Automatic Speaker Age Estimation and Gender Dependent Emotion Recognition

Shivaji J. Chaudhari
PG Scholar
Department of Computer Engineering
Dr. D. Y. Patil School of Engineering and Technology, Pune, India

Ramesh M. Kagalkar
Assistant Professor
Department of Computer Engineering
Dr. D. Y. Patil School of Engineering and Technology, Pune, India

ABSTRACT
Gender-dependent age, emotions (stress and feeling) are speaker qualities being examined in voice-based speaker voice processing system, these qualities or characteristics play important role in the Human and Computer Interaction (HCI). Grouping speaker attributes is an important task in the fields of Voice Processing, Sspeech Synthesis, Forensics, Language Learning, Assessment, furthermore Speaker Identification to increase the performance of voice processing system, also enhance the emotion identification depend on two-stage recognizer that identify the gender of speaker male or female And then recognize the emotions. Noise elimination technique eliminate the noisy sound from audio clip. Mel-Frequency Cepstral Coefficients (MFCCs) is a feature extraction technique broadly utilized as an important part of Automatic voice processing for unique feature extraction. The system contains the Gaussian mixture model (GMM) supervectors as features for a support vector machine (SVM) for the large data classification into different group based on the margin between the two different classes. Principal component analysis (PCA)is used to reduce the large dimension size of feature vector to improve the system performance and accuracy in HCI.

Keywords
Human and Computer Interaction (HCI); Mel-Frequency Cepstral Coefficients (MFCCs); Support Vector Machine (SVM); Gaussian Mixture Model (GMM); Principal Component Analysis (PCA).

1. INTRODUCTION
Many characteristics can be used for human to computer interaction (HCI) like human face, human voice, thumb impression, iris etc. It is very important to increase efficiency and performance of the human to computer interaction by identifying the human information using the human voice, each human has a unique voice and different features of voice used in the speaker identification, speaker verification, Identifying the information of speaker like gender, age, emotion, utterance, language etc. There is an increasing need to know not only what information a user speaking, but also how it is being spoken and its meaning. There are numerous speaker qualities that have helpful applications. The most prevalent of these incorporate gender, age, health, language, dialect, accent, emotional state, and attention state.

Speech synthesis speed changes according to the user's age and speech recognition systems can select a more appropriate language model. In call centers, classification or order of speakers into age classifications is utilized to perform client profiling, which is a premise for essential applications like particular market research (for the young), targeted advertising, and service customization. Several speech samples based age and gender estimation systems were proposed, using and combining different kinds of acoustic features and classification algorithms. Also Emotion recognition has many applications that appear in telecommunications, human robotic interfaces, smart call centers, and intelligent spoken tutoring systems.

Voice or speech processing mainly depends on the feature of voice signal which are extracted using the feature extraction algorithm. In current voice and speech processing system, MFCC has been widely used. The MFCC feature extraction algorithm gives high performance and accuracy in feature extraction of the voice signal. Gender dependent age estimation and emotion identification approach are used in this work. The estimation of a speaker's age is often performed based on groups of speakers in groups with a wider age range to improve the result of age and emotion identification. SVM is used for the classification of age group and emotion, for the age identification created seven age groups (Male_Young, Male_Adult, Male_Senior, Female_Young, Female_Adult, Female_Senior, and Child) to find the speaker age group and precise age using the exact matching. SVM work based on the GMM supervector model. For emotion identification consider six emotions happy, fear, sad, disgust, surprise and neutral, for each emotion one GMM model is created for supervector of signal feature which are extracted using the MFCC technique. PCA is an orthogonal linear transformation dimension reduction technique to reduce the high dimension of feature's vector. Supervised PCA (SPCA) is a PCA variant where the feature vectors are preprocessed before applying PCA on supervector. This system divides into two phases that is training phase and testing phase and above mentioned techniques used in both training phase and testing phase. Lastly, in voice processing performance decreases dramatically in noisy environments that may consist of a wide range of noise sources such as music, background noise, car, factory, cafeteria noise, etc. and where very low signal-to-noise ratios are encountered. Besides acoustical model adaptation and robust feature extraction approaches, noise reduction is considered as an effective approach for voice processing robust ASR system.

2. LITERATURE SURVEY
Gil Dobry, Ron M. Hecht [1] presents a novel dimension reduction method which improves the accuracy and the efficiency of speakers age estimation systems based on speech signal. Two different gender based age estimation approaches were implemented, the first age group (Senior, Adult, and Young) classification, and the second, accurate age estimation using regression technique.

Hugo Meinedo1, Isabel Trancoso [2] present gender detection is a very useful task for a wide range of applications. In the Spoken Language Systems lab of INESC-ID, the Gender Identification module is one of the basic components of our Voice processing system, where it is mainly used for speaker clustering, in order to avoid mixing speakers from different genders in the same cluster. Gender information (male or female) is also used for building gender-dependent acoustic models for speech recognition.

Mohamad Hasan Bahari, Hugo Van Hamme [3] introduces a new gender detection and an age estimation approach. To create this strategy, after determining an acoustic model for all speakers of the database, Gaussian mixture weights are extracted and concatenated to create a supervector for each speaker. Then, hybrid architecture of WSNMF and GRNN is developed using the supervectors of the training data set.

Ismail Mohd, Adnan Shahin [4] focused on improving emotion identification performance and accuracy based on a two-stage recognizer that is composed of gender recognizer followed by an emotion recognizer. This work is a gender dependent, text-independent and speaker-independent emotion recognizer. Both HMMs and SPHMMs have been used as classifiers in the two-stage architecture. The databases are collected database using different emotions sample and Emotional Prosody Speech and Transcripts database.

Chul Min Lee and Shrikanth Narayanan [5] explore the detection of domain-specific emotions using language and discourse information in conjunction with acoustic correlates of emotion in speech signals. The main focus is on detecting negative and non-negative emotions (happy or unhappy) using spoken language data obtained from a call center application and from another source.

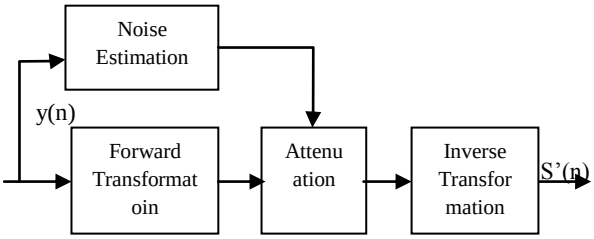
Tetsuya Takiguchi and Yasuo Ariki [8] investigate robust feature extraction using kernel PCA instead of DCT, where kernel PCA is applied to the mel-scale filter bank output, because the expectation is that kernel PCA will project the main speech element onto low-order features, while noise (reverberant) element onto high-order ones. The use of kernel PCA provides better performance for reverberant speech.

3. SYSTEM DESIGN

A First noise removal process applied to remove the noise from the audio file. The noise removal process is generally called signal denoising. It is also called an estimator because it forms an estimate S'(n) of the underlying signal S(n). In the case where S(n) is a speech signal, the noise removal process is referred to as speech enhancement. Denoising consists of two basic parts, namely a noise estimation process and a denoising algorithm.

1. The noise estimation process usually uses an algorithm that estimates the noise spectrum. White Gaussian noise is generated and then added to the clean signal to produce the noisy signal. The noise spectrum can, therefore, be directly calculated from the noise, instead of being estimated from the real-world data.

2. The denoising algorithm is the basic mechanism of denoising. It relies on the noise estimate and is comprised of three parts and all the three are derived, y(n) is noisy signal and S'(n) is clean signal.



3.1 MFCC Feature Extraction Algorithm

The extraction of the best feature or parametric representation of acoustic signal is a critical task to produce a better identification and recognition performance and accuracy. The proficiency of this stage is critical for the following stage since it affects its behavior[8].

The Mel-Frequency Cepstral Coefficients(MFCC) feature extraction method is a leading approach for speech feature extraction and current research aims to identify performance enhancements. MFCC is a popular technique because it is based on the known variation of the human ears critical frequency bandwidth. Following some steps are involved in MFCC,

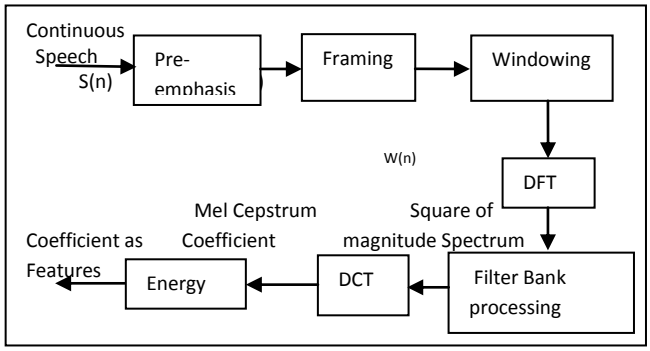


Fig 1: MFCC feature extraction technique

1) Pre-emphasis: This step processing of a signal through a filter which emphasizes higher frequencies. This process obtains similar amplitude. For higher frequency signal.
2) Framing: The process of creating the segment of speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec for windowing.
3) Windowing: Window is implicitly used when a sequence of N sample frame is retrieved from the signal. Windowing reduces the distortion, but it increases the signal shape alternation, most used window shape is hamming Hamming window is used as window shop for considering the next

block in the feature extraction processing chain and integrates all the closest frequency lines.

4) Discrete Fourier Transform: DFT is the standard method of spectral analysis relies on the Fourier transform, convert each frame of N samples from time domain into the frequency domain. DFT help reduces computational complexity to order N log(N).
5) Filter Bank Processing: Spectral feature is generally obtained as exit of filter banks, which properly integrates the spectrum at the defined frequency ranges.
6) DCT: Discrete cosine transform (DCT) produce the highly uncorrelated feature, DCT is used to achieve Mel-cepstrum coefficients; we are going to select 13 coefficients for recognition system.
7) Energy: The voice signal and the frame changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time Cepstral coefficient are joined to energy coefficient using the logarithm of the energy of the frame.

3.2 System Architecture

Proposed system architecture divided into two phases.

1. Training phase
2. Testing phase

There are two important datasets used in the system one is the training dataset contain different types of data, information sources and characteristics of available training data also in different environmental condition and medium. And the second is testing dataset is a new spec file to identify an age and emotion depends on the gender based on the training dataset and the variance in the dataset. MFCC feature extraction algorithm work at both side training and testing. Emotion recognition performance based on the two-stage approach (gender dependent emotion recognizer) has been significantly improved compared to that based on emotion recognizer without gender information.

Training Phase **a**ge detection is achieved using the age group estimation and precise age identification. While training the system the training dataset (T_D) is divided according to different gender dependent the age group like M_Y, M_A, M_S, F_Y, F_A, F_S, and also Training dataset is divided as per emotions type is the speaker speaking in the happy, sadness, disgust, fear, surprise and neutral (S, H, D, S, F, N) emotions.

$T_D = \{M_Y, M_A, M_S, F_Y, F_A, F_S\}$

Feature_database=MFCC (S_1, S_2 ……S_n)

Feature _database=MFCC(S, H, D, S, F, N)

At Training phase: Input

Training Dataset = (S_1, S_2……..S_N)

V= (F1, F2……..F_N)

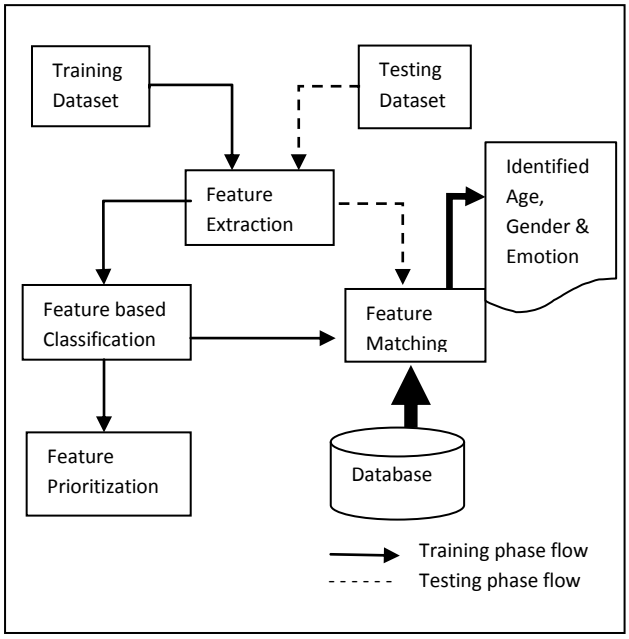


Fig 2: System architecture

At Testing Phase input is new speaker voice applied to feature extraction algorithm (MFCC) using the testing dataset and compare the matching result between the extracted features of training datasets according to the system requirement to train the system.

V is the extracted feature's vector, using the feature vector by concatenating all V create the supervector feature presentation using the SVM and GMM model for each type speaker characteristic the template will do that for you[9].

For the data or feature classification and supervector representation use the Gaussian Mixture model (GMM) supervectors as features for a support vector machine (SVM) model. SVM is an important technique for pattern classification. SVM creates the SVM model for the each group of age and emotion, an SVM is a discriminative classifier, An SVM classifier classifies the age of the speaker into divided seven age groups and SVM regression estimate the exact matching age of the testing dataset input models the boundary between the above mentioned groups[10].

SVM creates the supervector for feature representation using the GMM, standard MFCC feature vector with the GMM component mean features, reduced error rates on both tasks are achieved. Clustering perspective, most voice signal data cannot be adequately modeled by a single cluster Gaussian model. However, they can often be accurately modeled via a Gaussian mixture model (GMM) i.e., data distribution can be expressed as a mixture of multiple normal distributions.

Building the supervector for Each GMM model is represented by GMM supervector, formed by concatenating all the N Gaussians means,

$$S_v = \text{SVMGMM\_model}(v_1, v_2 \ldots\ldots v_N) \quad (1)$$

Where v_i is the mean vector of the ith Gaussian. The training supervectors are formed using the MAP adapted GMM models.

The GMM supervectors creation consists of several parts. First, the UBM is trained over a large speech database, then a GMM MAP adaptation is performed for each speaker to obtain a speaker specific model based on which the GMM supervector is created and finally, the dimension reduction is applied, PCA is the Dimension reduction technique orthogonal linear transformation that projects a set of vectors to a new basis whose components are linearly uncorrelated and arranged in a decreasing order of variance. This method assumes that most of the relevant information is found in the first coordinates of the projected space, since they contain most of the variance. A dimension reduction is then made by using only the first coordinates of the projected vectors such that and is the original feature vector dimension. Supervised PCA (SPCA) is a PCA variant where the feature Vectors are preprocessed before applying PCA. The Preprocessing consists of screening out coordinates having the lowest correlation with labels. DIM_{RM} is the Dimension reduction matrix, PCA applied on the S_v to reduce large dimensions of feature vector[1].

At Training side: Input: S_v
Output: DIM_{RM}
DIM_{RM}=PCA (S_v)

SVM classifiers are trained over the training data. SVM cross validation used find the relevant or a matching group for the testing data; find the score between the training data and testing data. SVM_C is the SVM classifier and S_N is the Score Normalization to calculate the matching score.

S_C = Feature {M_Y, M_A, M_S, F_Y, F_A, F_S}&{S, H, D, S, F, N}

S_C=SVM_C (DIM_{RM})

S_N=SVM_C(S_C)

Final Output: AGEGrp

AGE_{Grp} =SVM_C(S_C; S_N)

S_N is the score normalization to normalize the estimate mapping score using the SVM regression method and finds the precise age of testing phase speaker gender dependent age and emotion.

4. DATABASE

In the previous work Gil Dobry, Ron M. Hecht [1] used the voice or speech data used to train the UBM model was taken from the LDC's Switchboard corpus annotated with age and gender labels[13]. Classification Database While training the voice processing system Training dataset is divided into seven groups in the following Table .I.

Table 1. Classification of Speaker's Dataset

Classified Dataset Name	Age Range (Year)	Notation
Child	0-12	C
Male young	13-30	M_Y
Male Adult	30-50	M_A
Male Senior	>50	M_S
Female young	13-30	F_Y
Female Adult	30-50	F_A
Female Senior	>50	F_S

Classifications of training data sets, shown in Table I, were selected such that there is no speaker overlap between them. Collected the audio file from different above mentioned age group type and in each emotion and train the system on well classified database to increase the performance and accuracy. The database is collected from different source like an audio clip from movie, news channel, some dialog, and recorded audio clip.

5. RESULTS

The aim of noise reduction or elimination in speech is to minimize the noise level without distorting the speech signal quality. In many speech communication applications, the recorded and transmitted speech signals contain a considerable amount of acoustic background noise. Noise eliminaton technique reduce noise at a great level and give the final output as clean and noise free signal from noised voice signal shown in the Fig 4 and Fig 5

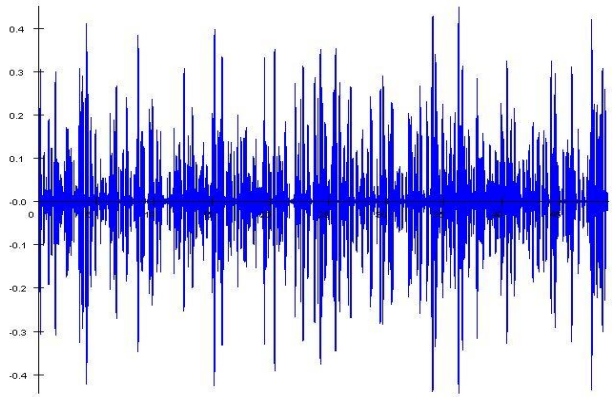


Fig 3: Noised voice signal

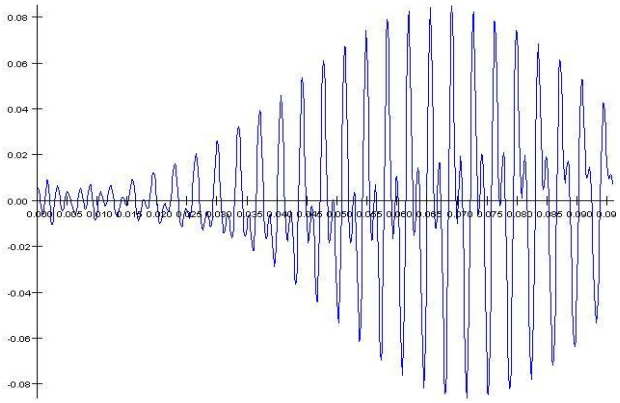


Fig 4:Noise free signal

MFCC plays important role in proposed system; feature of training data is extracted using the MFCC feature extraction technique. MFCC is a highly efficient algorithm for feature extraction. Input to the system is the audio file from any one of the age group and emotional stale from the table I with known information like age group, age and emotional state extract the feature and generate an XML file of that audio file, xml file contains the many feature it plays an important role in the training phase system and also the

testing phase to extract the feature of the newly arrived audio file.

Mainly database divided into two main database that is Male and Female and again the collected data divided into above mentioned seven sub database for efficient classification. Emotion identification performance based on a two-stage recognizer that is composed of gender recognizer followed by an emotion recognizer. And two databases, one is the collected database and second is emotional prosody speech and Transcripts database. For each type of emotion training dataset is collected and also from the emotional prosody speech dataset [4].

The database is divided into the seven groups as per the age group of human as shows in the table I to increase the automation performance in real time.

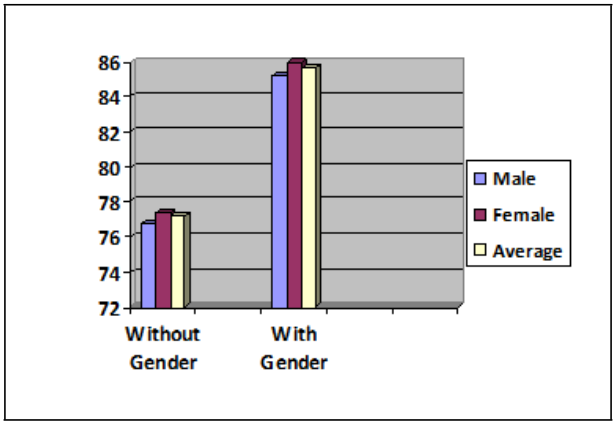


Fig 5: Difference between without gender dependant and with gender dependant system

A system giving high performance and accuracy as compare to exiting the system, because the proposed system mainly using the gender dependant dataset and dimension reduction technique to reduce the large size of feature vector complexity Fig 3 gives the difference between the without gender dependant system and with a gender dependant system for performance and accuracy.

Generate the model for each group using the SVM and GMM modeling technique for the classification of speaker into different group using the SVM classifiers and finds the exact match for the testing dataset feature with training dataset feature, and applying the Dimension reduction technique.

6. CONCLUSION

The proposed system develops to improving the human to computer interaction. The system identifies the speaker age, emotion depend on gender, this speakers characteristics is helpful in many applications like for advertisement, targeting to particular people, automatically identification of this features age, emotion to provide facility and service to customer in a call center, also the speaker's voice can be used as the biometric security. The system performing the noise elimination from noisy voice signal, reducing the complexity of processing large feature vector, supervector, by reducing the dimension of matrix and manage the result in efficiently to increase the accuracy and efficiency of system output.

Future scope of this system is to achieve greater performance and accuracy and also perform the processing on the multiple speakers simultaneously at the same time.

7. ACKNOWLEDGMENTS

It gives me immense pleasure to express my deepest sense of gratitude and sincere thanks to my highly respected and esteemed guide Mr. Ramesh Kagalkar, Asst. Professor Computer Engineering Deptt, DYPSOET, Pune and the Ms. Arti Mohanpurkar HOD of computer department DYPSOET, Pune.and also thanks to all staffs of Computer Engineering Department, DYPSOET, Pune for their valuable guidance, encouragement and help for completing this work. Their useful suggestions for this whole work and co-operative behavior are sincerely acknowledged

8. REFERENCES

[1] Gil Dobry, Ron M. Hecht, Mireille Avigal and Yaniv Z, SEPTEMBER, 2011. "Supervector Dimension Reduction for Efficient Speaker Age Estimation Based on the Acoustic Speech Signal",IEEE transction VOL. 19, NO. 7,

[2] Hugo Meinedo1 and Isabel Trancoso, 2008 "Age and Gender Classification using Fusion of Acoustic and Prosodic Features",Spoken Language Systems Lab, INESC-ID Lisboa, Portugal, Instituto Superior Tecnico, Lisboa, Portugal.

[3] Ismail Mohd Adnan Shahin, 2013"Gender-dependent emotion recognition based on HMMs and SPHMMs",Int J Speech Technol, Springer 16:133141.

[4] Mohamad Hasan Bahari and Hugo Van h, ITN2008 "Speaker Age Estimation and Gender Detection Based on Supervised Non-Negative Matrix Factorization", Centre for Processing Speech and Images Belgium.

[5] Chul Min Lee and Shrikanth S. Narayanan, 2005 "Toward Detecting Emotions in Spoken Dialogs", IEEE transaction 1063-6676.

[6] Tetsuya Takiguchi and Yasuo Ariki, 2006 "Robust feature extraction using kernel PCA",Department of Computer and System Engg Kobe University, Japan, ICASSP 1-4244-0469.

[7] Michael Feld, Felix Burkhardt and Christian Muller, 2010 "Automatic Speaker Age and Gender Recognition in the Car for Tailoring Dialog and Mobile Services",German Research Center for Artificial Intelligence, INTERSPEECH.

[8] Md Afzal Hossan, Sheeraz Memon and Mark A Gregory, "A Novel Approach for MFCC Feature extraction", RMIT university, Melbourne, Australia, IEEE, 2010.

[9] Ruben Solera-Ure, 2008 "Real-time Robust Automatic Speech Recognition Using Compact Support Vector Machines",TEC 2008-06382 and TEC 2008-02473.

[10] Marc Ferras, Cheung-Chi Leung, Claude Barras, and Jean-Luc Gauvain, 2010 "Comparison of Speaker Adaptation Methods as Feature Extraction for SVM-Based Speaker Recognition",IEEE Transaction 1558-7916.

[11] Wei HAN and Cheong fat CHAN, 2006 "An Efficient MFCC Extraction Method in Speech Recognition",Department of Electronic Engineering, The Chinese University of Hong Kong Hong Kong, 7803-9390-06/IEEE ISCAS.

[12] Arif Ullah Khan and L. P. Bhaiya, 2008 "Text Dependent Method for Person Identification through Voice Segment",ISSN- 2277-1956 IJECSE.

[13] Felix Burkhardt, Martin Eckert, Wiebke Johannsen and Joachim Stegmann, 2010"A Database of Age and Gender Annotated Telephone Speech", Deutsche Telekom AG Laboratories, Ernst-Reuter-Platz 7, 10587 Berlin, Germany.

[14] Lingli Yu and Kaijun Zhou, March 2014, "A Comparative Study on Support Vector Machines Classifiers for Emotional Speech Recognition", Immune Computation (IC) Volume:2, Number:1.

[15] Rui Martins, Isabel Trancoso, Alberto Abad and Hugo Meinedo, 2009, "Detection of Childrens Voices", Intituto Superior Tecnico, Lisboa, Portugal INESC-ID Lisboa, Portugal.

[16] Chao Gao, Guruprasad Saikumar, Amit Srivastava and Premkumar Natarajan, 2011, "OpenSet Speaker Identification in Broadcast News", IEEE 978-1-4577-0539.