is replaced by another randomly initialized hidden layer and a new random softmax layer on top, and the network is discriminatively trained again to full convergence. The process is repeated until the desired number of hidden layers is reached. This is similar to a greedy layer-wise training [3], but differs in that of [3] only by the updates of newly added hidden layers. DPT outperforms DNNs with unsupervised pre-training and DNNs without pre-training.

3.5. DNN-HMM:

The key difference between DNN-HMM and GMM-HMM is the using of DNN (instead of GMM) to estimate the observation probabilities. They actually use the DNN to model p(qt|ot), the posterior probability of the state given the observation vector ot , which is possible since p(qt) is easy to estimate from an initial state-level alignment of the training set.

1) DNN-HMM Training Procedure: The detailed training process for Speech recognition is as follows:

a) For each Speech class c(c = 1, . . . , C), a left to right GMM-HMM λc with Q states is trained using the training speech sentences of class c.

b) For each speech sentence O = (o1, o2, . . . , oT) in the training set c, the Viterbi algorithm of the GMM-HMM according is performed on λc to obtain an optimal state sequence (qc1, ..., qcT), and each state qct is assigned a label Li(i ∈ (1, . . . , C × Q)) according to a state label mapping table.

c) All the training sentences, together with their labeled state sequences are used as inputs to train a DNN, whose outputs are the posterior probabilities of the C ×Q output units. The training of the DNN is performed using BP algorithm with (i) the unsupervised pre-training, or (ii) the discriminative pre-training.

2) DNN-HMM Recognition Procedure: In the Speech recognition process, for an input speech sentence O = (o1, o2, . . . , oT), one should estimate the probability p(O|λc) for each Speech class c, and get the final recognition according to it. In GMM-HMM, this probability is obtained via the Viterbi algorithm with it.

In DNN-HMM, adopt the following procedure to calculate the probability p(O|λc).

a) The input feature sequence O is firstly input into the DNN, obtaining the posterior probabilities {p(Li|ot)}i=1,...,C×Q as outputs. Then the posterior probability p(qt = Sck|ot) can be obtained from p(Li|ot), by mapping the label Li to the state k of the model c, using a state-label mapping table.

b) According to the Bayesian principle, this calculate the likelihood probability p(ot|qt) as

$$p(o_t|q_t) = \frac{p(q_t|o_t)p(o_t)}{p(q_t)}$$

The prior probability of each state, p(qt), is calculated from (occurrences of) the training set, and p(ot) can be assigned a constant since the observation feature vectors are regarded as independent of each other.

c) For each Speech model λc, the Viterbi algorithm is performed to calculate the likelihood probability p(O|λc) according to it . However, here the probability bqt (ot) is replaced by p(ot|qt) .

4. Proposed Work

In many criminal cases, evidence might be in the form of recorded conversations, possibly over the telephone. Therefore, law enforcement agencies have been concerned about accurate methods to profile different characteristics of a speaker from recorded voice patterns, which facilitate to identify him/her or at least narrow down the number of suspects. Here we propose a new gender and age group recognition approach based on Hidden Markov Model (HMM). First, an acoustic model is trained for all speakers in a training database including male and female speakers of different age. Finally, Supervised HMM is applied to detect the gender and age group of unseen test speakers.

Proposed Work:

1. Designing a new HMM-based approach for speaker, gender and age estimation, which improves the accuracy of the state-of-the-art speaker age estimation methods with statistical significance.
2. Analyzing the effect of major factors influencing the automatic gender and age estimation systems.

Age and Gender Recognition architecture:

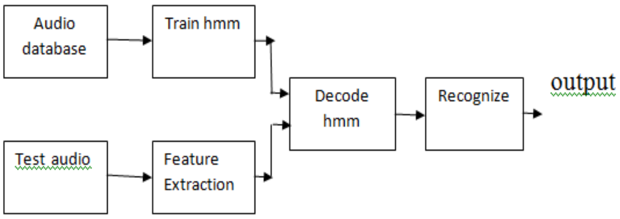

Fig: Reorganization of Age & Gender architecture.

A new gender and age group recognition approach based on Hidden Markov Model (HMM).The audio database are used in some samples audio recorded and the used as the six classification. After First, an acoustic model is trained for all speakers in a training database including male and female speakers of different age. Finally, Supervised HMM is applied to detect the gender and age group of unseen test speakers. Then the feature extraction techniques are used . last, that the the check decode hmm models and recognize the speaker in him /her gender or afe of that gender.

5. Conclusion

This concept of many cases, evidence might be in the form of recorded conversations. The accurate methods to profile different characteristics of a speaker from recorded voice patterns, which facilitate to identify him/her or at least narrow down the number of suspects. Here we propose a new gender and age group recognition approach based on Hidden Markov Model (HMM).Audio database is trained using HMM and some feature extraction from test audio are decoded . If the match is found age and gender is recognize.

The benefit of this HMM model is analyzing the effect of major factors influencing the automatic gender and age estimation systems.

References

[1] Longfei Li, Yong Zhao, Dongmei Jiang and Yanning Zhang, Fengna Wang, Isabel Gonzalez, Enescu Valentin and Hichem Sahli,”*Hybrid Deep Neural Network - Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition”, IEEE* 2013.

[2] Levent M. Arslan, *Member, IEEE,* and John H. L. Hansen, *Senior Member, IEEE,” Selective Training for Hidden Markov Models with Applications to Speech Classification”.* IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 7, NO. 1, JANUARY 1999.

[3] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly,Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury ,*“Deep nural network for aucostic modeling in speech recognition”* IEEE SIGNAL PROCESSING MAGAZINE , NOVEMBER 2012.

[4] Rubén Solera-Ureña, *Member, IEEE*, Ana Isabel García-Moral, Carmen Peláez-Moreno, *Member, IEEE*, Manel Martínez-Ramón, *Senior Member, IEEE*, and Fernando Díaz-de-María, *Member, IEEE,” Real-Time Robust Automatic Speech Recognition Using Compact Support Vector Machines*”, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 4, MAY 2012.

[5] Jung-Hui Im and Soo-Young Lee,” *Unified Training of Feature Extractor and HMM Classifier for Speech Recognition*”, IEEE SIGNAL PROCESSING LETTERS, VOL. 19, NO. 2, FEBRUARY 2012.

[6] IGOR BISIO, ALESSANDRO DELFINO, FABIO LAVAGETTO, MARIO MARCHESE, AND ANDREA SCIARRONE ,”*Gender-Driven Emotion Recognition Through Speech Signals for Ambient Intelligence Applications*”, VOLUME 1, NO. 2, DECEMBER 2013.

[7] De Zhang , Yunhong Wang and Bir Bhanu ,*” Age Classification Based on Gait Using HMM*”, 2010 International Conference on Pattern Recognition, IEEE 2010.

[8] Joseph Picone , *“Continuous Speech Recognition Using Hidden Markov Models”,* IEEE ASSP MAGAZINE JULY 1990.

[9] Nobuaki MINEMATSU, Mariko SEKIGUCHI , Keikichi HIROSE,” *AUTOMATIC ESTIMATION OF ONE’S AGE WITH HIS/HER SPEECH BASED UPON ACOUSTIC MODELING TECHNIQUES OF SPEAKERS”,IEEE 2002.*

[10] Hui Jiang, *Member, IEEE*, Xinwei Li, and Chaojun Liu, *Member, IEEE* , “Large Margin Hidden Markov Models for Speech Recognition” , IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 14, NO. 5, SEPTEMBER 2006.

[11] Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B. Mariño, “*Speech Emotion Recognition Using Hidden Markov Models”* , Eurospeech 2001.

[12] Florian Metze, Jitendra Ajmera, Roman Englert, Udo Bub ,Felix Burkhardt, Joachim Stegmann, Christian M¨uller ,Richard Huber ;Bernt Andrassy, Josef G. Bauer, Bernhard Little, “*COMPARISON OF FOUR APPROACHES TO AGE AND GENDER RECOGNITION FOR TELEPHONE APPLICATIONS*”.

[13] Kumar Rakesh , Subhangi Dutta and Kumara Shama ,“ *GENDER RECOGNITION USING SPEECH PROCESSING TECHNIQUES IN LABVIEW* ”, International Journal of Advances in Engineering & Technology, May 2011.

[14] Ana Isabel García-Moral, Rubén Solera-Ureña, *Student Member, IEEE*, Carmen Peláez-Moreno, *Member, IEEE,*and Fernando Díaz-de-María, *Member, IEEE* , “Data Balancing for Efficient Training of Hybrid ANN/HMM Automatic Speech Recognition Systems” , IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 3, MARCH 2011.

[15] Junichi Yamagishi, *Member, IEEE*, Takashi Nose, Heiga Zen, Zhen-Hua Ling, Tomoki Toda, *Member, IEEE,*Keiichi Tokuda, *Member, IEEE*, Simon King, *Senior Member, IEEE*, and Steve Renals, *Member, IEEE , “Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis*” , IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 17, NO. 6, AUGUST 2009.

[16] Wooil Kim, *Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*, “Time–Frequency Correlation-Based Missing-Feature Reconstruction for Robust Speech Recognition in Band-Restricted Conditions” , IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 17, NO. 7, SEPTEMBER 2009.

[17] Timothy J. Hazen, *Member, IEEE* , “*Visual Model Structures and Synchrony Constraints for Audio-Visual Speech Recognition*”, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 14, NO. 3, MAY 2006.

[18] Oliver Watts, Junichi Yamagishi, *Member, IEEE*, Simon King, *Senior Member, IEEE*, and Kay Berkling, *Senior Member, IEEE , “Synthesis of Child Speech With HMM Adaptation and Voice Conversion”,* IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 5, JULY 2010.

[19] Sven Ewan Shepstone, *Member* IEEE, Zheng-Hua Tan, *Senior Member* IEEE and Søren Holdt Jensen, *Senior Member* IEEE , “*Audio-based Age and Gender Identification to Enhance the Recommendation of TV Content”,IEEE 2013.*

[20] Jen-Chun Lin, Chung-Hsien Wu, *Senior Member, IEEE*, and Wen-Li Wei , *“Error Weighted Semi-Coupled Hidden Markov Model for Audio-Visual Emotion Recognition”,* IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 14, NO. 1, FEBRUARY 2012.