Speech Enhancement using Hmm and Snmf(Os)

BarinderPal Singh
M .Tech Student
Information Technology
Chandigarh Engineering College
Landran, Mohali

Dr. Shashi Bhushan
Professor & Head of Department
Information Technology
Chandigarh Engineering College
Landran, Mohali

Abstract—The speech enhancement is the process to enhance the speech signal by reducing the noise from the signal as well as improving the quality of the signal. The speech signal enhancement requires various techniques associated with the signal noise removal as well as the signal patch fixation in order to enhance the frequency of the speech signal. In this paper, we have proposed the new speech enhancement model for the speech enhancement with the amalgamation of the various speech processing techniques together. The proposed model is equipped with the Supervised sparse non-negative matrix factorization (S-SNMF) along with hidden markov model (HMM) and noise reducing filter to overcome the problem of the signal enhancement by reducing the missing values and by enhancing the signal on the weak points detected under the application of the HMM. The experimental results have proved the efficiency of the proposed model in comparison with the existing model. The improvement of nearly 50% has been recorded from the parameters of peak signal to noise ratio (PSNR), mean squared error (MSE), signal to noise ratio (SNR) etc.

Keywords: *Speech enhancement, noise reduction, sparse non-negative matrix factorization, hidden markov model, hybrid speech enhancement model.*

*****

1. INTRODUCTION

“Speech Enhancement” refereed as to improve quality or intelligibility of speech signal. Speech signal is often degraded by additive background noise like babble noise, train noise, restaurant noise etc. In such noisy environment listening task is very difficult at the end user. Many times speech enhancement is used for pre processing of speech for computer speech recognition system. Digital noise reduction for audio signals has been an area of investigation since computers became powerful enough to manipulate digital audio in a practical way. Subsequent enhancements and generalizations have been motivated by the stricter fidelity requirements of commercial and archival restoration of old recordings. The speech signal degradations may be attributed to various factors; viz. disorders in production organs, different sensors (microphones) and their placement (hands free), acoustic non-speech and speech background, channel and reverberation effect and disorders in perception organs. Considerable research recently has examined ways to enhance speech, mostly related to speech distorted by background noise (occurring at the source or in transmission)-both wideband (and usually stationary) noise and (less often) narrowband noise, clicks, and other non-stationary interferences.

Speech enhancement aims at improving the performance of speech communication systems in noisy environments. Speech enhancement may be applied, for example, to a mobile radio communication system, a speech recognition system, a set of low quality recordings, or to improve the performance of aids for the hearing impaired. The interference source may be a wide-band noise in the form of a white or colored noise, a periodic signal such as in hum noise, room reverberations, or it can take the form of fading noise. The first two examples represent additive noise sources, while the other two examples represent convolutional and multiplicative noise sources, respectively. The speech signal may be simultaneously attacked by more than one noise source.

Most cases assume noise whose pertinent features change slowly (i.e., locally stationary over analysis frames of interest), so that it can be characterized in terms of mean and variance (i.e., second-order statistics), either during non-speech intervals (pauses) of the input signals or via a second microphone (called reference microphone) receiving little speech input [1]. The auditory system is more sensitive to the presence than absence of energy, and tends to ignore many aspects of phase. Thus speech enhancement algorithms often focus on accurate modelling of peaks in the speech amplitude spectrum, rather than on phase relationships or on energy at weaker frequencies. Voiced speech, with its high amplitude and concentration of energy at low frequency, is more perceptually important than unvoiced speech for preserving quality. Hence, speech enhancement usually emphasizes improving the periodic portions of speech.

In this research, we are proposing a new-age proposed model for the speech signal enhancement and noise removal. At first, a detailed literature survey will be conducted to find the most used and popular speech enhancement algorithms. The selected best algorithms (2 or 3 maximum) will be then implemented using the Matlab simulator. The results analysis will include the best suitable parameters for the selection of the best algorithm for development purposes. The most common parameters among the development professionals are always execution time, accuracy, voice quality and number of trained samples.

2. LITERATURE REVIEW

Berdugo, B. et. al. [1] proposed a new approach called minima controlled recursive averaging (MCRA) for noise estimation. The noise estimate was updated by averaging the past spectral values of noisy speech which was controlled by a time and frequency dependent smoothing factors. These smoothing factors were calculated based on the signal presence probability in each frequency bin separately. This probability was in turn calculated using the ratio of the noisy speech power spectrum to its local minimum calculated over a fixed window time .

Cohen et. al. [2] presented methods that incorporated the fact that speech might not be present at all frequencies and at all times. Authors provided an estimate of the probability that speech is absent at a particular frequency bin. In this research, MMSE magnitude estimator under the assumed Laplacian model and uncertainty of speech presence has been described & considered a two-state model for speech events. According to this two state model, either speech is present at a particular frequency bin (hypothesis H1) or not (hypothesis H(0).

Malah et. al. [6] derived the MMSE STSA estimator, based on modeling speech and noise spectral components as statistically independent Gaussian random variables. Authors a nalyzed the performance of the proposed STSA estimator and compared it with a STSA estimator derived from the Wiener estimator. Authors also examin

Nasser et. al. [8] proposed an improved MCRA noise variance estimator improvements. For objective results, the improvement in segmental SNR was reported for white Gaussian noise, car interior noise and F16 cockpit noise for various noise levels from-5 to 10 dB. In all the cases, the MCRA approach showed a higher performance compared to weighted averaged method. Also, the methods were compared with a subjective study of spectrogram of enhanced speech and informal listening tests. The tracking ability of the algorithms was tested by authors by comparing the spectrograms of enhanced speech for a signal recorded in a car by suddenly turning on the defroster in full.

3. DESIGN AND IMPLEMENTATION

In this section we will compare the results of the input speech signal and after applying HMM algorithm. The original Speech of the signal and the HMM approaches are shown. The results have been obtained in the form of various performance parameters. At very first the original signal has been obtained both before and after enhancement. The signal enhancement is the technique which is used to improve the quality of the speech signal. The speech signal enhancement has been performed by using the amalgamation of the hidden markov model with the non-negative matrix factorization. The non-negative matrix factorization has returned the feature extracted from the speech signal which is further used as the core sample to improve the quality of the speech signal using the HMM model.

3.1. BASIC DESIGN OF THE PROPOSED SYSTEM

The basic design of the proposed system defines the flow of work and also defines the algorithm level design and pseudo code. In flow of work, the voice or speech signal is acquired and loaded into the run time memory. Then the preprocesing technique is used by applying the median filter which is used to remove the salt and pepper noise or Gaussian noise. Non- negative matrix factorization (NMF) is applied to minimize the size of feature descriptor size and then the voice signal vectorization is done and the target features are extracted from the image and marked as the detected objects. The features that are extracted are passed to the SNMF (S) and SNMF (O) with LSA for the signal quality enhancement. In the later step, the enhanced speech signal is returned to the user.

3.2. Hidden Markov Model

The Hidden Markov Model (HMM) is a popular statistical tool for modeling a wide range of time series data. Because Voice is a time domain based data, Hence HMM is used for voice processing i.e. speech recognition, speaker recognition or speaker verification. In the context of natural language processing (NLP), HMMs have been applied with great success to problems such as part-of-speech tagging and noun-phrase chunking. For the Feature Extraction in HMM, the best adaptable algorithm is MFCC (Mel Frequency Cepstral Coefficients).

Probabilistic models such as Hidden Markov Models (HMMs) have also been used for text-independent speaker recognition. These methods suffer in two ways. One is that they require long exemplars for effective modelling. Second, the HMMs model temporal sequencing of sounds,which ‘for text-independent tasks … contains little speaker-dependent information’ (Reynolds and Rose 1995: 73).A different kind of implicit segmentation was pursued in Klevans and Rodman (1997) using a two-level cascading segregating method. Accuracies in the high 90s were achieved in closed-set tests over populations (taken from the TIMIT database) ranging in size from 25 to 65 from similar dialect regions. However, no open-set results were attempted.

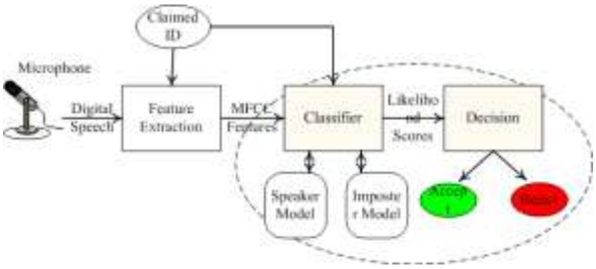

FIGURE 1: Work Flow of Hidden Markov Model for Automatic Speaker Recognition

The hidden markov models are the most successful methods among the stochastic methods produced for the speech enhancement. The speech is divided into the phenomes (the sub-features extracted from the speech signal), and then enhanced using the given HMM algorithm. The HMM is equipped of the data training module before its deployment on any of the speech signal enhancement applications.

Algorithm 1: HMM Model

1. Initialize the process of speech signal with signal acquisition.
2. The following recursive application of HMM is used by using following algorithm design:

a. For time t=1,2,3….N

b. For states s=1,2,3…..M

c. Compute the following

d. $\delta_t(i) = max[\delta_{t-1}(j).\ aji\ ]\ .\ f_i(\boldsymbol{x}(t));$ *max(.) performed over all j;*

e. $\psi_t(i) = arg\{max[\delta_{t-1}(j)\ .\ a_{ji}]\};$ *max(.) performed over all j;*

f. End step b (States)

g. End step a (Time)

3. Retrieve the most likely states in the final form.
4. Retrieve the sequence of the shortlisted states.
5. Return the enhanced speech signal after restructuring the states back in the original waveform for the given signal.

3.3. SUPERVISED SPARSE NON-NEGATIVE MATRIX FACTORIZATION

The idea of using L1-norm regularization for the purpose of achieving sparsity of the solution has been successfully utilized in a variety of problems [18]. We impose the sparsity on the H factor so that it could indicate the clustering membership. The modified formulation is given as:

$$\min_{W,H} \frac{1}{2}\left[\|A - WH^T\|_F^2 + \eta\|W\|_F^2 + \beta\sum_{j=1}^{n}\|H(j,:)\|_1^2\right] \quad s.t. \quad W,H \geq 0$$

where H(j, :) is the i-th row vector of H. The parameter η > 0 controls the size of the elements of W, and β > 0 balances the trade-off between the accuracy of approximation and the sparseness of H. A larger value of β implies stronger sparsity while smaller values of β can be used for better accuracy of approximation. In the same framework of the NMF based on the alternating nonnegative least squares, sparse NMF is solved by iterating the following nonnegativity constrained least square problems until convergence:

$$\min_{H}\left\|\begin{pmatrix} W \\ \sqrt{\beta}e_{1\times k}\end{pmatrix}H^T - \begin{pmatrix} A \\ 0_{1\times n}\end{pmatrix}\right\|_F^2, \quad s.t. \quad H \geq 0$$

where $e_{1\times k} \in R1^{\times k}$ is a row vector having every element as one, and 01×n is a zero vector, and

$$\min_{W}\left\|\begin{pmatrix} H \\ \sqrt{\eta}I_k\end{pmatrix}W^T - \begin{pmatrix} A^T \\ 0_{k\times m}\end{pmatrix}\right\|_F^2, \quad s.t. \quad W \geq 0$$

where I_k is an identity matrix of size k × k and $0_{k\times m}$ is a zero matrix of size k × m.

The non-negative matrix factorization is used to extract the features from the given non-negative form data. The non-negative matrix factorization (NMF) working like the following in the proposed model simulation:

Algorithm 2: Supervised sparse non-negative matrix factorization (S-SNMF)

1. Initialize the W and H factors. Methods for choosing, or seeding, the initial matrices W and H for various algorithms.
2. Uniqueness. Sufficient conditions for uniqueness of solutions to the NMF problem can be considered in terms of simplicial cones.
3. Updating the factors. Devising efficient and effective updating methods when columns are added to the data matrix A.

RESULT ANALYSIS AND COMPARISON

The major objective of this section in this research project is to compare three major speech enhancement techniques i.e. Hidden Markov Model (HMM), Gaussian Mixture Model (GMM) and Vector Quantization (VQ). Vector quantization is a classical quantization technique from signal processing which allows the modeling of probability density functions by the distribution of prototype vectors. The Hidden Markov Model (HMM) is a popular statistical tool for modeling a wide range of time series data. Because Voice is a time domain based data, Hence HMM is used for voice processing i.e. speech recognition, speech enhancement or speaker verification. The Gaussian mixture model (GMM) of speakers described in Reynolds and Rose (1995) is an implicit segmentation approach in which like sounds are (probabilistically) compared with like. The acoustic features are of the mel-cepstral variety (with some other preprocessing of the speech signal).

We have implemented almost 60 percent of the work which includes the MATLAB implementation of Vector Quantization and Hidden Markov Model for the text-independent speech enhancement model. We have used eight training samples for each simulated model.

	TECHNIQUE	MSE	PSNR	T.TIME	PCC	RMSE	NAE
User1. Wav	GMM	0.0153	86.278	0.7810	13055	0.2512	0.1066
	VQ	0.0153	86.278	1.2708	13055	0.3016	0.0733
	HMM-SNMF(OS)	0.00163	95.99	0.7163	61183	0.000	0.000
User2.	GMM	0.0135	86.837	0.7541	14847	0.1783	0.0262
	VQ	0.0135	86.278	1.3358	14847	0.2791	0.0636

Wav							
	HMM-SNMF(OS)	0.0000737	89.45596196	0.888704388	13567	0	0
User3. Wav	GMM	0.0137	86.761	0.8274	14591	0.1842	0.2723
	VQ	0.0137	86.761	1.4917	14591	0.2901	0.0668
	HMM-SNMF(OS)	0.0000751	89.3732367	0.243498893	13311	0	0
User4. Wav	GMM	0.0150	86.362	0.7030	13311	0.0185	0.0401
	VQ	0.0150	86.362	1.3328	13311	0.0185	0.0721
	HMM-SNMF(OS)	0.0000673	89.8474832	0.027378905	14847	0	0
User5. wav	GMM	0.0596	86.375	0.7996	13055	0.1968	0.0406
	VQ	0.05976	86.278	1.2875	13055	0.3034	0.0738
	HMM-SNMF(OS)	0.0000550	90.72578675	0.904163741	18175	0	0
User6. wav	GMM	0.0150	86.362	0.7352	13311	0.1960	0.0304
	VQ	0.0150	86.362	1.2875	13311	0.3034	0.0709
	HMM-SNMF(OS)	0.0000673	89.8474832	0.229676037	14847	0	0
User7. Wav	GMM	0.0135	86.837	0.7659	14847	0.1781	0.0263
	VQ	0.0135	86.837	1.3053	14847	0.2808	0.0632
	HMM-SNMF(OS)	0.0000698	89.69508353	0.669281808	14335	0	0
User8. Wav	GMM	0.0156	86.362	0.7028	13311	0.1989	0.0312
	VQ	0.0156	86.362	1.2877	13311	0.28835	0.0694
	HMM-SNMF(OS)	0.0000685	89.77195182	0.366673251	14591	0	0

Table 1: Audio File Quality Comparison

	Signal to Noise Ratio			
Signal Index	0dB	5dB	10dB	15dB
1	6.081368	7.156993	8.296138	9.471909
2	6.030599	6.788403	7.613279	8.602677
3	6.260992	7.328589	9.284813	9.575865
4	5.883268	6.742693	7.56469	8.644814
5	5.942089	6.777363	8.203958	8.912068
6	6.148561	7.115684	8.002534	8.945226
7	6.007926	8.172346	8.166489	9.08904
8	5.958074	6.616629	7.413572	8.468705
9	7.074829	7.802071	8.143961	8.815945
10	5.694435	6.434984	7.435729	8.276943
11	5.963504	7.454314	8.933787	8.977834
12	6.331815	6.921058	7.672662	8.405754

13	5.822206	6.551134	8.37752	8.767295
14	7.607185	7.25425	8.613513	9.516134
15	7.732839	8.010687	8.408529	9.373987
16	6.140482	6.871498	8.184132	9.144828
17	5.850217	6.490764	7.714079	8.518309
18	6.305775	7.095695	8.502272	9.14922
19	6.218061	7.327009	8.426445	9.565741
20	5.562342	6.412198	7.410401	8.401402
21	6.261457	7.975026	8.487629	9.231393
22	5.9838	6.866855	8.287938	9.20627
23	7.51993	6.990357	8.082636	8.957059
24	6.326411	7.180318	8.79208	9.10966
25	7.363006	7.812679	8.830428	9.187847
26	7.138356	6.921424	7.49627	8.062626
27	7.160928	6.624817	7.518727	8.257745
28	5.980055	7.577008	8.450533	8.815203
29	6.110755	7.714168	8.50345	8.77047
30	7.192342	6.609294	8.311231	8.560891

Table 2: The SNR based evaluation of data with different levels of noise

SNR	**Noisy**	**NMF**	**SNMF (S)**	**SNMF (T)**	**SNMF (ST)**	**SNMF(OS)**
0 dB	1.30	1.67	**1.70**	1.67	1.70	6.42
5 dB	1.77	2.07	2.10	2.10	**2.12**	**7.16**
10 dB	2.06	2.32	2.38	2.34	**2.39**	**8.17**
15 dB	2.39	2.53	2.60	2.56	**2.63**	**8.91**

Table 3: The independent analysis of the speech enhancement schemes

SNR	LSA	LSA+NMF	LSA+SNMF (S)	LSA+SNMF (ST)	HMM-SNMF(OS)
0 dB	1.68	1.96	1.94	1.98	6.49
5 dB	2.23	2.36	2.42	2.43	7.59
10 dB	2.48	2.58	2.62	2.65	9.16
15 dB	2.80	2.81	2.88	2.89	9.84

Table 4: The result analysis of speech enhancement schemes in combination

CONCLUSION AND FUTURE WORK

The proposed model has been equipped with the supervised sparse non-negative matrix factorization (S-SNMF) model to reduce the size of the signal being improved in order to enhance the execution speed of the proposed model. In the proposed model, the S-SNMF has been improved and developed as the improvement in the baseline sparse non-negative matrix factorization (SNMF) model. For the better performance, the proposed model has been combined with the intelligent hidden markov model (HMM) algorithm to fix the missing elements in the speech signal in order to apply the smoothen and enhanced speech signal and to reduce the noise levels in the speech signal. The proposed model has been found efficient in the terms of performance parameters of PSNR, SNR, MSE, PCC, RMSE etc. The proposed model has been found with almost SNR of 9.8

against the previous SNR of 2.8 on the signal with 15dB of noise

REFERENCES

[1] Berdugo, B. and Cohen, I.,―Noise estimation by minima controlled recursive averaging for robust speech enhancement , IEEE Signal Proc. Letters, vol. 9, no. 1, pp. 12-15, Jan. 2002.

[2] C. Breithaupt and R. Martin, ―MMSE estimation of magnitude-squared DFT coefficients with super-gaussian priors,‖ in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 848-851, 2003.

[3] H. Xuchu, S. Guo, H. Cui, K. Tang and Y. Li, ―Speech Enhancement for Non-Stationary Noise Environments‖, International Conference on Information Engineering and Computer Science, ICIECS 2009, pp. 1-3, 19-20 Dec. 2009.

[4] *Hao-Teng Fan1, Jeih-weih Hung 1, Xugang Lu2, Syu-Siang Wang3, and Yu Tsao3,"* Speech enhancement using segmental non negative matrix factorization, 2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP).

[5] I. Cohen, ―Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator,‖ IEEE Signal Processing Lett., vol. 9, pp. 113 -116, Apr. 2002.

[6] Malah, D., Cox, R.V. and Accardi, A.J., ―Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments,‖ Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp.789-792,15-19 Mar 1999

[7] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 208-211, 1979.

[8] Nasser IEEE, "Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2140 – 2151, oct 2013.

[9] R. Martin and C. Breithaupt, ―Speech enhancement in the DFT domain using Laplacian speech priors,‖ in International Workshop on Acoustic Echo and Noise Control (IWAENC), pp. 87–90, Sept. 2003.

[10] R. Martin, ―Speech enhancement using a minimum mean-square error short-time spectral amplitude estimation,‖ in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 504—512, 2002.

[11] S. Kamath, P. Loizou, "A multi-band spectral subtractionmethod for enhancing speech corrupted by colored noise,"in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. IV-4164, 2002.

[12] S. Boll, "Suppression of acoustic noise in speech using spectralsubtraction," *IEEE Transactions on Acoustics, Speechand Signal Processing*, 27(2), pp. 113–120, 1979.

[13] Y. Ephraim,‖ Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator‖, IEEE Transactions On Acoustics, Speech and Signal Processing 0096-35 18/8S/0400-0443, 1985.

[14] Y. Hu and P. C. Loizou, "Evaluation of objective qualitymeasures for speech enhancement," *IEEE Transactions onAudio, Speech, and Language Processing*, 16(1), pp. 229-238, 2008.4520

[15] Tomi Kinnunen, Ilja Sidoroff, Marko Tuononen, Pasi Fränti, Comparison of clustering methods: A case study of text-independent speaker modeling, ScienceDirect, 2011.

[16] Cemal Hanilci, Figen Ertas, Comparison of the impact of some Minkowski metrics on VQ/GMM based speaker recognition, ScienceDirect, 2010.

[17] Tomi Kinnunen, Juhaani Saastamoinen, Ville Hautamaki, Mikko Vinni, Pasi Franti, Comparative evaluation of maximum of Posteriori vector quantization and gaussian mixture models in speaker verification, ScienceDirect, 2009.

[18] Man-Wai Ma, Hon-Bill Yu, A study of voice activity detection techniques for NIST speaker recognition evaluations, ScienceDirect, 2013.