NON-LINEAR DIMENSION REDUCTION OF GABOR FEATURES FOR NOISE-ROBUST ASR

Hitesh Anand Gupta, Anirudh Raju, Abeer Alwan

Department of Electrical Engineering, University of California Los Angeles, USA
{hiteshag, anirudh90}@ucla.edu, alwan@seas.ucla.edu

ABSTRACT

It has been shown that Gabor filters closely resemble the spectro-temporal response fields of neurons in the primary auditory cortex. A filter bank of 2-D Gabor filters can be applied to either the mel-spectrogram or power normalized spectrogram to obtain a set of physiologically inspired Gabor Filter Bank Features. The high dimensionality and the correlated nature of these features pose an issue for ASR. In the past, dimension reduction was performed through (1) feature selection, (2) channel selection, (3) linear dimension reduction or (4) tandem acoustic modelling. In this paper, we propose a novel solution to this issue based on channel selection and non-linear dimension reduction using Laplacian Eigenmaps. These features are concatenated with Power Normalized Cepstral Coefficients (PNCC) to evaluate if the two are complementary and provide an improvement in performance. We show a relative reduction of 12.66% in the WER compared to the PNCC baseline, when applied to the Aurora 4 database.

Index Terms— Gabor filter-bank, Laplacian Eigenmaps, Multi-layer perceptron

1. INTRODUCTION AND RELATED WORK

There have been several methods to improve the robustness of automatic speech recognition (ASR) performance in the case of variability in the speech signal. Sources of this variability are attributed to extrinsic sources (e.g. background noise, channel noise) and intrinsic sources (e.g. speaking rate, gender, age, mood of speaker, etc.). Although human auditory perception is highly robust to most of these variations, ASR systems are not.

Significant research has been carried out to identify spectro-temporal receptive fields (STRFs) which are an approximation for the spectral-temporal representation of the sound that "excites" a neuron. The use of physiologically inspired 2-dimensional Gabor filters to approximate the STRF of neurons in the primary auditory cortex was first proposed in [1].

The Gabor filter-bank is used to extract a wide range of spectro-temporal modulation frequencies from a speech signal. The filter bank output is high dimensional, with correlated information. Typical HMM/GMM based back end systems use GMMs with diagonal covariance matrices, and hence require the features to be uncorrelated. In addition, a big challenge here is to find a low dimensional representation of the important spectro-temporal information, while limiting redundancy.

In the past, several methods have been proposed to perform dimension reduction on the $\text{GBFB}_{\text{full}}$ (full Gabor Filter Bank output):

1. Feature Selection Algorithm [1] - use a Feature Finding Neural Network (FFNN) [2].

2. Channel Selection [3] - selects a representative set of frequency channels by utilizing the fact that filter outputs between neighbouring channels are correlated when the filter has a large spectral extent. From here on, the Gabor filter bank output after channel selection is referred to as "GBFB" features.

3. Linear Dimension Reduction [3] - Principal Component Analysis (PCA) is applied to either GBFB or $\text{GBFB}_{\text{full}}$ to reduce dimension and decorrelate the features.

4. Tandem Acoustic Model - The high dimensional Gabor Features (either $\text{GBFB}_{\text{full}}$ or GBFB) rather than MFCC features, which were first used in a tandem framework [4], are used as an input to a Multi-layer Perceptron (MLP) that is trained to produce phoneme posteriors [5]. These are then decorrelated and treated as acoustic feature vectors (typically concatenated with other standard features). Though this isn't traditionally thought of as a dimension reduction method, it is in essence, moving the GBFB features to the low dimensional phoneme posterior space.

   The Tandem HMM system has also been used in a stream based setting [6][7], where the features are split up into streams based on the location in the spectro-temporal response field. Each stream is independently processed with an MLP to produce phone posteriors. These are concatenated and then reduced to a lower dimension using PCA.

Work supported in part by DARPA via SRI.

Typically, either one or many of the above methods are used as a solution to deal with the high dimensionality of the $GBFB_{full}$ features. In the following sections, a non-linear dimension reduction technique based on the application of Laplacian Eigenmaps to the GBFB features is proposed as a novel solution to this problem. The novel features proposed in this paper are referred to as $GBFB_{LE}$. This is different from previous work as the GBFB features are not being used in a Tandem Acoustic Model [8]. Moreover, previous efforts that perform dimension reduction on the features only involve linear methods such as PCA [3].

The $GBFB_{LE}$ features are concatenated with PNCC to evaluate whether these features are complementary and provide a boost in performance. The recognition performance of the proposed system is compared with $GBFB_{PCA}$ and PNCC baseline. The proposed method provides a 12.66%(relative) reduction in word error rate compared to the PNCC baseline on the Aurora 4 database.

It is imperative to mention that, in the past, non-linear dimension reduction techniques have been applied to speech data, on either log power-spectra [9], or on the MFCCs [10]. However, the effect of non-linear dimension reduction on the Gabor Filter Bank feature space has not been explored before. Moreover, this paper involves a novel usage of an MLP as a solution to the out-of-sample extension problem which is encountered in graph based non-linear dimension reduction techniques.

In Section 2, the proposed method is explained followed by the experimental setup in Section 3. The results along with their interpretations are presented in Section 4, and a summary is provided in the final section.

2. PROPOSED METHOD

There are two main parts of the proposed system as shown in Figure 1. The left portion of the figure deals with the feature extraction and is described in Section 2.1. The one on the right is the offline training phase where a mapping from a high dimensional Gabor space to a low dimensional manifold is learnt, which is used in the feature extraction phase. This is described in Section 2.2.

2.1. Power Normalized Gabor and Cepstrum Features

Gabor features are computed by processing the spectro-temporal representation of an input signal by several two-dimensional modulation filters. This filter bank is convolved with the spectrogram in order to extract the Gabor features. It is relatively robust to both extrinsic variability (additive noise), and against intrinsic variability (speaker variability) [3]. A Gabor filter is the product of an envelope function and a complex sinusoid carrier, as described in [3]. Further details can be found in that paper.

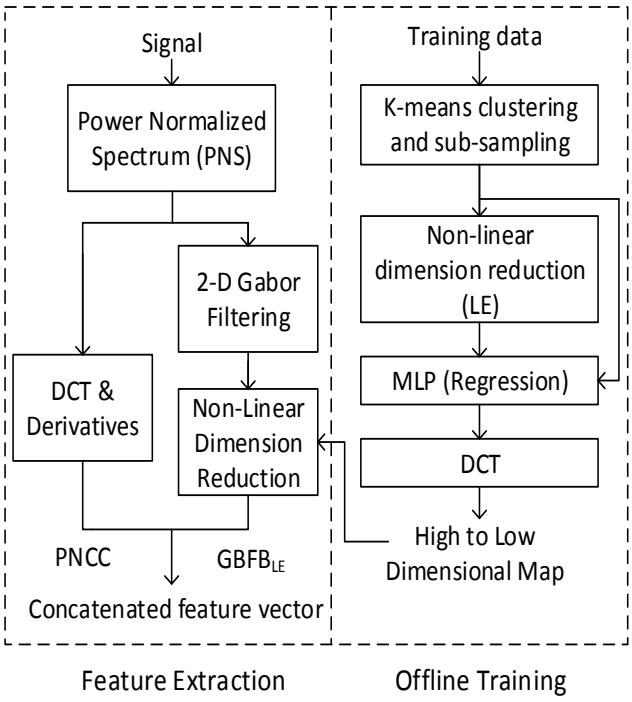

Fig. 1. Flowchart of proposed system

These Gabor Filters can be applied to any spectrogram, such as the mel-spectrogram or the Power Normalized Spectrogram (PNS). The PNS involves the use of a mel filter bank, "medium-time" and "short time" processing, and power-law non-linearity as proposed in [11].

The Gabor features are obtained from the PNS as follows:

1. 2D Gabor filtering is applied to the PNS to obtain the full Gabor feature ($GBFB_{full}$) vector. It was shown in [8] that Gabor filters applied to a PNS performs better in comparison to mel-spectrogram.

2. $GBFB_{full}$ is reduced to a 564 dimensional vector (GBFB feature) through channel selection as described in [3]. Mean Variance Normalization (MVN) is carried out on this feature vector.

3. The high dimensional GBFB feature is projected onto a low dimensional manifold using a mapping which is learnt in the offline training phase, described in Section 2.2.2. The resultant feature is the $GBFB_{LE}$ feature.

A 39 dimensional PNCC feature vector is obtained by concatenating the static coefficients with the velocity and acceleration coefficients. $GBFB_{LE}$ and PNCC (39 dim) are concatenated to obtain a final feature vector. The dimension of $GBFB_{LE}$ is selected based on empirical experiments described in Section 4.

2.2. Non-Linear Dimension Reduction

Laplacian Eigenmaps is used to project the high dimensional GBFB feature onto a low dimensional manifold and the mapping to the low dimension is learnt using an MLP.

2.2.1. Laplacian Eigenmaps

To discover the intrinsic low dimensional manifold that the high dimensional data lie in, a graph based non-linear dimension reduction technique known as Laplacian Eigenmaps (LE) [12] is used. It has locality preserving properties based on pairwise distances between neighbouring points.

The algorithm as given in [12], is outlined below.

1. Build a graph using the given data points. Construct the adjacency matrix (W) by connecting nodes *i* and *j* if *i* is within *N* neighbours of *j*, or *j* is within *N* neighbours of *i* (*N*-nearest neighbour criterion). The edge weights are set to 1 or 0 based on whether the two nodes are connected or not.

2. Using the adjacency matrix W, the Laplacian matrix L is computed as $L = D - W$, where D is the degree matrix (diagonal matrix with entries equal to row sum of W). The eigenvectors (f) and eigenvalues (λ) of the generalized eigenvector problem of L are computed using:
$$Lf = \lambda Df \quad (1)$$

3. Let the eigenvector solutions from the previous step be f_0...., f_m ordered according to their eigenvalues (in increasing order). Discard the eigenvector with eigenvalue 0 and use the next m eigenvectors to embed the data in an m-dimensional space.

   These f_is i.e. $(f_1$...., $f_m)$ are the new data points embedded in an m-dimensional space.

In order to reduce the computational complexity of the Laplacian Eigenmaps dimension reduction, it is performed on a representative subset of the total training data. This representative subset is obtained through *k*-means clustering on the training data. Once the low dimensional representation has been obtained for this data subset, the features are decorrelated using DCT.

2.2.2. Multi-Layer Perceptron for out-of-sample extension

The issue with Laplacian Eigenmaps is that the data points are directly embedded in a low dimensional space, without providing an explicit mapping to project onto the new space. Hence, it cannot embed new data points into this low dimensional manifold. The problem of embedding new data points onto this low dimensional space is usually referred to as the out of sample extension problem. In the past, the Nystrom Approximation [13] has been used as a solution to this problem. However, in this work, an MLP is used in a regression setting for learning a non-linear mapping between the input (high dimensional feature) and output (low dimensional feature). This is done by using a linear activation function in the output layer, as opposed to a sigmoid function, which provided lower WERs than the Nystrom approximation.

3. EXPERIMENTAL SETUP

The proposed system is tested on the Aurora4 medium vocabulary speech database which is derived from the WSJ0 5k-word closed vocabulary dictation task. The training and testing data consist of 7138 utterances from 83 speakers and 330 utterances from 8 speakers respectively. The test set consists of the following noise types: airport, babble, car, restaurant, street and train, with SNRs ranging from 5 to 15 dB. Clean speech is used for training whereas both clean and noisy speech is used for testing. The HMM Toolkit [14] is used for both training and testing. The HMMs are modelled as cross-word triphones, each with 3 states and 8 Gaussian mixtures with diagonal covariance per state. The standard 5k bigram language model is used for the evaluation.

The PNS is obtained using a 40 channel mel filter-bank and a compression factor of 1/15 for the non-linearity, instead of log compression. PNCCs are obtained by first performing DCT on PNS and retaining the first 13 coefficients followed by MVN. The first and second derivatives are computed (13 dimensions each) and concatenated with the static coefficients to give the PNCC feature vector.

2D Gabor filtering is applied on the PNS using spectral and temporal distance values of 0.3 and 0.2 [3] which results in a 564 dimensional feature vector after channel selection. MVN is also carried out on the GBFB feature. *k*-means clustering is performed on a concatenated feature matrix of the MFCC features using 1000 clusters. MFCC are used for clustering in place of PNCC as they are a better cepstrum representation for clean speech, which is used for training. We use a 5% random sample of each cluster of MFCC points, and use the corresponding GBFB features for the offline training. Laplacian Eigenmaps is used to reduce the dimension of data by choosing the *n* smallest eigenvectors. The value of *n* is determined empirically in the next section. The adjacency matrix is built using the 12 nearest neighbours to each point. The mapping to the low dimensional space is learnt using an MLP, which consists of a single hidden layer with 400 nodes. The cost function used was the regularized mean square error.

4. RESULTS AND DISCUSSION

The PNCC (with MVN) baseline is compared with different configurations based on the Gabor features. The GBFB features reduced using PCA and LE are referred to as GBFB_{PCA} and GBFB_{LE} respectively. In Table 1, the WERs are shown

WER (in %)	Gabor Dim.	Total Dim.	Clean	Airport	Babble	Car	Restaurant	Street	Train	Average
(1) PNCC (with MVN)	0	39	12.65	34.00	34.82	18.44	38.15	35.14	36.33	29.93
(2) $GBFB_{PCA}$	39	39	17.67	38.20	40.54	21.86	43.56	42.18	41.29	35.04
(3) $GBFB_{PCA}$ + PNCC	27	66	15.54	32.92	34.06	20.36	37.75	36.35	36.58	30.50
(4) $GBFB_{LE}$ + PNCC	27	66	12.33	31.57	33.46	16.78	35.91	34.00	34.75	28.40
(5) $GBFB_{LE}$ + PNCC	13	52	11.83	**29.78**	31.87	**15.65**	**33.16**	32.26	32.60	26.73
(6) $GBFB_{LE}$ + PNCC	10	49	11.99	29.98	31.18	15.84	34.22	31.72	33.03	26.85
(7) $GBFB_{LE}$ + PNCC	7	46	**11.23**	30.08	**30.45**	15.67	33.36	**30.56**	**31.66**	**26.14**

Table 1. Results comparing WERs of proposed system with PNCC baseline, including first and second derivatives, and $GBFB_{PCA}$ on the Aurora 4 task. Bold numbers represent best performance.

for the PNCC (with MVN) baseline (1), PCA reduced Gabor features ($GBFB_{PCA}$) (2), $GBFB_{PCA}$ concatenated with PNCC (3) and LE reduced Gabor features ($GBFB_{LE}$) concatenated with PNCC (4-7), in clean and noisy speech. The $GBFB_{PCA}$ features reduced to 27 dimensions perform much better when concatenated with the 39 dimensional PNCC vector, than if $GBFB_{PCA}$ features are used by themselves. Though $GBFB_{LE}$ perform very poorly when used as standalone features, they provide complimentary information to PNCC features. Hence, the $GBFB_{LE}$ features are concatenated with PNCC from here on.

Comparing the performance of the concatenated $GBFB_{PCA}$ features (3) with the concatenated $GBFB_{LE}$ features (4), it is observed that the $GBFB_{LE}$ features provide a 2.10% absolute improvement over $GBFB_{PCA}$. Therefore, it is clear that non-linear dimension reduction provides an improved performance over linear dimension reduction methods. This is perhaps due to the fact that the Gabor features lie on a curved low-dimensional manifold. Since PCA fits the data into an ellipsoid while projecting onto a low dimensional flat subspace, it is unable to match the performance of the non-linear technique, Laplacian Eigenmaps.

The ideal dimension of the low dimensional manifold to embed into is not known. Hence, empirical experiments are performed to determine this value. The performance of the $GBFB_{LE}$ features (concatenated with PNCC) is evaluated for different values of the target dimension. It was seen that the performance of our system tends to vary based on the target dimension, which was varied over the set $\{100, 50, 27, 13, 10, 7\}$. The best performance was obtained for a $GBFB_{LE}$ dimension of 7. A similar search wasn't carried out for $GBFB_{PCA}$ because $GBFB_{LE}$ outperformed $GBFB_{PCA}$ at dimension 27. Overall, our new system provides a relative improvement of 12.66% (absolute: 3.79%) in the word error rate (WER) over the PNCC baseline.

5. SUMMARY

In this paper, a novel method based on non-linear dimension reduction technique, Laplacian Eigenmaps, of the PNS-Gabor features with an MLP for out-of-sample extension has been proposed. The $GBFB_{LE}$ feature (reduced to 7 dimensions) concatenated with PNCC (39 dimensions) gave a relative improvement of 12.66% in WER over the PNCC baseline.

In the future, it is worth investigating the effect of the complexity of the recognition task (small, medium or large vocabulary), the parameters used in Laplacian Eigenmaps, and the architecture of the MLP, on the WER. These relationships could be analyzed to give new insights into the structure of the Gabor feature space. It is likely that the mapping learnt using the Aurora 4 database could be applied to other speech recognition databases as well.

6. REFERENCES

[1] Michael Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Proceedings of INTERSPEECH*, 2003, pp. 2573–2576.

[2] Tino Gramss, "Word recognition with the feature finding neural network (FFNN)," in *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, 1991, pp. 289–298.

[3] Marc René Schädler, Bernd T Meyer, and Birger Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, pp. 4134–4151, 2012.

[4] Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proceedings of ICASSP*, 2000, vol. 3, pp. 1635–1638.

[5] Shuo-Yiin Chang, Bernd T Meyer, and Nelson Morgan, "Spectro-temporal features for noise-robust speech recognition using power-law nonlinearity and power-bias subtraction," in *Proceedings of ICASSP*, 2013, vol. 15, pp. 7063–7067.

[6] Sherry Y Zhao and Nelson Morgan, "Multi-stream spectro-temporal features for robust speech recognition," in *Proceedings of INTERSPEECH*, 2008, pp. 898–901.

[7] Suman V Ravuri and Nelson Morgan, "Easy does it: Robust spectro-temporal many-stream ASR without fine tuning streams," in *Proceedings of ICASSP*, 2012, pp. 4309–4312.

[8] Bernd T Meyer, Constantin Spille, Birger Kollmeier, and Nelson Morgan, "Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition," in *Proceedings of INTERSPEECH*, 2012, pp. 1259–1262.

[9] Viren Jain and Lawrence K Saul, "Exploratory analysis and visualization of speech and music by locally linear embedding," in *Proceedings of ICASSP*, 2004, vol. 3, pp. 984–987.

[10] Ayyoob Jafari and Farshad Almasganj, "Using Laplacian eigenmaps latent variable model and manifold learning to improve speech recognition accuracy," *Speech Communication*, vol. 52, no. 9, pp. 725–735, 2010.

[11] Chanwoo Kim and Richard M Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proceedings of ICASSP*, 2012, pp. 4101–4104.

[12] Mikhail Belkin and Partha Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[13] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar, "Sampling techniques for the Nystrom method," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 304–311.

[14] Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland, "The HTK book," *Cambridge University Engineering Department*, vol. 3, pp. 175, 2002.