A Study of Speech Emotion and Speaker Identification System using VQ and GMM

Sushma Bahuguna[1], **Y. P. Raiwani**[2]
[1] **BCIIT (Affiliated to GGSIPU)**
New Delhi, India

[2] **CSE, HNB Garhwal University,**
Srinagar, Uttarakhand, India

Abstract

This paper describes a text independent, closed set, speaker identification system to identify the speaker along with the emotional expression (Emo-voice Model) of the particular speech sentence. The system is evaluated on recorded sample sentences of native Hindi speakers in five basic emotions. Spectral Features, Mel Frequency cepstral coefficients have been used to implement emo-voice models using Vector Quantization and Gaussian Mixture modeling techniques for selected sample sentences using MATLAB. The VQ model trained with K-mean algorithm achieves as much as 82.7% of speaker identification with correct emotion accuracy whilst GMM model trained with EM algorithm achieves 87.9% of speaker identification with correct emotion accuracy. The statistical approach of Emo-voice Models could be used to extend the application field of voiceprint recognition technology.

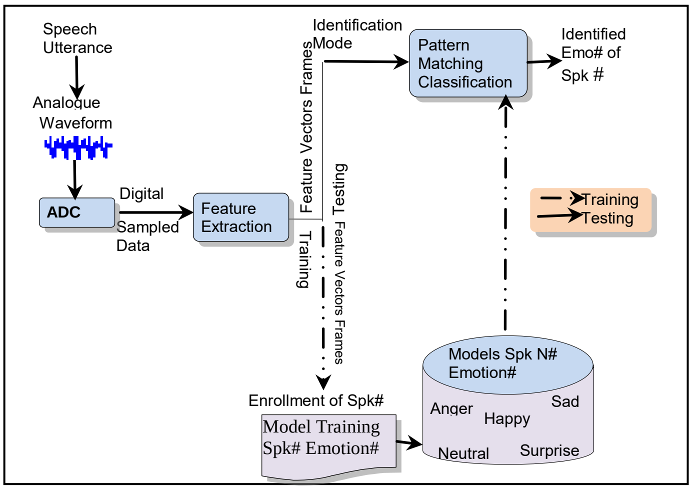
Keywords: *Emo-voice model, EM, GMM, K-mean, VQ*

1. Introduction

The present work explores a real time text-independent, closed-set, speaker identification system comparing a speech signal from an unknown speaker to a database of known speakers to classify speaker and speaker's emotion on the basis of individual information included in speech waves . It is an application based on behavioral and physiological characteristics of the speaker's voice in which unique features of speech are analyzed to identify the speaker speaking in different emotions. Features extracted from the converted digital symbols are stored as character template of the person which is stored in computer database and speaker emotion classification is processed inside the Identification System. The system operates in Training mode and Identification mode. In training mode we make feature model of the voice and using the information of training mode we isolate and identify the speaker in the identification mode. Figure 1 depicts System overview of the Emo-voice model.

The input speech passes through Feature extraction and Feature matching stages in order to classify the speaker along with the expressed emotion. Mel Frequency Cepstral Coefficients (MFCC) have been used for feature extraction. Vector Quantization (VQ) and Gaussian Mixture Model (GMM) techniques are used to explore speaker identification application using MATLAB. In present study 25 sample sentences of 8 native Hindi Speakers (four males and four females) of different age groups were recorded in five basic emotions namely Anger (A), Happiness (H), Neutral (N), Sadness (Sa) and Surprise (S). Chosen sentences were commonly used in everyday communications and recording was done by electrets microphone in partially sound treated room. To judge the emotional content of each speaker for each sentence, a listening test was conducted for the voice samples. 850 sentences were correctly classified for particular emotion category by the listeners was selected for the study.

Fig 1: System overview of Emo-voice Model



2. Development of Speaker Identification Systems

Speaker identification machine using filter banks and correlating two digital spectrograms named voiceprint analysis was invented in the 1960's by [1] at Bell Labs and was improved by [2] using linear discriminators. Formant analysis was introduced by [3] at Texas Instruments. The different scholars used various statistical parameters for speaker features extraction independent of phonetic context including instantaneous spectra covariance matrix, averaged autocorrelation, long-term averaged spectra, spectrum and fundamental frequency histograms and linear prediction coefficients.

The first automated speaker verification providing operational security was built by Texas Instruments system. The Bell Labs invented systems to work over dialed-up telephone lines. [4] found that combination of cepstral coefficient and their first and second polynomial coefficient can be used to increase robustness against distortion by the telephone system. The cepstrum-based features later became standard for speaker recognition as well as speech recognition. The state-of-the-art text independent speaker verification techniques associate one or several parameterization level with world model normalization and one or several score normalizations [5].

One of the key technologies developed in the 1980s is Hidden Markov Model (HMM) approach [6]. It is a doubly stochastic process with an underlying not observable stochastic process which can be observed through another stochastic process that produces a sequence of observations. For text-independent speaker recognition, nonparametric and parametric probability models were recommended. Vector quantization (VQ) was used as a nonparametric model [7]. A set of short-time training feature vectors of a speaker can be efficiently compressed to a small set of representative points, a so-called VQ codebook. Scholars worked on HMM as a parametric model and [8] proposed using an ergodic HMM which allow all possible transitions between states. [9] developed concept by using an eight state ergodic autoregressive HMM represented by continuous probability density functions with two to eight mixture components per state. [10] moduled Gaussian Mixture Model (GMM) using a single-state HMM, as a robust parametric model.

3. MFCC Features

The spectrum of the original signal was computed with the Fourier transform and the obtained spectrum was mapped in Mel, making use of appropriate overlapping windows. The logarithm was calculated for each obtained function.

The discrete cosine transform was calculated and coefficients were the amplitudes of the resulting spectrum. The size of each frame was 25 ms in order to make the contained wave stationary. Each frame had been subjected to the Hamming window to minimize the discontinuities at the edges of the frame. For each frame 20 MFCCs were calculated. The block diagram for computing MFCC is given in Figure 2.

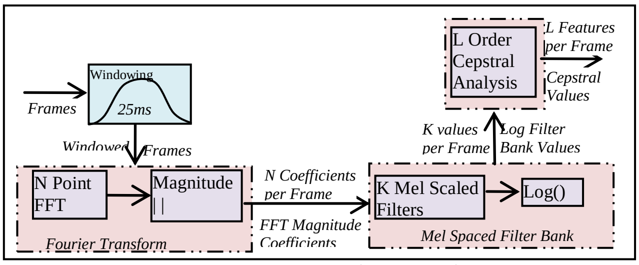


Fig 2: Analysis block diagram for MFCC feature vectors

3.1 Fourier Transform

The basis of performing Fourier transform is to convert the convolution of the glottal pulse and the vocal tract impulse response from time domain into multiplication in the frequency domain [11]. The fast Fourier Transform (FFT) is performed to obtain complex spectral values from each frame. We obtained 256 complex spectral values uniformly spaced from 0 to Fs/2, applying a 512-point FFT, where Fs represent sampling frequency. Phase information is ignored and only the FFT magnitude spectrum is considered in speech processing.

3.2 Mel-spaced filter bank values

Complex spectral values (N = 256) represent too much spectral information and by smoothing of spectrum to only $K = 20$, or so, more efficient representation may be achieved. So, K filter bank values are obtained by conversion of N FFT magnitude coefficients. Tones with different frequencies form speech signal and a subjective pitch can be measured on the ‘Mel’ scale for each tone with an actual Frequency f, measured in Hz. The *Mel-frequency* scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. Mels for a given frequency f in Hz can be computed as:

$$mel(f)= 2595*log10(1+f/700) \qquad (1)$$

A filter bank with one filter for each desired Mel frequency component can be used to stimulate the subjective spectrum. The filter bank has a triangular band pass frequency response and constant Mel-frequency interval determines the spacing and bandwidth by a constant Mel-frequency interval. Filter bank values are obtained by Cross-wise multiplication of the N FFT magnitude coefficients by the K triangular filter bank weighting function and then accumulating the results from each triangle. The centers of the triangle filter banks are spaced according to the Mel scale. Mel filter banks are non-uniformly spaced on the frequency axis, so we have more filters in the low frequency regions and less number of filters in high frequency regions [12]. We applied Mel filter banks on spectrum getting after Fast Fourier transform operation for windowed signal and signal processed based on human perception.

(2)

Where :

S(l) :Mel spectrum,
S(K) :Original spectrum.
M(K) :Mel filter bank.

l=0, 1... l - 1, where l is the total number of Mel filter banks
n / 2 = Half FFT size.

Now, we proceed to the next stage to get the cpestrum or the mel frequency cpestrum coefficients.

3.3 Cepstral analysis

The final step is to convert the log Mel spectrum back to time resulting Mel frequency cepstrum coefficients (MFCCs). For the given frame analysis the cepstral representation provides good illustration of the local spectral properties. Mel spectrum coefficients are converted to the time domain using the Discrete

Cosine Transform (DCT). The speech signal represented as a convolution between slowly varying vocal tract impulse response (filter) and quickly varying glottal pulse (source) producing speech spectrum having spectral details (high frequency) and spectral envelop (low frequency). These spectral envelope and spectral details has to be separated from the spectrum. We converted the multiplication of the magnitude of the Fourier transform into addition and then by taking the inverse DCT of the logarithm of the magnitude spectrum, the glottal pulse and the impulse response were separated. We used DCT because the signal is real (we took Magnitude) with mirror symmetry. The DCT does not needs complex arithmetic and implements the same function as the FFT more efficiently by taking advantage of the redundancy in a real signal. Moreover it is more efficient computationally. The following equation was used to calculate MFCCs [13].

$$- \quad — \qquad ...(3)$$

Where n=1, 2….K

The number of Mel cepstrum coefficients is chosen 20. The first component, c~0, represents the mean value of the input signal which carries little speaker specific information that is excluded from the DCT. Since the log power spectrum is real and symmetric, inverse FFT reduces to a Discrete Cosine Transform (DCT). By this process a set of Mel-frequency cepstrum coefficients is computed for each speech frame of 25 ms with overlap which is called acoustic vector. These acoustic vectors are used to represent and recognize the voice characteristic of the speaker and each input utterance is transformed into a sequence of acoustic vectors.

4. Modeling with VQ

In Vector Quantization technique using the extracted features, a codebook for each speaker was generated and clustering the feature vectors was done using the K-means algorithm. Centroids were randomly chosen among feature vectors by algorithm and then nearest centroid, and the new centroids were calculated for the new clusters. The process was followed continuously till there was no more change in the cluster-center assignment (mean square error) between the feature vectors and the cluster-centroids below a certain threshold[14]. So, objective of the K-means is to minimize total intra-cluster variance, V:

$$(4)$$

Where, there are k clusters S_i, i = 1, 2... k and μ_i is the centroid or mean point of all the points. The K-means algorithm determines the centroid, distance of each object to centroids and finds closest centroid.

In identification phase, a sequence of feature vectors representing an unknown speaker was compared with the codebooks in the database. Distortion measure was computed for each codebook and the speaker with the lowest distortion was chosen.

$$(5)$$

The average of the Euclidean distances defines the sum of squared distances between vector and its centroid (distortion measures):

$$- \qquad (6)$$

Where C_{min} denotes the nearest codeword x_t in the codebook C and d (.) is the Euclidean distance. Thus, each feature vector in the sequence X is compared with all the codebooks and the codebook with the minimized average distance is chosen to be the best. The formula used to calculate the Euclidean distance [14] can be defined as following:

The Euclidean distance between two points and

$$\qquad …(7)$$

The speaker with the lowest distortion distance is chosen to be identified as the unknown person.

4.1 System Implementation

Figure 3 depicts the VQ training and testing framework. In enrollment mode speaker with known identity is enrolled into the system's database. The feature vectors consisting of 20 MFCC from each frame were subjected to vector quantization and 16 code books were generated. The resulting set of code vectors was stored in the speaker database. In identification mode unknown speaker gave a voice input and the system made a decision about the speaker's identity.

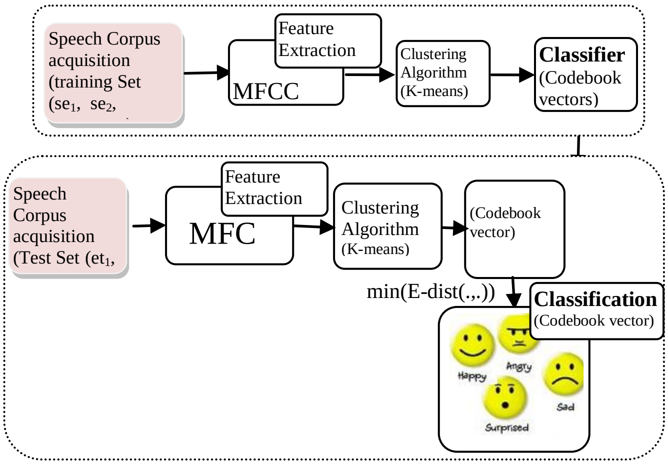

Speech Corpus acquisition (training Set (se1, se2,
Feature Extraction
MFCC
Clustering Algorithm (K-means)
Classifier (Codebook vectors)
Speech Corpus acquisition (Test Set (et1,
Feature Extraction
MFC
Clustering Algorithm (K-means)
(Codebook vector)
min(E-dist(.,.))
Classification (Codebook vector)
Happy
Angry
Sad
Surprised

Fig 3: VQ training and testing framework

The Euclidean distance between feature vector and code book for each speaker was calculated and the speaker with the smallest average minimum distance was picked.

5. Modeling with GMM

A Gaussian Mixture technique [10] is a parametric probability density function represented as a weighted sum of Gaussian component densities. It is a hybrid technique between Gaussian uni-modal and a vector quantization as it uses a discrete set of Gaussian functions, each with their mean and covariance matrix, to allow a better modeling capability. A Gaussian mixture density $p(x|\lambda)$ is defined as a sum of M component densities as governed by the output probability expression.

(8)

Where x = observation vector of size D × 1, is the prior probability or mixing weight of i^{th} Gaussian component and

The i^{th} multivariate Gaussian distribution with mean and covariance $\sum_i$ is defined as:

——— – (9)

Where μ_i=a vector of size D × *1* and = matrix of size *D* × *D*, $|\sum_i|$ = determinant of $\sum_i$.

Given training vectors and a GMM configuration, the values for the GMM model parameters λ
are calculated based on training data. The likelihood of ***X*** given the model λ is for a training vector sequence ***X*** = $\{x_1, x_2, \ldots,$ $\}$ is

.....(10)

The likelihood function in (*10*) is a nonlinear function of $\lambda\hat{o}$ and we need an iterative *expectation maximization* (EM) algorithm[15] to find the values for the parameters in the model λ that maximize the likelihood function. To guarantee monotonic convergence to the set of optimal parameters is the key feature of the algorithms.. Figure 4 depicts GMM ML scoring.

The basic design of the EM algorithm begins with an initial model λ, to estimate a new model $\bar{\lambda}$ in a way that $\bar{}$ $|\lambda)$. The new model becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached. The initial model is derived by using binary VQ estimation. On each EM iteration, following formula's re-estimation guarantee a monotonic increase in the model's likelihood value.

Mixture Weights

– – (11)

Means

– ——— ....(12)

Variances (diagonal covariance)

....(13)

——— - –

Where μ refer to arbitrary elements of the vectors respectively.

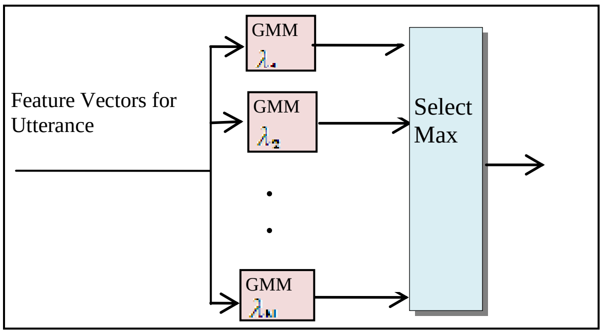


Fig. 4: GMM ML Scoring

In testing phase, speaker-specific models learned in the training phase are used to identify the speaker of the test vector sequences. The identification is made based on the maximum *a posteriori* probability: the speaker of a test vector sequence is decided to be the one whose GMM model gives the highest posterior probability [16]. The maximum a posteriori estimate for the speaker of a test vector sequence $X = \{x_1, \ldots, x_T\}$ is:

——— .....(14)

The term is a prior probability of the j^{th} speaker, and can be assumed to be equal (1/S) for all speakers. The denominator $X)$ is also equal for all the speakers, hence we can write:

....(15)

By assuming the observations to be independent this can be written as a sum of log-densities:

.....(16)

The probability density $p(X_t|\lambda_j)$ is the same as in (*8*). The segment is identified to be spoken by the speaker whose model λ_s maximizes (*16*).

5.1 System Implementation

GMM using 8 to 64 mixtures were used to analyze organized matrix of Mel-Cepstral coefficients for each frame. A set of statistical data i.e. mean vector, a covariance matrix and a weight vector constituted template. The training step stored each template generated from speaker's utterances analysis into the system which represented the knowledge base for the training phase. In testing phase, whole knowledge base of the system was compared to the test template of each speaker using log-likelihood. The output comprises the similarity estimation of each test with respect to each profile stored into the system. Figure 5 depicts the GMM training and testing framework.

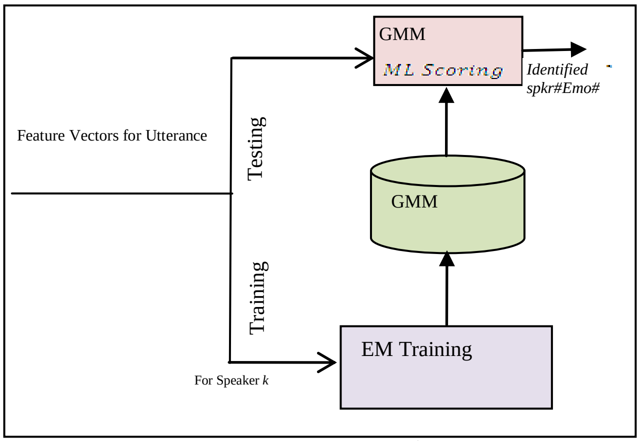


Fig 5: GMM training and testing framework

6. EXPERIMEMTAL RESULTS

In order to evaluate performance of Text Independent, closed set Emo-voice models based on vector quantization and Gaussian Mixture model techniques, as a real time application, database of recorded sample sentences was subjected to the training and testing sessions. Two sentences per speaker per emotion were used for training, comprising identification sentences. In both experiments a pre-emphasis filter with Co-efficient 0.97 was applied from each 25ms frame at every 10ms and generated 20 MFCC.

6.1 VQ: Speaker with Emotion Classification accuracy

The VQ Emo-Voice model is composed of attributes MFCCs and Vector Quantization technique with Performance evaluation as a function of 20 MFCC and a codebook size of 16. Table 1 illustrate output of 850 voice sample sentences by VQ Emo-Voice model specifying classified class of each speaker along with emotion (*Spk*$_1$-*Spk*$_8$)

Table I: Speaker with emotion identification(vq) %

Emo	Spk$_1$	Spk$_2$	Spk$_3$	Spk$_4$	Spk$_5$	Spk$_6$	Spk$_7$	Spk$_8$
A	89	94	85	88	93	44	73	88
H	90	94	90	95	90	85	75	92
N	84	50	85	94	80	90	90	100
Sa	37	94	89	95	94	100	67	69
S	47	85	85	50	85	94	100	80

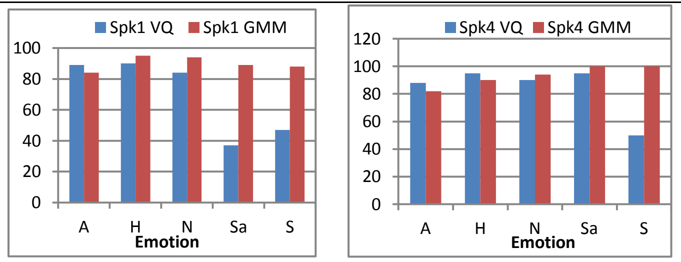
6.2 GMM: Speaker with Emotion Classification accuracy

The GMM Emo-Voice model is Composed of attributes MFCCs and Gaussian mixture technique. The number of mixtures experimented from 8 to 64. With current set of 20 MFCC the best performance for GMM was with 16 mixtures and performance deterioration was observed when more mixture component was added. Table 2 illustrate output of sample sentences by GMM Emo-voice model specifying correct classified class of each speaker along with emotion (*Spk*$_1$-*Spk*$_8$) .

TABLE II : SPEAKER WITH EMOTION IDENTIFICATION(GMM) %

Emo	Spk$_1$	Spk$_2$	Spk$_3$	Spk$_4$	Spk$_5$	Spk$_6$	Spk$_7$	Spk$_8$
A	84	83	95	82	79	89	67	81
H	95	87	90	90	90	95	95	92
N	94	89	90	94	70	95	95	90
Sa	89	82	89	100	100	93	78	92
S	88	90	75	100	90	75	69	95

The experiment has performed classification of speaker with emotions using Vector Quantization and Gaussian Mixture Model techniques. The results show the superior performance of GMM with 87.9% of identification accuracy in comparison with VQ with 82.7% of identification accuracy. Fig.6 (a) and (b) shows a comparison of VQ and GMM model depicting GMM having comparatively better performance than VQ. VQ have higher accuracy in some cases, hybrid model of VQ and GMM may be used to improve identification accuracy.



(a) Speaker No.1 (b) Speaker No. 4

Fig.6 comparison of results of VQ and GMM

7. CONCLUSION

The proposed approaches yielded an overall speaker with correct emotion classification accuracy of 87.9% and 82.7% for Gaussian technique and vector quantization technique respectively. The experimental results of Emo-voice models are limited to recognize the speaker based on the devices used for recording the corresponding speech files. Fusion of different prosodic and spectral features may also be considered for achieving more accurate recognition. Moreover, hybrid modeling approach by using Vector Quantization and Gaussian Mixture techniques may lead to more accurate classification of any later study. In sum, from the attain accuracy it is believe that proposed application is feasible for speaker along with emotion identification in real time.

References

[1] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition", Journal of Acoustical Society of America, vol. 35, pp. 354-358, 1963.

[2] S. Pruzansky and M. V. Mathews, "Talker recognition procedure based on analysis of variance" , Journal of Acoustical Society of America, vol. 36, pp. 2041-2047, 1964.

[3] G. R. Doddington,"A method of speaker verification", Journal of Acoustical Society of America, vol. 49, 139 (A), 1971.

[4] S. Furui, “Cepstral analysis technique for automatic speaker verification”, IEEE Trans. Acoustics, Speech Signal Processing, ASSP-29, pp. 254- 272, 1981.

[5] S. Furui,” Fifty years of progress in speech abd speaker recognition”, ECTI transaction on computer and interaction technology,vol. 1(2), 2005.

[6] J. Ferguson et al., “Hidden Markov Analysis: An Introduction”, in J.D. Ferguson (ed.) Hidden Markov models for speech, Princeton, NJ: IDA-CRD, 1980

[7] A. E. Rosenberg and F. K. Soong, “Evaluation of a vector quantization talker recognition system in text independent and text dependent models”, Computer Speech and Language vol. 22, pp. 143-157, 1987.

[8] B. Poritz, “Linear predictive hidden Markov models and the speech signal”, Proc. ICASSP 82, pp. 1291-1294, 1982.

[9] N. Tishby, “On the application of mixture AR Hidden Markov Models to text independent speaker recognition”, IEEE Trans. Acoust., Speech, Signal Processing, ASSP- vol. 30(3) , pp. 563-570, 1991.

[10] R. C. Rose and D. A. Reynolds, “Text- independent speaker identification using automatic acoustic segmentation”, In proceedings of the international conference on Acoustics, speech and Signal processing, April 1990, pp 293-296, 1990.

[11] J.Walker and P. Murphy, “A review of glottal waveform analysis”: Y. Stylianou, M. Faundez-Zanuy, A. Esposito (Eds.): WNSP 2005, LNCS 4391, pp. 1–21, 2007.

[12] www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf by K. Prahallad , Speech Technology: A Practical Introduction.

[13] D.A. Reynolds and R.C. Rose, Robust Text Independent Speaker Identification using Gaussain Mixture Speaker models”, IEEE Transation on speech and audio processing, Vol 3, No. 1 Jan 1995, pp. 72-83

[14] “K means Analysis”: http://www.clustan.com/k-means_analysis.html.

[15] The EM Algorithm for Gaussian Mixtures, probabilistic Learning: Theory and Algorithms, CS274A: http://www.ccs.neu.edu/ home/jaa/CS6140.13F/Homeworks/ HW05/8-em.pdf

[16] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models”, Speech Communication, vol. 17(1-2), pp. 91–108, 1995.