A Novel Method for Indexing and Retrieval of Speech using Gaussian Mixture Model Techniques

R. Thiruvengatanadhan
Assistant Professor
Department of CSE
Annamalai University

P. Dhanalakshmi
Associate Professor
Department of CSE
Annamalai University

ABSTRACT
Large speech databases such a Television broadcasts, TV programs, radio broadcasts, CDs and DVDs are available online these days. Research related to speech indexing and retrieval has received much attention in recent days due to the huge multimedia data storage capabilities. The goal of speech indexing and retrieval system is to provide the user with capabilities to index and retrieve the speech archives in an efficient manner. In this paper, we propose a method for indexing and retrieval of the speech. The speech activity is identified using voice activitiy detection and each complete speech dialogue is separated into individual words by marking each word¨s segment through the Root Means Square (RMS) energy envelope. Then the features namely Perceptual Linear Prediction (PLP), Power Normalized Cepstral Coefficient (PNCC), Subband Coding (SBC) and Sonogram extracted from each of the individual word. For retrieval, a novel method is proposed using Gaussian mixture models. The probability that the query feature vector belongs to the Gaussian is computed. The average Probability density function is computed for each of the feature vectors in the database and the retrieval is based on the highest probability.

Keywords
Voice Activity Detection, Root Mean Square, Gaussian Mixture Model, Probability density function

1. INTRODUCTION
Large speech databases such a Television program, radio broadcasts, CDs and DVDs are available online and the necessity to organize such huge databases becomes essential these days [1]. As Large Vocabulary Continuous Speech Recognition (LVCSR) is imperfect, automatic speech transcripts contain errors. Due to storage constraints, research related to speech indexing and retrieval has received much attention [2]. As storage has become cheaper, large collection of spoken documents is available online, but there is a lack of adequate technology to explain them. Manual transcription of speech is costly and also has privacy constraints [3]. Hence, the need to explore automatic approaches to search and retrieve spoken documents has increased. Moreover, a wide variety of multimedia data is available online and paves the way for development of new technologies to index and search such media [4].

In this work, voice activity detection is performed using the Root Mean Square (RMS) energy of the signal to form the temporal envelope for isolating the individual words from the continuous speech signal.

Figure 1 shows the overall architecture of the proposed method for speech indexing and retrieval. A novel method is developed for speech indexing where a GMM is constructed for each query feature vector and the probability density function is computed for all isolated words in the speech database against the GMM of the speech query.

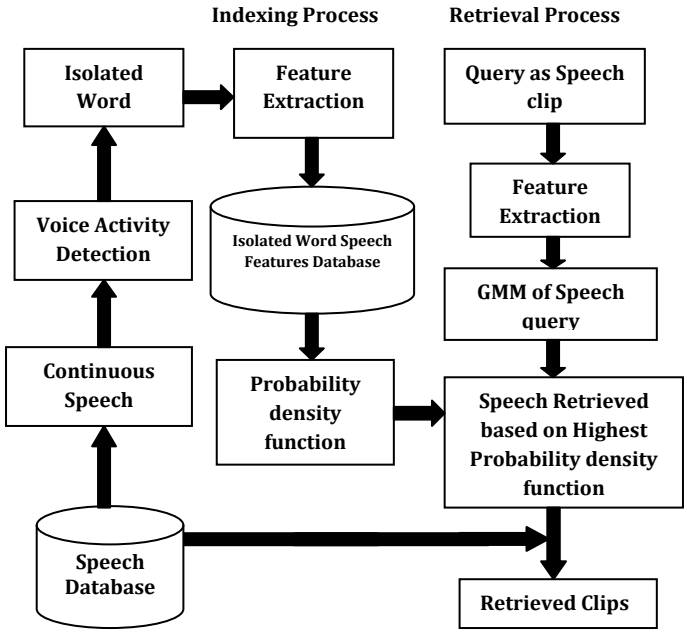

Fig 1: Proposed Methods for Speech Indexing and Retrieval System.

2. VOICE ACTIVITY DETECTION (VAD)
VAD is a technique for finding voiced segments in speech and plays an important role in speech mining applications. VAD ignores the additional signal information around the wordunder consideration [5]. It can be also viewed as a speaker independent word recognition problem. The basic principle of a VAD algorithm is that it extracts acoustic features from the input signal and then compares these values with thresholds usually extracted from silence. Voice activity is declared if the measured values exceed the threshold. Otherwise, no speech activity is present [6].

It identifies where the speech is voiced, unvoiced or sustained and makes smooth progress of the speech process. Figure 2 shows the isolated word separation.

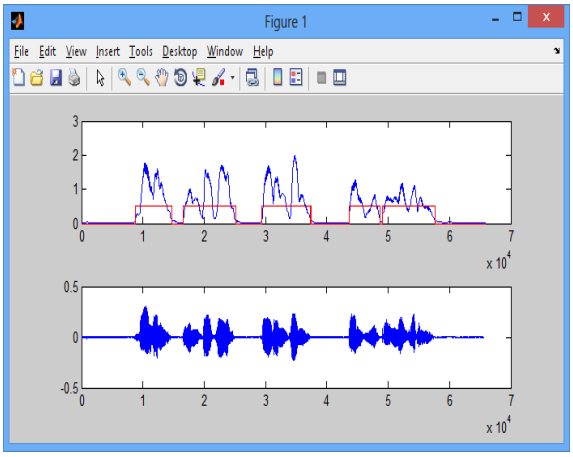

Fig 2: Isolated Word Separations

A frame size of 20 ms, with an overlap of 50%, is considered for VAD. RMS is extracted for each frame using Equation (1). In this work 0.4 is the threshold, below which the frames are considered to be silent frames.

The RMS is computed as follows.

$$\text{RMS} = \sqrt{\sum_{n=0}^{m} x^2(n)} \qquad (1)$$

where, x(n) represents the discrete time signal, m represents the number of samples in a frame.

3. ACOUSTIC FEATURE EXTRACTION

An important objective of extracting the features is to compress the speech signal to a vector that is representative of the meaningful information it is trying to characterize. In these works, acoustic features namely PLP, PNCC, SBC and sonogram features are extracted.

3.1 Perceptual Linear Prediction

Hermansky developed a model known as PLP. It is based on the concept of psychophysics theory and discards unwanted information from the human pitch [7]. It resembles the procedure to extract LPC parameters except that the spectral characteristics of the speech signal are transformed to match the human auditory system.

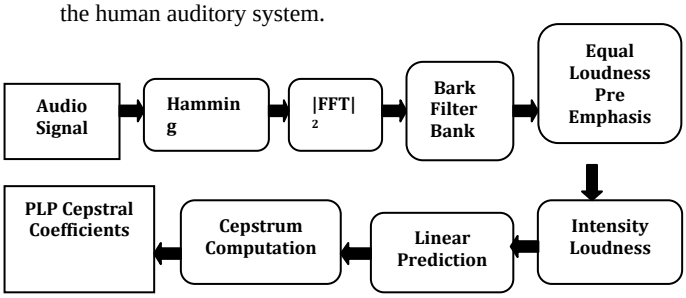

Fig 3: PLP Parameter Computations.

PLP is the approximation of three aspects related to perceptron namely resolution curves of the critical band, curve for equal loudness and the power law relation of intensity loudness. The process of PLP computation is shown in Figure 3. The audio signal is hamming windowed to reduce discontinuities. The Fast Fourier Transform (FFT) transforms the windowed speech segment into the frequency domain [8]. The power spectrum for the Fourier transform coefficients is calculated using Equation (2).

$$P(\omega) = \text{Re}(S(\omega))^2 + \text{Im}(S(\omega))^2 \qquad (2)$$

where P(ω) is the Power spectrum, Re(S(ω)) represents the real part and Im(S(ω)) represents the imaginary part of Fourier transform

The bark transformation is used to warp the spectrum P(ω) along its frequency axis ω into the bark frequency Ω in Equation (3).

$$\Omega(\omega) = 6\ln\left[\frac{\omega}{1200\pi} + \sqrt{\left(\frac{\omega}{1200\pi}\right)^2 + 1}\right] \qquad (3)$$

Critical band is the frequency bandwidth created by the cochlea, which acts as an auditory filter. The cochlea is the hearing sense organ in the inner ear. Bark scale corresponds to 1 to 24 critical bands. The power spectrum of the critical band masking curve and auditory warped spectrum are convoluted to simulate the human hearing resolution. The equal loudness pre-emphasis needs to compensate the unequal perception of loudness at varying frequencies.

In this work, a 9th order LP analysis is used to approximate the spectral samples and hence obtained a 9-dimensional feature vector for a speech signal of frame size of 20 milliseconds is obtained.

3.2 Power Normalized Cepstral Coefficients

Power Normalized Cepstral Coefficients (PNCC) is well known for the accuracy of automatic speech recognition systems, even in high-noise environments [9], [10].

PNCC algorithm a pre-emphasis is done to enhance the energy of the high frequency signal. Each frame is hamming windowed and then Short Time Fourier Transform (STFT) is computed. The magnitude squared STFT is computed for frequencies which are positive. Spectral power is computed by combining the frequency response with filter banks, which are gammatone shaped [11]. Short time spectral power is the squared gammatone summation. Spectral smoothing across the channels is helpful in processing schemes such as PNCC where there are nonlinearities that vary in their effect from channel to channel. Mean power normalization is performed to minimize the impact of amplitude scaling in PNCC by dividing the incoming power with a running average of the overall power. A nonlinear function describes the relationship between incoming signal amplitude in a given frequency channel and the corresponding response of the processing mode. The power-law function is chosen for PNCC processing. Inverse Discrete Fourier Transform (IDFT) is applied. Finally mean normalization is performed.

The PNCC features are computed using the algorithm mentioned above. Typically, the first 13 PNCCs are used as acoustic features.

3.3 Subband Coding

Stress is termed as perceptually induced deviation in the production of speech from that of the conventional production of speech. The excitation plays a vital role in determining the stress information present in the speech signal rather than vocal tract in the linear modeling [12]. Based on the knowledge of stress and its types, the additional information has been incorporated into the speech system which increases the performance of the system.

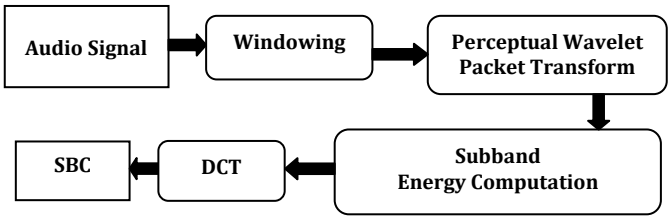

Fig 4: SBC Feature Extractions.

Subband Coding (SBC) incorporates the excitation in the speech signal whereas mel-scale analysis incorporates properties of human auditory system [13]. In this work a set of features are extracted based on the multi-rate subband analysis or wavelet analysis of stressed speech. The Discrete Cosine Transform (DCT) of subband energy for each frame in the speech signal is extracted using perceptual wavelet packet transform.

This wavelet packet transform can be achieved by two filter banks: low pass filter and high pass filter respectively [14]. The current work is focused to obtain the high energy information in the cascaded filter bank with its wavelet packet tree [15]. Figure 4 shows the block diagram of the extraction procedure of SBC feature. As per the SBC features are extracted from audio signal. In this work, the number of SBC parameters is chosen as 12.

3.4 Sonogram

Pre-emphasis is performed for the speech signal followed by frame blocking and windowing. The speech segment is then transformed using FFT into spectrogram representation [16]. Bark scale is applied and frequency bands are grouped into 24 critical bands. Spectral masking effect is achieved using spreading function. The spectrum energy values are transformed into decibel scale [17]. Equal loudness contour is incorporated to calculate the loudness level. The loudness sensation per critical band is computed. STFT is computed for each segment of pre-processed speech. A frame size of 20 ms is deployed with 50% overlap between the frames. The sampling frequency of 1 second duration is 16 kHz. The block diagram of sonogram extraction is shown in Figure 5.

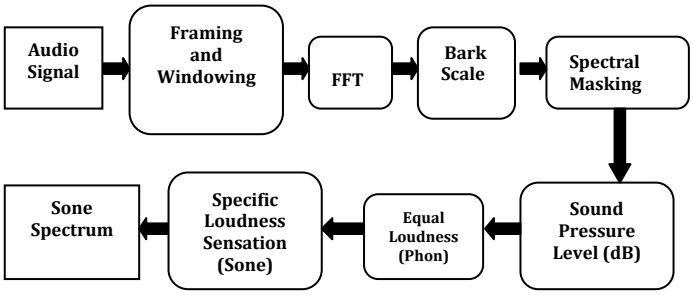

Fig 5: Sonogram Feature Extractions.

A perceptual scale known as bark scale is applied to the spectrogram and it groups the frequencies based upon the perceptive pitch regions to critical bands. The occlusion of one sound to another is modelled by applying a spectral masking spread function to the signal [18]. The spectrum energy values are then transformed into decibel scale. Phone scale computation involves equal loudness curve which represents different perceptron of loudness at different frequencies respectively. The values are then transformed into a sone-scale to reflect the loudness sensation of the human auditory system [19].

4. TECHNIQUES FOR SPEECH INDEXING AND RETRIEVAL

In this Section, Gaussian Mixture Model (GMM) is used to index speech clips. Acoustic features namely PLP, PNCC, SBC and Sonogram are extracted from speech audio clips. GMM is used for creating the index and retrieval is made depending on the maximum probability density function.

4.1 Gaussian Mixture Model

Parametric or non-parametric methods are used to model the distribution of feature vectors. Parametric models are based on the shape of probability density function [20], [21], [22]. In non-parametric modeling only minimal or no assumption regarding the probability density function of feature vector is made [23], [24]. The basis for using GMM is that the distribution of feature vectors extracted from a class can be modeled by a mixture of Gaussian densities.

GMM's represent the feature vectors using Gaussian components and are characterized by the mean vector and the co-variance matrix [25]. Even in the absence of other information,

GMM models have the capability to form an arbitrarily shaped observation density [20]. For a D dimensional feature vector x, the mixture density function for category s is defined as

$$p(\frac{\boldsymbol{x}}{\lambda^s}) = \sum_{i=1}^{M} \alpha_i^s f_i^s(\boldsymbol{x}) \quad (4)$$

The mixture density function is a weighted linear combination of M component uni-modal Gaussian densities $f_i^s(.)$.Every Gaussian density function $f_i^s(.)$ is categorized by the mean vector μ_i^s and covariance matrix $\sum_i^s$ using

$$f_i^s(x) = \frac{1}{\sqrt{(2\pi)^d |\sum_i^s|}} \exp\{-\frac{1}{2}(x-\mu_i^s)^T (\sum_i^s)^{-1}(x-\mu_i^s)\} \quad (5)$$

where, $(\sum_i^s)^{-1}$ and $|\sum_i^s|$ denote the inverse and determinant of the covariance matrix $\sum_i^s$,respectively. The iterative Expectation Maximization (EM) algorithm is used to estimate the parameters of GMM.

EM algorithm is one of the most popular clustering algorithms used to estimate the probabilistic models for each Gaussian component. The Expectation step (E-step) and Maximization step (M-step) are iterated till the convergence of the parameter [4]. EM algorithm finds out maximum likelihood estimation of parameters.

5. PROPOSED METHOD FOR INDEXING AND RETRIEVAL OF SPEECH CLIPS

The algorithm for indexing and retrieval of speech audio clips is described below:

5.1 Algorithm for Indexing of Speech Clips

Step 1: Collect 100 speech clips $s_1, s_2, \ldots, s_{100}$ each of 5 to 10 seconds duration from television broadcast news channels.

Step 2: VAD is performed using RMS energy to mark the envelope of beginning and end of the words. Hence, VAD isolates the individual words.

Step 3: 9-dimensional PLP features are extracted from each word in all 100 speech clips to form the speech index.

Step 4: Repeat steps 1 to 3 for extracting features using 13-dimensional PNCC, 12-dimensional SBC and 22-dimensional sonogram respectively.

5.2 Algorithm for Retrieval of Speech Clips using Index

Step 1: A query speech clip of 1 to 2 seconds duration is extracted from the speech wave file.

Step 2: PLP features are extracted from the speech query.

Step 3: A GMM is fit using the acoustic features extracted.

Step 4: Compute the probability density function for features extracted from each word in the speech index to the query GMM.

Step 5: The speech clip containing words which have the maximum probability density function are retrieved.

Step 6: Repeat steps 2 to 5 for retrieval of the features using PNCC, SBC and sonogram respectively.

5.3 Performance Measures

5.3.1 Accuracy of retrieval

Accuracy of retrieval is a performance measure for speech audio indexing system. It is measured using Equation (6).

$$R = \frac{K}{Q} \times 100 \quad (6)$$

where, R is the accuracy of retrieval, K is the number of speech clips retrieved in the top 'n' ranked list and Q is the total number of queries.

5.3.2 Average number of clips retrieval

The performance of speech indexing and retrieval system is measured by the average number of speech clips retrieved for each query as given in Equation (7).

$$R = \frac{N}{Q} \quad (7)$$

where, *R* represents the average number of speech clips retrieved, *N* is the number of retrieved clips for each speech query and *Q* is the total number of queries.

5.3.3 Retrieval based on rank

A new measure called Rank *n* of a retrieved result is defined as the possibility that the query speech is found in the top *n* clips.

6. EXPERIMENTAL RESULTS

For speech indexing and retrieval, experiments are conducted to study the performance of the speech retrieval algorithms in terms of the average number of speech clips retrieved.

6.1 Database for Speech Indexing

Experiments are conducted for indexing speech audio using Television broadcast speech data collected from Tamil news channels using a tuner card. A total dataset of 100 different speech dialogue clips, ranging from 5 to 10 seconds duration, sampled at 16 kHz and encoded by 16-bit is recorded. Voice activity detection is performed to isolate the words in each speech file using RMS energy envelope. For each speech file, a database of the isolated words is obtained using VAD.

6.2 Acoustic Feature Extraction

For each speech clip ranging from 5 to 10 seconds duration PLP, PNCC, SBC and sonogram features are extracted.
A frame size of 20 ms with a frame shift of 10 ms is used. Thereby 9 PLP, 13 PNCC, 12 SBC and 22 sonogram features are extracted for each word in the speech database.

6.3 Creation of Index

In this experiment, the index is created for each isolated word in the speech database. If a word is of 1 sec duration, a feature vector of size 100×9 is obtained for PLP features. PLP features are extracted for all words in the speech file. Similarly, PLP features are extracted for all the files in the speech database and this forms the index for speech retrieval. The same process is repeated for PNCC, SBC and sonogram features.

6.4 Retrieval of Speech Clips using Index

For retrieval, the keyword of interest is given as a query speech clip of duration 1 to 2 seconds. PLP features are extracted for the query and a GMM is constructed for the same. For every word in the speech database, the probability density function of the feature vectors belonging to the query GMM is computed. Retrieval is based on the maximum probability density function. Similarly experiments are conducted to analyze the performance of speech indexing and retrieval using PNCC, SBC and sonogram features.

Figure 6 shows the performance of speech retrieval for different ranks.

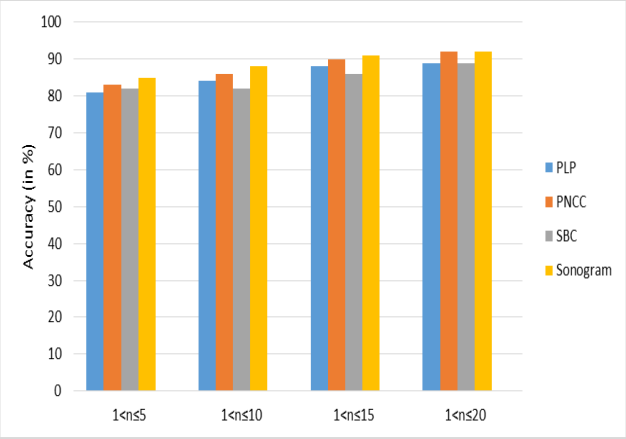

Fig. 6. Performance of Speech Indexing and Retrieval for Different Ranks

Table 1. Average Number of Clips Retrieved for Each Query Using Various Feature Sets.

Features	Average clips retrieved per query
PLP	1.0
PNCC	1.0
SBC	2.0
Sonogram	2.0

Table 1 shows the average number of clips retrieved for each query speech using various feature sets. It can be seen that PLP and PNCC gives an improvement in performance compared to SBC and Sonogram features.

7. CONCLUSION

In this paper, a method is proposed for indexing and retrieval of speech. Voice activity detection is performed to isolate the words in each speech file using RMS energy envelope. For creation of index, PLP, PNCC, SBC and sonogram features are extracted for all words in the speech database. For retrieval, acoustic features are extracted from the speech query and a GMM is constructed using EM algorithm. The probability density function of the feature vectors in the indexed speech database belonging to the query GMM is computed. Retrieval is based on the maximum probability density function. Performance of speech indexing and retrieval system is evaluated for a number of queries and the method achieves an overall accuracy of 91.0%. In feature number of clips can be increased. Different acoustic features can be extracted. Indexing techniques to enhance the retrieval performance may be implemented.

8. REFERENCES

[1] YaliZheng, Chisaki, Y. and Usagawa T., 2013, Speech/Music Indexing for Audio Life Logs from Portable Device Record, IEEE International Conference on Advanced Computer Science and Information Systems, pp. 173-178.

[2] Tsung-Hsien Wen, Hung-Yi Lee, Pei-hao Su and Lin-shan Lee, 2013, Interactive Spoken Content Retrieval by Extended Query Model and Continuous State Space Markov Decision Process, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8510-8514.

[3] Iswarya, P. and Radha, V, 2014, Speech and Text Query Based Tamil - English Cross Language Information Retrieval system, International Conference on Computer Communication and Informatics, pp. 1-4, Coimbatore.

[4] [4] Chien-Lin Huang, Chiori Hori and Hideki Kashioka, 2013, Semantic Inference Based on Neural Probabilistic Language Modeling for Speech Indexing, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8480-8484.

[5] [5] Ivan Markovi, Sre´ ckoJuri´ Kavelj and Ivan Petrovi, 2013, Partial Mutual Information Based Input Variable Selection for Supervised Learning Approaches to Voice Activity Detection, Applied Soft Computing Elsevier, vol. 13, pp. 4383-4391.

[6] Khoubrouy, S. A. and Panahi, I.M.S., 2013, Voice Activation Detection using Teager-Kaiser Energy Measure, International Symposium on Image and Signal Processing and Analysis, pp. 388-392.

[7] Peter M. Grosche, 2012, Signal Processing Methods for Beat Tracking, Music Segmentation and Audio Retrieval, Thesis, Universit¨at des Saarlandes.

[8] PetrMotlcek, 2003, Modeling of Spectra and Temporal Trajectories in Speech Processing, Ph.D thesis, Brno University of Technology.

[9] Chanwookim and Stern, R. M., 2012, Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4101-4104.

[10] Xin Yan and Ying Li, 2012, Anti-noise Power Normalized Cepstral Coefficients for Robust Environmental Sounds Recognition in Real Noisy Conditions, Fourth International Conference on Computational Intelligence and Communication Networks, pp. 263-267.

[11] Arcos Gordillo, C., Grivet, M.A. and Alcaim, A, 2014, PNCC Features and FNN - MAP Compensation Techniques for Continuous Speech Recognition, IEEE International Telecommunications Symposium, pp. 1-5.

[12] Venkatramaphani kumar S and K V Krishna Kishore, 2013, An Efficient Multimodal Person Authentication System using Gabor and Subband Coding, IEEE International Conference Computational Intelligence and Computing Research, pp. 1-5.

[13] Zhu Leqing, Zhang Zhen, 2010, Insect Sound Recognition Based on SBC and HMM, International Conference on Intelligent Computation Technology and Automation, IEEE, pp. 544-548.

[14] Chaya. S, Ramjan Khatik, Siraj Patha and Banda Nawaz, 2014, Subband Coding of Speech Signal Using Scilab, *International* Journal of Electronics & Communication (IIJEC), vol. 2, Issue 5.

[15] Mahdi Hatam and Mohammad Ali Masnadi-Shirazi, 2015, Optimum Nonnegative Integer Bit Allocation for Wavelet Based Signal Compression and Coding, Information Sciences Elsevier, pp. 332-344.

[16] Xiaowen Cheng, Jarod V. Hart, and James S. Walker, 2008, "Time-frequency Analysis of Musical Rhythm," Notices of AMS, vol. 56, no. 3.

[17] Ausgef¨uhrt, 2006, Evaluation of New Audio Features and Their Utilization in Novel Music Retrieval Applications, Master's thesis, Vienna University of Technology.

[18] Eberhard Zwicker and Hugo Fastl, 1999, Psychoacoustics-Facts and Models, Springer Series of Information Sciences, Berlin.

[19] Schroder M. R., B. S. Atal, and J. L. Hall, 1979, Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear, Journal of the Acoustical Society of America, vol. 66, pp. 1647-1652.

[20] Tang, H., Chu, S. M., Hasegawa-Johnson, M. and Huang, T. S., 2012, Partially Supervised Speaker Clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 5, pp. 959-971.

[21] Reddy, P. R., Rout, K. and Rama Murty, K. S., 2014, Query Word Retrieval from Continuous Speech using GMM Posteriorgrams, International Conference on Signal Processing and Communications, pp. 1-6.

[22] Menaka Rajapakse and Lonce Wyse, 2005, Generic Audio Classification using a Hybrid Model Based on GMMs and HMMs, IEEE International Multimedia Modelling Conference, pp. 53-58

[23] Zanoni, M., Ciminieri, D., Sarti, A. and Tubaro, S, 2012, Searching for Dominant High-Level Features for Music

Information Retrieval, European Signal Processing Conference, pp. 2025-2029.

[24] Chunhui Wang, Qianqian Zhu, Zhenyu Shan, Yingjie Xia and Yuncai Liu, 2014, Fusing Heterogeneous Traffic Data by Kalman Filters and Gaussian Mixture Models, IEEE International Conference on Intelligent Transportation Systems, pp. 276-281.

[25] Rafael Iriya and Miguel Arjona Ramírez, 2014, Gaussian Mixture Models with Class-Dependent Features for Speech Emotion Recognition, IEEE Workshop on Statistical Signal Processing, pp. 480-483.