EFFECT OF NUMBER OF MIXTURE COMPONENTS OF GMM AND FEATURE VECTOR DIMENSIONS IN NON-INTRUSIVE SPEECH QUALITY EVALUATION

[1]RAJESH KUMAR DUBEY, [2]ARUN KUMAR

[1]Jaypee Institute of Information Technology, Noida India,
[2]Indian Institute of Technology Delhi, India
E-mail: [1]rajeshk_dubey@yahoo.com, [2]arunkm@care.iitd.ac.in

Abstract— A meaningful objective model for the estimation of non-intrusive speech quality can be established by utilizing the speech production model and the auditory perception phenomena of the human auditory system. To supplement the subjective mean opinion score (MOS), the estimation of objective MOS is configured using the principle of human auditory perception models and the speech production model. In this work, the Lyon's auditory features, mel-frequency cepstral coefficients (MFCC) and features corresponding to the vocal tract resonances such as line spectral frequencies (LSF) are concatenated to make the feature vector. The size of feature vectors are reduced using principal component analysis (PCA) and reduced size feature vectors are used to compute the objective MOS using GMM probabilistic approach. The effect of number of mixture components in GMM and the dimensions in different reduced size feature vectors made up of the combinations of meaningful speech features such as Lyon's auditory features, MFCC and LSF has been studied that which one leads to better objective MOS in terms of increased correlation with the subjective MOS. These feature vectors also include the first and second differences of MFCC and LSF features. The training of Gaussian Mixture Model (GMM) to obtain its parameters using these reduced size feature vectors has been done using expectation maximization (EM) algorithm for different speech databases. The performance evaluation in terms of correlation between the subjective MOS and the objective MOS using different reduced size feature vectors for different GMM mixture components are compared. The results are also compared with ITU-T Recommendation P.563, the standard for non-intrusive speech quality estimation.

Index Terms— Speech quality, Gaussian mixture model, Auditory features, mel-frequency cepstral coefficients, Line spectral frequencies.

I. INTRODUCTION

In modern telecommunication networks and any other systems using speech processing algorithms, the quality evaluation of speech is essential to monitor and maintain the quality of service from design and development point of view. The ideal method for speech quality evaluation is subjective listening test according to absolute category rating (ACR) method called subjective MOS [1], but it is time taking, expensive and impractical for system automation point of view. Thus, to supplement the subjective listening tests, speech quality evaluations are done objectively. There are several objective speech quality evaluation techniques in the literature. The quality of speech measured objectively is called objective MOS and the performance of objective speech quality evaluation algorithms are given in terms of correlation between the subjective MOS and the objective MOS. The objective quality evaluation of speech may be intrusive (double ended) or non-intrusive (single ended). The intrusive method uses both the original clean speech and degraded received speech, while non-intrusive method uses only degraded received speech for quality evaluation. The ITU-T Recommendation P.563 is the standard for non-intrusive speech quality evaluation [2]. In [3], low complexity non-intrusive speech quality assessment algorithm is presented, where objective speech quality evaluation has been done by GMM mapping using local and global features of speech obtained from speech coders without considering any degradation model. The training of GMM has been done by EM algorithm [4] using these local and global speech features. The human auditory system and perception phenomena of speech signal in form of temporal envelope representation of speech have been used in auditory non-intrusive quality estimation ANIQUE model [5]. A detailed description of computational model by Lyon for human auditory system as filtering, detection and compression in cochlea is given in [6]. A meaningful non-intrusive speech quality evaluation algorithm is developed using several combination of features obtained from human speech perception phenomena of auditory model and speech production model is given in [7].

The comparison of the performance of two different types of speech features such as mel-frequency cepstral coefficients (MFCC) and reconstructed phase spaces (RPS) generally used in speech recognition problem is presented in [8]. The GMM mapping by use of the combinations of MFCC, perceptual linear prediction (PLP) coefficients and line spectral frequencies (LSF) features is given in [9] for non-intrusive speech quality evaluation. To include the effect of temporal masking and localized distortion of speech utterances in listening and perception phenomena, multi-resolution auditory model (MRAM) features have been used for GMM mapping in [10] for non-intrusive speech quality evaluation. In this work, the feature vectors are obtained by making the combinations of speech features such as Lyon's

auditory features, MFCC and LSF features. The size of feature vectors is reduced using principal component analysis (PCA) and GMM mapping has been done to compute the objective MOS value using reduced size feature vector and GMM parameters. The effects of the number of mixture components and the dimensions of the reduced size feature vectors are investigated in terms of correlation between the subjective MOS and the objective MOS. Results are also compared with the ITU-T standard.

II. FEATURES OF SPEECH SIGNALS

The different features of speech signals used in this work are Lyon's auditory features, MFCC, LSF features, and magnitude of 1st difference of MFCC and LSF features. The detail description of these features is give below:

i. Lyon's auditory features

The important psychoacoustic phenomena of the human peripheral auditory system are absolute hearing threshold, sensation of loudness, and two types of auditory masking, viz. simultaneous masking and temporal masking. To get the complete computational auditory model for speech perception, these psychoacoustic models of individual auditory phenomena must be individually modeled. The physiological structure of the human ear closely links with these auditory phenomena. The schematic model of human auditory mechanism adapted from [11] is shown in Figure 1.

The human auditory system is divided mainly into three main parts: outer, middle and inner ear. The outer ear performs directional filtering, the middle ear is for impedance transformation from air to fluid and the inner ear is for transformation of fluid borne sound to neural excitation. The cochlea is a snail shell like structure filled up with watery fluid in a cross-section of oval and round window and is a frequency-place separator, which separates the speech sound according to frequency from low to high. The dynamic compression of the intensity of the signal is performed by the basilar membrane (BM) of the cochlea according to frequencies. For the sensation of the signal, a neural spike at the base of auditory nerve is only generated when the stereocilia of the inner hair cell are bent one way and no spikes are generated other way in the central processing block. The vibrations of fluid of the cochlea give motion to the basilar membrane and organ of corti to the thousands of hair cells, which convert the sound vibration motion to the electrical signal in auditory nerve and communicated via neurotransmitters to thousands of nerve cells according to frequency from low to high, which reaches the brainstem for further processing. The cognitive mapping and pattern recognition is done in the brain for perception of speech sound.

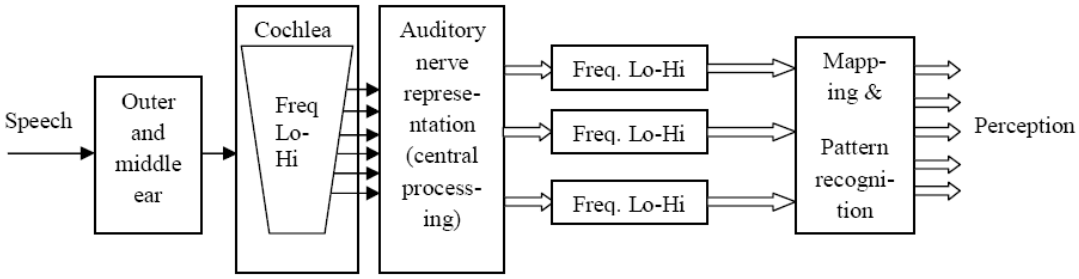

Figure 1: Human auditory mechanism [11].

In this work, Lyon's auditory model is used to represent human auditory phenomena [6], which is a well known computational model using filtering in the 1st stage by filter banks of broadly tuned cascade of low-pass filters, detection as half wave rectification at 2nd stage and compression as automatic gain control in the cochlea at 3rd stage as shown in Figure 2. The outputs of these stages approximately represent the neural firing rates.

Input Speech

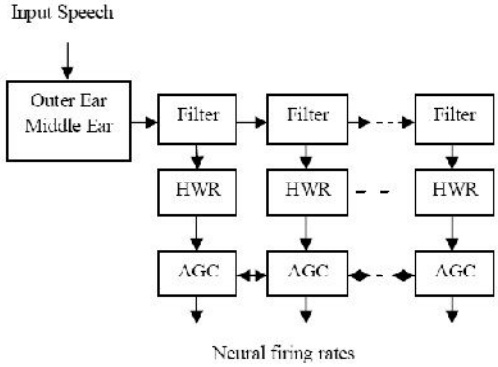

Figure 2: Lyon's auditory model

The short-time autocorrelation (STA) of each of the output of the auditory model is computed by:

$$r_{xi}(m,\tau) = \frac{1}{N}\sum_{n=0}^{N-|\tau|-1} x_{wi}(n,m).x_{wi}(n-|\tau|,m) \quad (1)$$

where, $x_{wi}(n,m) = x_i(n).w(m-n)$ (2)

and r_{xi} is the short time autocorrelation function, x_{wi} is the windowed speech signal by Hamming window, $x_i(n)$ is the input speech signal, i is the channel index, m is the discrete time index, and τ is the autocorrelation lag.

Each of the outputs of Lyon's auditory model is short time windowed and auto-correlated. Thus, number of STAs obtained is equal to the number of channels. The high energy active speech frame is windowed by Hamming window and is passed through the 64-channel Lyon's model producing a 64-dimensional Lyon's feature vector. It is implemented with the open source auditory toolbox [12]. For each channel output, the mean, variance, skewness and kurtosis are calculated over the frame to capture the statistics of computed Lyon's auditory features and appended in a single vector producing a 256-dimensional feature vector. Then, principal component analysis (PCA) is done to optimally reduce the vector dimensionality to 14 [13]. The 14 principal components were observed to retain more than 98% of the total energy of active speech when tested on the ITU-T P. Supplement-23 database [14]. The distribution of energy over number of principal components is shown in Figure 3.

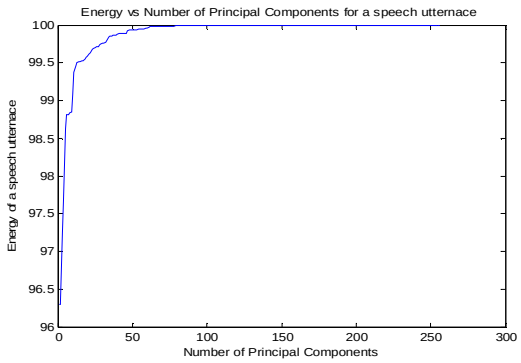

Figure 3: Distribution of energy vs. number of principal components for an active speech.

ii. Mel-frequency cepstral coefficients

The MFCC are one of the most widely used feature representation of the speech signal frame that captures the vibrations of BM of the human ear's critical bandwidth with frequency. The MFCC technique uses two types of filters: linear frequency spaced filters below 1000 Hz and logarithmic frequency spaced filters above 1000 Hz frequency [15]. The MFCCs are less susceptible to the variations in the speech waveform due to the varying physical conditions of speaker's vocal cord and are more sensitive to the different external degradations of speech sound and are an effective perceptual representation of the quality variations of speech signal. Also, MFCC has a de-correlating effect due to the use of Discrete Cosine Transform (DCT) on the log mel-spectrum coefficient calculated over the frames and thus suitable for learning algorithm.

iii. Line spectral frequencies features

The LSF features also offer an alternative efficient spectral envelope representation form for speech as borne out by its extensive use in speech coding algorithms. They carry intrinsic information of the formant structure of phoneme which is related to the resonance frequencies of the vocal tract of the speaker during articulation [16]. A 10th order LPC analysis is done over the frames of active speech of duration 16 ms windowed using a Hamming window of same length to obtain a 10values of LSF features. To obtain a 10-dimensional LSF feature vector for entire speech utterance, the mean values over the frames are computed.

III. FEATURE VECTORS AND GMM PROBABILISTIC APPROACH

The combination of 14-dimensional Lyon's feature vectors, 13-dimensional MFCC, 10-dimensional LSF features, and magnitude of 1st & 2nd differences of MFCC and LSF are used for training of GMM and computation of objective MOS for degraded speech utterances.

The subjective MOS score θ_j from MOS labeled speech databases is appended to the feature vector Ψ, and used for the training of a joint GMM using EM algorithm [4] to obtain parameters of the joint GMM $\Pi(\omega^{(k)},\mu^{(k)},\sum^{(k)})$ with $k=1,2,3\ldots$, M mixture components, where $\omega^{(k)}$, $\mu^{(k)}$, and $\sum^{(k)}$are the mixture weight, mean, and covariance matrix respectively of the k-th mixture component. These parameters are further used to estimate the objective MOS of the test speech utterance along with the feature vectors. The objective MOS estimate $\hat{\theta}$ is obtained using MMSE criterion:

$$\hat{\theta} = \hat{\theta}(\psi) = \arg\min_{\hat{\theta}(\psi)} E\left\{(\theta - \hat{\theta}(\psi))^2\right\} = E\{\theta/\psi\} \quad (3)$$

The modeling of the joint density function of the feature vector variables along with the subjective MOS scores as a GMM facilitates the estimation:

$$f(\psi/\Pi) = \sum_{k=1}^{K} \omega^{(k)} N\left(\psi/\mu_{\psi}^{(k)}, \Sigma_{\psi\psi}^{(k)}\right) \quad (4)$$

where, $N(\psi/\mu^{(k)},\Sigma^{(k)})$ are the multivariate Gaussian densities, with $\mu^{(k)}$being the mean vectors and $\Sigma^{(k)}$ the covariance matrices of the k^{th} mixture components of Gaussian density.

IV. RESULTS AND DISCUSSION

The silence region is removed from the narrowband input speech signal and the active speech regions are extracted for further processing using voice activity detection (VAD) algorithm.

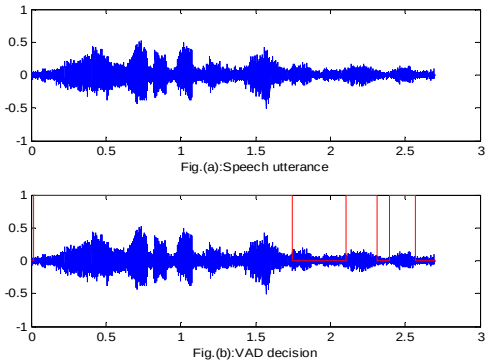

Figure 4: Noisy speech and VAD decision.

The waveform of a noisy speech utterance and VAD decision shown by label '1' for active speech and label '0' for silence region is given in Figure 4. The features (Lyon's feature, MFCC and LSF) are computed for each frame of active speech segmented into a frame length of 16 ms with 50% frame overlap. To observe the effect of feature vector dimension *d*, adequacy of training database and the potential of over-fitting with GMM mapping, the following feature vectors are used in the experiment:

(i) Feature vector1: 14-dimensional Lyon's features, 10-dimensional LSF features, 13-dimensional MFCC, magnitude of 1st & 2nd differences of LSF features, and magnitude of 1st & 2nd differences of MFCC.

(ii) Feature vexctor2: 14-dimensional Lyon's features, 10-dimensional LSF features, 13-dimensional MFCC, magnitude of 1st differences of LSF features, and magnitude of 1st differences of MFCC.

The number of training parameters will be equal to *M* x *[1+d+d. (d+1)/2]*, for *M* Gaussian mixture components and a feature vector of dimension *d [3]*. By increasing the number Gaussian mixture components there is a substantial increase in the number of training parameters. The number of training parameters will also increase with an increase in the dimension of the feature vectors. When number of training parameters increases, the requirement of the size of training database also increases. In this work, the ITU-T P. Supplement-23 databases consisting of 1328 speech files of duration 3 seconds each and all sampled at 8 kHz with different types of degradation has been used [14]. The database is having two experiments based on ACR subjective rating, Experiment-1 and Experiment-3. In Experiment-1, there are three sub-experiments A, D, and O for three different types of languages each having 176 speech files and in Experiment-3, there are four sub-experiments namely A, C, D, and O for four different types of languages each having 200 speech files. The results in terms of correlation coefficients between the subjective MOS and the objective MOS for condition averaged case are given in Table 1, where MOS values are averaged value for same condition of degradation of speech signal**.** It shows that for *d*=60 and *d*=83 there is no significant variation in the weighted average of the correlation coefficient with *M*=12 Gaussian mixture components, the respective values being 0.915 and 0.914.

Table 1: The correlation coefficients for condition averaged MOS using different dimensions of feature vectors with different number of GMM mixture components for ITU-T P. Supplement-23 database.

Data of Different Expts.	ITU-T Rec. P.563	Feature vector1 With *M=8 & d=83*	Feature vector2 with *M=12 & d=60*	Feature vector1 with *M=12 & d=83*	Feature vector1 with *M=16 & d=83*
1(A) French	0.759	0.910	0.938	0.937	0.924
1(D)-Japanese	0.701	0.949	0.934	0.933	0.916
1(O)-Am. English	0.790	0.952	0.951	0.941	0.932
3(A)-French	0.768	0.883	0.886	0.894	0.905
3(C)-Italian	0.762	0.843	0.894	0.894	0.890
3(D)-Japanese	0.801	0.912	0.922	0.923	0.908
3(O)-Am. English	0.788	0.870	0.890	0.881	0.863
Weighted Average	0.768	**0.901**	**0.915**	**0.914**	**0.904**

Table 2: Comparison of correlation for condition averaged MOS using different size of training data with 60-dimensional feature vector1 and 12-GMM mixture components for same database.

Data of Different Expts.	Training data size = 800	Training data size = 1000	Training data size = 1100	Training data size = 1150	Training data size = 1195
1(A)-French	0.926	0.933	0.921	0.933	0.938
1(D)-Japanese	0.940	0.927	0.932	0.935	0.934
1(O)-Am English	0.946	0.947	0.944	0.945	0.951
3(A)-French	0.878	0.835	0.866	0.890	0.886
3(C)-Italian	0.860	0.869	0.866	0.872	0.894
3(D)-Japanese	0.704	0.906	0.915	0.920	0.922
3(O)-Am English	0.887	0.857	0.904	0.909	0.890
Weighted Average	**0.874**	**0.894**	**0.905**	**0.914**	**0.915**

Thus, for the best performing feature vector of dimension 60 i.e. Lyon's features + MFCCs + LSFs + ΔMFCC + ΔLS, the number of training parameters will be 22692 with 12 Gaussians mixture components. Each parameter of GMM such as mean and covariance matrices are computed as averages over 90% of 1328 speech utterances in ITU-T P. Supplement-23 database in ten-fold cross-validation process.

To check the adequacy of training database size the results are computed in terms of correlation coefficients by increasing the size of training database and given in **Table 2**. It is observed that when the training data size increases, there is an improvement in the average value of the correlation coefficients, but for a training data size of 1150, the

GMM parameters saturates and remain approximately constant. There is no further change in the weighted average value of the correlation coefficients. For the training data size 1100, 1150, and 1195, the weighted average values of the correlation coefficients are 0.905, 0.914, and 0.915 respectively. Thus, it can be inferred that the size of database is sufficient to train the GMM parameters used for the estimation of objective MOS.

CONCLUSIONS

The methods for computing different speech features from active speech such as Lyon's auditory features, MFCC, LSF features and combining the relevant features along with their subjective MOS score for GMM training has been investigated in this work to observe the effect of the number of dimensions of feature vectors and the number of mixture components of GMM in non-intrusive speech quality evaluation problem. The non-intrusive speech quality estimation has been done as a conditional probability using GMM parameters and different speech feature vectors. It is concluded that 12 mixture components of GMM are sufficient for a speech database of 1328 speech utterances.

REFERENCES

[1] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," 1996.

[2] ITU-T Rec. P.563, "Single ended method for objective speech quality assessment in narrow-band telephony applications," 2004.

[3] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low complexity non-intrusive speech quality assessment," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1948-1956, 2006.

[4] A. P. Dempster, N. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1-38, 1977.

[5] D. S. Kim, "ANIQUE: An auditory model for single ended speech quality estimation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 13, no. 5, pp. 821-831, 2005.

[6] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Palo Alto, CA, 1982, pp. 1282-1285.

[7] R. K. Dubey and A. Kumar, "Non-intrusive speech quality assessment using several combinations of auditory features," *Int. Journal of Speech Technology*, Springer, vol. 16, no. 1, pp. 89-101, 2013.

[8] N. Parmar and R. K. Dubey, "Comparison of performance of the features of speech signal for non-intrusive speech quality assessment", in *Proc. of Int. Conf. on Signal Processing and Communication*, Noida, India, 2015, pp. 243-248.

[9] R. K. Dubey and A. Kumar, "Non-intrusive objective speech quality assessment using a combination of MFCC, PLP and LSF features", in *Proc. of Int. Conf. on Signal Processing and Communication*, Noida, India, 2013, pp. 297-302.

[10] R. K. Dubey and A. Kumar, "Non-intrusive speech quality estimation using multi-resolution auditory model features", *IET Signal Processing*, vol. 9, no. 9, pp. 638-346, 2015.

[11] M. B. Sachs, C. C. Blackburn, and E. D. Young, "Rate-place and temporal place representations of vowels in the auditory nerve and anteroventral cochlear nucleus," *Journal of Phonetics*, vol. 16, pp. 37–53, 1988.

[12] 12.http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[13] K. Audhkhasi and A. Kumar, "Two scale auditory features based non-intrusive speech quality evaluation," *IETE Journal of Research*, vol. 56, no. 2, pp. 111-118, 2010.

[14] ITU-T Rec. P.Supplement-23 "ITU-T coded-speech database", 1998.

[15] W. Han, C. F. Chan, C. S. Choy, and K. P. Pun, "An efficient MFCC extraction method in speech recognition," in *Proc. IEEE Int. Symposium on Circuits and Systems*, Island of Kos, 2006, pp. 145-148.

[16] E. Bozkurt, E. Erzin, C. E. Erdem, and A. T. Erdem, "Use of line spectral frequencies for emotion recognition from speech," in *IEEE Int. Conf. on Pattern Recognition*, 2010, pp. 3708-3711.

★★★