Compressive Sensed Speech Recognition

Usham V. Dias[1], Jeswil E. Mascarenhas[2], and Lioshka J. Dias[3]
[1]Dept. of Electrical Engineering, Indian Institute of Technology, Delhi, India
[2]Padre Conceicao College of Engineering, Verna, Goa, India
[3]Siemens, Energy Automation, Goa, India
Email: {ushamdias, jeswil.2693, lioshkajoelladias}@gmail.com

***Abstract*—This paper implements cepstral feature classification of compressive sensed speech signal and compares it with results obtained without compressive sensing. The Orthogonal Matching Pursuit algorithm is used for reconstruction with Gaussian as the sensing matrix and Discrete Cosine Transform as the sparsifying basis. The paper also uses post processing of compressive sensed signal to improve the accuracy of classification. The K-Nearest Neighbor classifier was tested for different distance models, with City Block giving the highest accuracy of 57.14% and 91.43%, with and without compressive sensing respectively. Post-Processing of compressive sensed signal using a median filter improves the accuracy from 57.14% to 80%.**

***Index Terms*—compressive sensing, speech recognition, orthogonal matching pursuit, K-nearest neighbor, cepstrum**

I. Introduction

Compressive sensing is a data acquisition technique which tries to acquire data as linear projections below Nyquist rate. It was proposed by Donoho and Candes [1], [2], as was mainly aimed to be an imaging technique. But its application for 1-D signal acquisition is promising [3], [4]. Audio compressive sensing using four sensing patterns namely Gaussian, Bernoulli+/-1, Bernoulli 0/1 and Hadamard was demonstrated in [5]. It assumed Discrete Cosine Transform (DCT) as the sparsifying basis and Orthogonal Matching Pursuit as the reconstruction algorithm.

This paper focuses on classification of the words uttered which was acquired assuming the Compressive sensed paradigm.

II. Speech Recognition

Based on the type of utterances that can be recognized, a speech recognition system can be classified as system for Isolated words, Connected words, Continuous speech and Spontaneous speech [6]. Feature extraction for speech can be classified into temporal and spectral techniques. Temporal techniques include Power estimation and Fundamental frequency estimation while Spectral techniques include Critical Band Filter Bank Analysis, Cepstral Analysis, Mel Cepstrum Analysis, Linear Predictive Coding (LPC) Analysis and Perceptually Based Linear Predictive Analysis [7].

Manuscript received August 10, 2015; revised April 10, 2016.

This paper uses cepstral analysis for feature extraction of isolated words as shown in Fig. 1. Speech signal is composed of an excitation source and a vocal tract system. The output speech is the convolution of the source and vocal tract response, which needs to be separated from the speech for analysis. The objective of *cepstral analysis* is to achieve this separation without any a priori knowledge about the source and/or system [8].

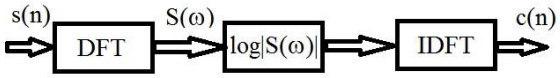

Figure 1. Block diagram for cepstral coefficient computation

Let $e(n)$ be the excitation sequence given to the vocal tract filter sequence h(n); then the speech sequence s(n) is given by

$$s(n) = e(n) * h(n) \tag{1}$$

Equation (1) can be represented as a multiplication operation in the frequency domain, given by (2) where E(ω), H(ω) and S(ω) are the Fourier transform of e(n), h(n) and s(n) respectively.

$$S(\omega) = E(\omega)H(\omega) \tag{2}$$

The magnitude of the spectrum of S(ω) given in (3) is further represented logarithmically given by (4) which converts the multiplication operation to a summation operation in the frequency domain.

$$|S(\omega)| = |E(\omega)||H(\omega)| \tag{3}$$

$$\log|S(\omega)| = \log|\mathrm{E}(\omega)| + \log|H(\omega)| \tag{4}$$

Separation is done by taking the IDFT of the log spectra given in (4), which transforms it to the frequency domain or cepstral domain which is similar to time domain. The cepstral domain representation of the signal is given in (5).

$$\begin{aligned} \mathrm{c(n)} &= \mathrm{IDFT}(\log|S(\omega)|) \\ &= \mathrm{IDFT}(\log|\mathrm{E}(\omega)| + \log|H(\omega)|) \\ &= ec(n) + hc(n) \end{aligned} \tag{5}$$

The vocal tract component is represented by the slowly varying components concentrated near the lower frequency region while the excitation component is represented by the fast varying components at the higher frequency region.

doi: 10.18178/ijsps.4.6.483-486

Some of the classifiers that can be used for classification of speech features include template matching, K-Nearest Neighbor algorithm (KNN), Artificial Neural Network (ANN) and Fuzzy classifiers. This paper uses KNN for classification with different distance metrics like Euclidean distance, City block, Cosine distance and Correlation distance given by (6)-(9) respectively:

$$D = \sqrt{\sum_{k=0}^{n}(\mathrm{X}k - Yk)^2} \quad (6)$$

$$D = \sum_{k=0}^{n} |\mathrm{X}k - \mathrm{Y}k| \quad (7)$$

$$D = \left(1 - \frac{XY'}{\sqrt{(XX')\,(YY')}}\right) \quad (8)$$

$$D = 1 - \frac{(X-\bar{X})(Y-\bar{Y})'}{\sqrt{(X-\bar{X})(X-\bar{X})'}\sqrt{(Y-\bar{Y})(Y-\bar{Y})'}} \quad (9)$$

where X=(X1, X2, …., Xn) and Y=(Y1, Y2, …., Yn) are the points between which distance is measured while $\bar{X}$ and $\bar{Y}$ are the mean of X and Y respectively.

The K-Nearest Neighbor Algorithm is described below [9]. A training set is created which stores the features of the speech that needs to be classified.

Let the training set be **Xmxn** where **m** is the number of feature vectors stored and **n** is the dimension of the feature vector.

Let the $\mathbf{x_i}=[\mathbf{x_i^{(1)}}\ \mathbf{x_i^{(2)}}\ \ldots..\ \mathbf{x_i^{(n)}}]$ be the n-dimensional feature vector of real numbers for $1< i < m$.

Let $\mathbf{y_i}$ be the class label **{1 … C}** for all i, where C is the number of classes.

Let $\mathbf{x_{new}}$ be the new feature vector at the input whose class, $\mathbf{y_{new}}$, has to be determined.

Find k closest vectors to $\mathbf{x_{new}}$ w.r.t a distance metric given by equation 6 through 9 and store their class in $\mathbf{y_{new}}$.

Final classifier output $\mathbf{y^{KNN}}$=majority vote among the classes stored in $\mathbf{y_{new}}$ based on the k-nearest points.

III. Compressive Sensing

Based on the initial work carried out in [5], this paper also tries to acquire the input using the compressive sensing paradigm as shown in Fig. 2. The sensing matrix used is Gaussian and the sparsifying basis assumed is DCT. An acquisition window of size 256 was used to capture the signal.

$$\boldsymbol{Y} = \boldsymbol{\Phi X} \quad (10)$$

In (10), 'Φ' is the Gaussian sensing matrix, X is the windowed speech signal i.e. 256 samples which is assumed to be sparse in the DCT domain and Y is the acquired projections or measurements which is much less in dimension than the dimension of the signal X.

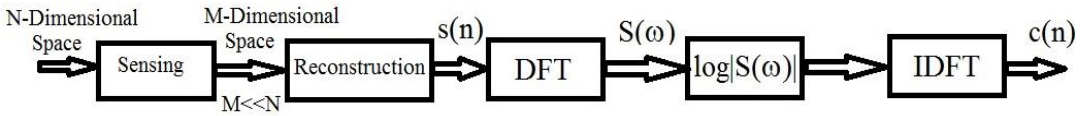

Figure 2. Compressive sensed speech acquisition and cepstral computation

The Orthogonal Matching Pursuit Reconstruction Algorithm is used to reconstruct X from Y [10] and is described below.

Let Φ be the sensing matrix.

Let Ψ the sparsifying basis.

Let $A=\Phi\Psi^{-1}$ be an MxN matrix.

Let $A_\lambda \in R^M$ be the λth column of A.

Let $A_S \in R^{MxS}$ be the Matching Pursuit estimation matrix where S is the sparsity which is also equal to the number of iteration.

Let $y \in R^M$ be the acquired data

Let $yp \in R^M$ be the estimate of y.

Let $yr \in R^M$ be the residual error in estimating y.

Let $x' \in R^N$ be the estimate of x.

Let $x'_k \in R$ be the k^{th} element of x'.

1. $A_S = 0$, yr=y, k=0.
2. k=k+1;
3. Take dot product of yr with every column of A and find the index λ which corresponds to the highest dot product value and which was not selected earlier.
4. Augment A_S with A_λ
5. Perform least squares estimation to calculate x' using A_S.

$$x' = (A_S^T A_S)^{-1} A_S^T y$$

6. Take the projection of x' onto A_S to calculate yp

$$yp = A_S x'$$

7. Update the residual yr

$$yr = y - yp$$

8. Perform steps 2 to 7 until k=s.

Finally we get x' which is the sparse signal in Ψ domain using sensing matrix Φ.

In [11], it was found that median filtering of the reconstructed data based on OMP provided quality improvement to the tune of 200 extra measurements per block in terms of relative error. This idea is used to provide higher classification accuracy in this paper as shown in Fig. 3.

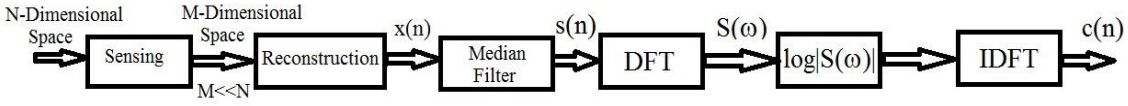

Figure 3. Improved compressive sensed speech reconstruction using median filter

IV. Experimental Results

The database created consists of seven words viz. Forward, Reverse, Right, Left, Up, Down and Stop. Each word has 10 samples, 5 samples are used for training and the remaining 5 for testing. Audio reconstruction of compressive sensed data is done for block size of 256 at a time. A total of 43 blocks are reconstructed for one voice sample which represents the word. Gaussian sensing matrix and Orthogonal Matching Pursuit reconstruction algorithm was used. Discrete Cosine Transform is used as the sparsifying basis. For the database created the average sparsity to preserve 99.9% of the energy was found to be 22 coefficients per block, hence sparsity was assumed to be 22 per block and measurements taken equals four times the sparsity [5]. The average reconstruction error and average time for reconstruction is **0.218821** and **0.133487** respectively. The classification was performed using K-Nearest Neighbour algorithm using different distance models. The number of votes considered (K) is five and three for comparison. The classification accuracy using KNN v/s cepstrum feature size for different distance models in KNN algorithm is shown in Fig. 4-Fig. 7.

For K=5, highest classification accuracy of 91.43% is achieved using City Block distance model with Cepstrum feature size of 60 coefficients as indicated in Fig. 4. For K=3, highest accuracy of 91.43 is achieved using City Block distance model with Cepstrum feature size of 30 coefficients as indicated in Fig. 5.

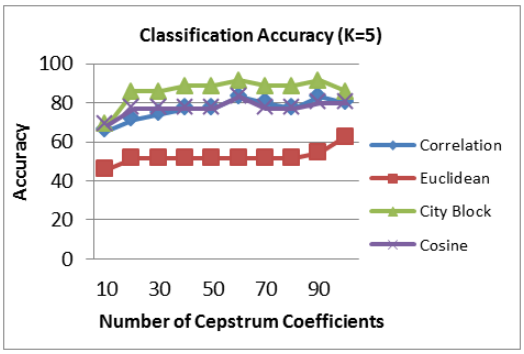

Figure 4. Classification accuracy for database with K=5.

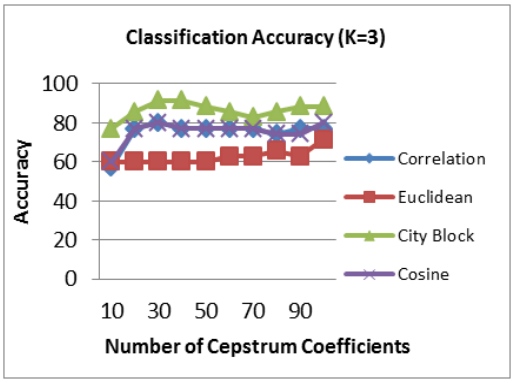

Figure 5. Classification accuracy for database with K=3.

In case of Compressive Sensed Audio signal classification using KNN with K=5, the highest accuracy achieved was 54.29% using City Block and Correlation distance model with cepstrum feature size of 50 and 30 coefficients respectively as indicated in Fig. 6. For K=3, highest accuracy of 57.14% is achieved using City Block distance model with cepstrum feature size of 30 coefficients as indicated in Fig. 7.

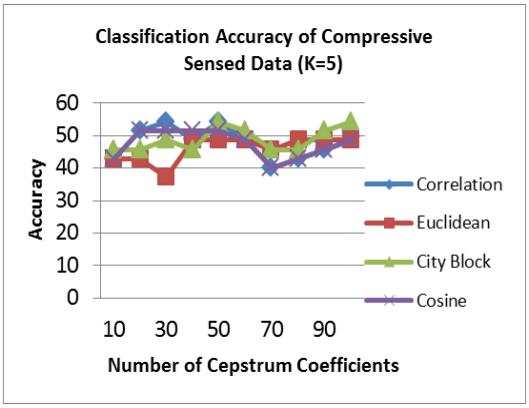

Figure 6. Classification accuracy for data acquired by compressive sensing, K=5.

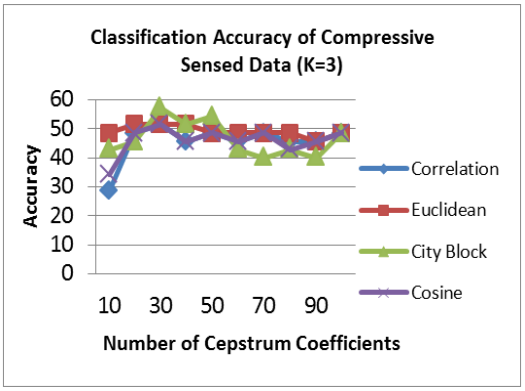

Figure 7. Classification accuracy for data acquired by compressive sensing, K=3.

The compressive sensed data was filtered using a median filter of size 5, to improve the quality of the signal below filtering. The results for K=5 and K=3 are shown in Fig. 8-Fig. 9. The highest accuracy obtained was 80% in both cases for city block distance with a feature size of 100 and 60 respectively.

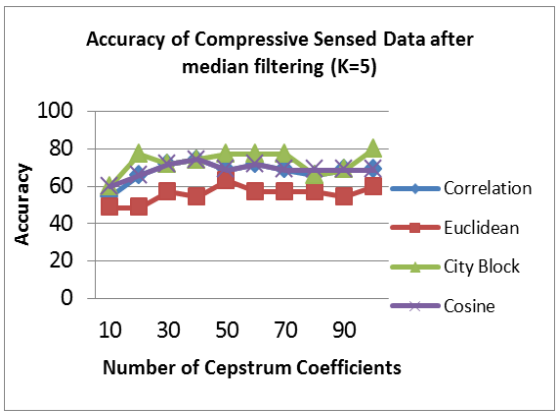

Figure 8. Classification accuracy for data acquired by compressive sensing followed by median filtering, K=5.

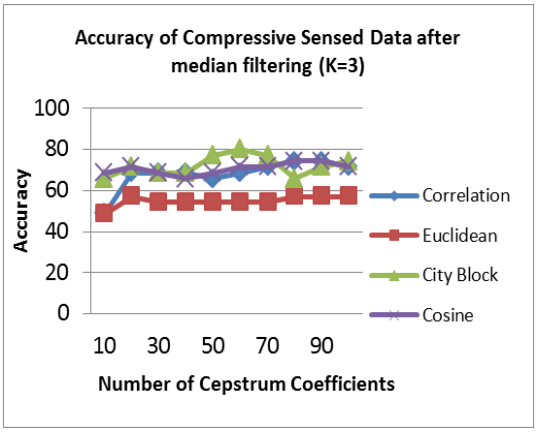

Figure 9. Classification accuracy for data acquired by compressive sensing followed by median filtering, K=3.

V. Conclusion

Based on the results obtained, City Block distance model for classification using KNN gave the best possible result in all cases. The best combination was the use of K=3 with 30 cepstrum coefficients giving an accuracy of 91.43%. In case of compressive sensed Audio classification, the same combination worked giving an accuracy of 57.14%. The classification accuracy can be considerably increased by filtering the reconstructed data. In the work carried out, an accuracy of 80% is achieved by the use of median filter as opposed to the highest of 57.14% without filtering.

Acknowledgment

This research was done while the authors were a part of the Dept. of Electronics and Telecommunication, Padre Conceicao College of Engineering, Verna-Goa.

References

[1] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289-1306, 2006.
[2] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203-4215, 2005.
[3] M. G. Christensen, J. Ostergaard, and S. H. Jensen, "On compressed sensing and its application to speech and audio signals," in *Proc. IEEE Forty-Third Asilomar Conference on Signals, Systems and Computers*, 2009, pp. 356-360.
[4] D. Wu, W. P. Zhu, and M. N. S. Swamy, "A compressive sensing method for noise reduction of speech and audio signals," in *Proc. IEEE 54th International Midwest Symposium on Circuits and Systems*, 2011, pp. 1-4.
[5] U. Dias; "Audio compressive sensing using orthogonal matching pursuit algorithm," *International J. of Multidispl. Research & Advcs. in Engg.*, vol. 6, no. 1, pp. 129-135, January 2014.
[6] S. S. Therese and C. Lingam; "Review of feature extraction techniques in automatic speech recognition," *International Journal of Scientific Engineering and Technology*, vol. 2, no. 6, pp. 479-484, 2013.
[7] M. P. Kesarkar and S. P. Rao, "Feature extraction for speech recognition," M.Tech. Credit Seminar Report, Electronic Systems Group, EE. Dept., IIT Bombay, submitted November 2003.
[8] Sakshat virtual labs, IITG. (May 2014). [Online]. Available: http://iitg.vlab.co.in/?sub=59&brch=164&sim=615&cnt=1
[9] X. Zhu. K-nearest-neighbor: An introduction to machine learning. Computer Sciences Department University of Wisconsin, Madison. [Online]. Available: http://www.cs.sun.ac.za/~kroon/courses/machine_learning/lecture 2/kNN-intro_to_ML.pdf
[10] U. Dias and M. E. Rane, "Block-based compressive sensed thermal image reconstruction using greedy algorithms," International Journal of Image, Graphics and Signal Processing, vol. 10, no. 10, pp. 36-42, 2014.
[11] U. Dias and S. Patil, "Compressive sensing based microarray image acquisition," in *Proc. International Conference for Convergence of Technology*, Pune, India, April 2014.

Usham V. Dias has received his Bachelor's Degree in Electronics and Telecommunication from Padre Conceicao College of Engineering, Verna, Goa in 2009 and Master's Degree in Electronics and Telecommunication from Vishwakarma Institute of Technology, Pune, Maharashtra in 2013.
He is currently pursuing his PhD at the Indian Institute of Technology, Delhi. He was previously employed as an Assistant Professor on contract in the Dept. of Electronics and Telecommunication, Padre Conceicao college of Engineering, Verna-Goa. His research interests include Compressive Sensing, Sub-Nyquist sampling, Signal & Image processing.

Jeswil Evrald Mascarenhas has received his Bachelor's Degree in Electronics and Telecommunication from Padre Conceicao College of Engineering, Verna, Goa in 2015.
He has worked as an intern at Bosch Packaging Limited, Verna, Goa. His research interests include Speech signal & Image processing, Embedded Systems and VLSI technologies.

Lioshka Joella Dias has received her Bachelor's Degree in Electronics and Telecommunication from Padre Conceicao College of Engineering, Verna, Goa in 2015.
She currently works as a Graduate Trainee Engineer at Siemens, Energy Automation, Goa. She has previously worked as an intern at Bosch Packaging Limited, Verna Goa. Her research interests include Speech signal & Image processing.