COMBINING WINDOW PREDICTIONS EFFICIENTLY - A NEW IMPUTATION APPROACH FOR NOISE ROBUST AUTOMATIC SPEECH RECOGNITION

Qun Feng Tan and Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory (SAIL)
University of Southern California
Department of Electrical Engineering
Los Angeles, CA 90089, U.S.A.
qtan@usc.edu, shri@sipi.usc.edu

ABSTRACT

This paper introduces a new optimization-based approach to Sparse Imputation/spectral denoising for robust Automatic Speech Recognition (ASR) applications. In particular, we propose an algorithm which couples frame-level optimization and strategic reconciliation of the predictions in a tight manner. We demonstrate that the proposed algorithm outperforms the current state-of-the-art two-step strategy of first optimizing and then averaging across windows, while maintaining the complexity advantages of efficient techniques like the Elastic Net. Our algorithm is also theoretically able to better exploit the properties of a collinear dictionary, which occurs with spectral exemplars from most speech corpora. Through experiments on the Aurora 2.0 noisy digits database, we demonstrate that this new technique achieves significant performance gains (7.67% on average over various SNR levels) over just simply averaging across large number of predictions.

Index Terms— Optimization, Denoising, Robustness, Automatic Speech Recognition

1. INTRODUCTION

Compressed sensing/Sparse Representation techniques have recently been employed in spectral denoising for speech recognition applications. Prior works [1, 2] have proposed the use of L_1 optimization techniques and demonstrated appreciable gains in speech recognition rates over the original noisy conditions. The L_1 optimization considers an objective function of the following form:

$$\min_{\boldsymbol{a}} \|\mathbf{D}\boldsymbol{a} - \boldsymbol{F}\|_2^2 + \lambda \|\boldsymbol{a}\|_1 \quad (1)$$

In Equation (1), $\boldsymbol{F}$ is the observed feature vector, $\boldsymbol{a}$ is the vector of activations, and $\mathbf{D}$ is the dictionary. This process is known as "Sparse Imputation" [1]. In this paper, we propose a technique to further improve upon the Sparse Imputation process.

2. RELATED WORK

Kanevsky et al. [3] have suggested the use of the Elastic Net algorithm for robust phoneme recognition. In previous work [4], we have also suggested the use of Elastic Net [5] as a solution to better exploit the properties of a dictionary of collinear spectral exemplars in the Missing Data Techniques (MDT) setting, and also provided a study of why sparsity is important in the regularization framework for spectral denoising. In addition, we [6] showed the importance of grouping spectral atoms for improving speech recognition.

When dealing with utterances of different lengths, prior works [1, 4, 6] have used the classical sliding window approach to extract fixed-duration frames for input to an appropriate regularization algorithm for noise removal. Subsequently, to reconcile predictions (i.e. recombine the individual imputed results from the contributing windows) from the analysis frames, an averaging strategy was proposed whereby the predictions are all added up and then averaged by the number of overlapping frames. While such a strategy improves performance compared to when just optimizing on non-overlapping frames, a more robust way of combining the predictions is possible, which we will introduce in this paper.

The main contribution of this paper is to propose an extension to the simple averaging reconciliation approach for spectral denoising. Instead of just simple averaging, we propose an alternative framework which tightly couples the frame-level optimization and the local prediction reconciliation step. Following the evaluation practice set by prior works, we will use the Aurora 2.0 noisy digits database for our denoising experiments. We will demonstrate that the proposed framework yields an appreciable improvement in ASR accuracies.

The structure of the paper is as follows: Section 3 details the framework for speech spectral denoising and also our algorithm formulation. Besides describing our algorithm in detail, we also provide justification as to why this algorithm is well-suited to perform in the setting of spectral exemplars. Section 4 provides a description of our ASR system setup/settings and also provides the results of our experiments with interpretation. Section 5 concludes with possible future work and extensions to our proposed system.

3. FRAMEWORK AND ALGORITHMIC DESCRIPTION

Fig. 1 shows the schematic of the ASR pipeline with a breakdown of the feature extraction module. In this paper, we propose improvement to the regularization block by a more robust reconciliation method.

3.1. Feature Extraction Procedure

At the feature extraction stage in the ASR pipeline, after the spectral features are extracted, we obtain a matrix of features of dimensions $N_B \times T_F$, where N_B stands for the number of frequency bands in the extraction process and T_F, the duration of the utterance in number of analysis frames.

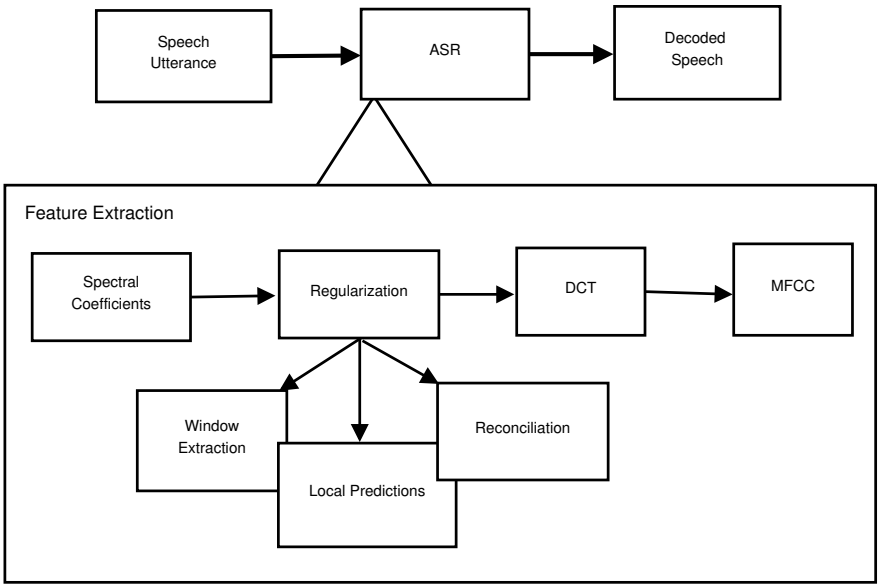

Fig. 1. Schematic of the Spectral Denoiser. The Regularization block can be broadly split into 3 steps: 1) Windowing 2) Regularization of the Windows (local optimization) 3) Reconciliation of predictions

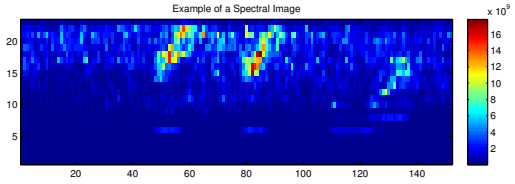

Fig. 2. The diagram above shows the spectral plots of a particular noisy utterance. We can think of this as an image and employ denoising techniques to clean up the noise artifacts evident in the image

3.2. Linear formulation of Problem

If $\boldsymbol{F}$ is a feature vector of spectral exemplars (prior to taking the logarithm), we assume the following linear model for our problem:

$$\boldsymbol{F} = \mathbf{D}\boldsymbol{a} \tag{2}$$

$\mathbf{D}$ is composed of a dictionary of spectral exemplars in this setting.

3.3. Denoising problem formulation

We can treat the problem at hand as an image denoising problem, where the feature values can be likened to the pixel intensities of an image (as in Fig. 2).

The problem formulation is given by the following equation:

$$\min_{\boldsymbol{a}} \|\mathbf{D}\boldsymbol{a} - \boldsymbol{F}\|_2^2 + \lambda \|\boldsymbol{a}\|_1 \tag{3}$$

We will now proceed to describe the procedure by which we obtain $\boldsymbol{F}$.

When $\mathbf{D}$ is comprised of spectral exemplars extracted from speech, $\mathbf{D}$ will have a tendency of being collinear, since spectral images have energy localizations in similar regions for similar sounding utterances. Thus, we [4] demonstrated the need for more robust solutions to this problem, and showed the effectiveness of the Elastic Net formulation:

$$\min_{\boldsymbol{a}} \|\mathbf{D}\boldsymbol{a} - \boldsymbol{F}\|_2^2 + \lambda_1 \|\boldsymbol{a}\|_1 + \lambda_2 \|\boldsymbol{a}\|_2^2 \tag{4}$$

For each utterance, due to different durations, we will have different values of T_F. To ensure that we have matrices of equal dimensions each time we run our optimization, we adopt the strategy of a sliding window matrix extraction. In particular, we define a window of size $N_B \times T$ and shift the window at regular intervals which is predetermined. The vector $\boldsymbol{F}$ is then determined by linearizing the extracted window.

To represent the window extraction process algebraically, let us define a matrix $\mathbf{R}_i$ which is of dimensions $T_F \times T$. i indicates the window count. Thus, by putting Identity matrix blocks and zero blocks in appropriate locations in $\mathbf{R}_i$, we can write the window extraction process as follows:

$$\mathbf{F}_s = \mathbf{F}_o \mathbf{R}_i \tag{5}$$

In Equation (5), $\mathbf{F}_s$ refers to the window extracted subset and $\mathbf{F}_o$ refers to the original feature matrix.

Prior works have attempted to reconcile the predictions by an averaging strategy: at the end of all regularizations for a particular utterance, the predictions are summed appropriately and then divided by the number of times the sliding window overlaps at the particular location. While this has yielded good results, we propose a more tightly coupled way of optimization which further improves upon the averaging framework as shown in a later section (Sec 4).

3.4. Signal Reliability Masks

Before we describe the proposed optimization formulation, we briefly review Signal Reliability Masks since we will be integrating them in our optimization formulation. In this paper, we adopt a simple binary mask [7] which is simply a matrix of dimensions $N_B \times T_F$ consisting of zeros and ones with zero indicating an unreliable component and one indicating a reliable component.

In particular, let us define a matrix $\mathbf{E}$ which extracts the reliable rows of the feature vector $\boldsymbol{F}$ with respect to the reliability indication by the binary mask. Then, the linear model as given in Equation (2) can be rewritten as follows:

$$\boldsymbol{F}_{\text{reliable}} = \mathbf{E}\mathbf{D}\boldsymbol{a} \tag{6}$$

For subsequent notational convenience, we define:

$$\mathbf{D}_{\text{reliable}} = \mathbf{E}\mathbf{D} \tag{7}$$

Equation (4) can be rewritten as:

$$\min_{\boldsymbol{a}} \|\mathbf{D}_{\text{reliable}}\boldsymbol{a} - \boldsymbol{F}_{\text{reliable}}\|_2^2 + \lambda_1 \|\boldsymbol{a}\|_1 + \lambda_2 \|\boldsymbol{a}\|_2^2 \tag{8}$$

3.5. A Novel Formulation of the Optimization Problem

In section 3.3, we drew the analogy that the spectral denoising problem can be visualized as an image denoising problem. Elad et al [8] have proposed an alternative framework for image denoising which deals with the problem of image patches. We will likewise be motivated by that framework and propose a new system which reconciles the individual predictions by optimization.

Let us define a new matrix $\mathbf{G}$ which has the same dimensions as $\mathbf{F}_o$, representing the denoised version of $\mathbf{F}_o$. In addition, denote the number of extracted windows that we have by N_W. A natural generalization of the Maximum A Posteriori (MAP) estimate in this case will be the following optimization:

$$\begin{aligned} \min_{\boldsymbol{a}_{i\in\{1\ldots N_W\}},\mathbf{G}} \quad & \lambda \quad \|\mathbf{G} - \mathbf{F}_o\|_2^2 + \sum_{i=1}^{N_W} \|\mathbf{D}\boldsymbol{a}_i - \mathbf{G}\mathbf{R}_i\|_2^2 \\ & + \quad \lambda_{0i} \|\boldsymbol{a}_i\|_0 \end{aligned} \tag{9}$$

Note that the formulation in Equation (9) is an NP-hard problem (due to the L_0 norm term), and thus a convenient convex relaxation can be formulated as follows:

$$\begin{aligned} \min_{\boldsymbol{a}_{i\in\{1\ldots N_W\}},\mathbf{G}} \quad & \lambda \quad \|\mathbf{G} - \mathbf{F}_o\|_2^2 + \sum_{i=1}^{N_W} \|\mathbf{D}\boldsymbol{a}_i - \mathbf{G}\mathbf{R}_i\|_2^2 \\ & + \quad \lambda_{1i} \|\boldsymbol{a}_i\|_1 \end{aligned} \tag{10}$$

Since Equation (9) is convex, there are a variety of fast solutions. However, to be better able to handle a dictionary $\mathbf{D}$ of collinear spectral exemplars, we hereby propose the following formulation which naturally ties in with the formulation of the Elastic Net:

$$\begin{aligned} \min_{\boldsymbol{a}_{i\in\{1\ldots N_W\}},\mathbf{G}} \quad & \lambda \quad \|\mathbf{G} - \mathbf{F}_o\|_2^2 \\ & + \quad \sum_{i=1}^{N_W} \|\mathbf{D}\boldsymbol{a}_i - \mathbf{G}\mathbf{R}_i\|_2^2 \\ & + \quad \lambda_{1i} \|\boldsymbol{a}_i\|_1 + \lambda_{2i} \|\boldsymbol{a}_i\|_2^2 \end{aligned} \tag{11}$$

To solve the formulation in Equation (11), we decouple the expression into a series of smaller optimization problems (by optimizing on each unknown sequentially). In particular, we employ the Elastic Net algorithm to solve the following series of N_W optimization problems:

$$\min_{\boldsymbol{a}_i} \|\mathbf{D}\boldsymbol{a}_i - \boldsymbol{F}\|_2^2 + \lambda_{1i} \|\boldsymbol{a}_i\|_1 + \lambda_{2i} \|\boldsymbol{a}_i\|_2^2 \tag{12}$$

From the $\boldsymbol{a}_i$'s obtained from the Elastic Net, we can now fix them and proceed to optimize for $\mathbf{G}$. From Equation (11) we can see that we need to further solve the following optimization problem:

$$\min_{\mathbf{G}} \lambda \|\mathbf{G} - \mathbf{F}_o\|_2^2 + \sum_{i=1}^{N_W} \|\mathbf{D}\widehat{\boldsymbol{a}}_i - \mathbf{G}\mathbf{R}_i\|_2^2 \tag{13}$$

For subsequent notational convenience, let us denote $J = \lambda \|\mathbf{G} - \mathbf{F}_o\|_2^2 + \sum_{i=1}^{N_W} \|\mathbf{D}\widehat{\boldsymbol{a}}_i - \mathbf{G}\mathbf{R}_i\|_2^2$.
We can write:

$$\begin{aligned} \|\mathbf{D}\widehat{\boldsymbol{a}}_i - \mathbf{G}\mathbf{R}_i\|_2^2 \quad = \quad & \widehat{\boldsymbol{a}}_i{}^T\mathbf{D}^T\mathbf{D}\widehat{\boldsymbol{a}}_i - \widehat{\boldsymbol{a}}_i{}^T\mathbf{D}^T\mathbf{G}\mathbf{R}_i \\ - \quad & \mathbf{R}_i^T\mathbf{G}^T\mathbf{D}\widehat{\boldsymbol{a}}_i \\ + \quad & \mathbf{R}_i^T\mathbf{G}^T\mathbf{G}\mathbf{R}_i \end{aligned} \tag{14}$$

Hence, taking the partial derivative of J w.r.t $\mathbf{G}$:

$$\frac{\partial J}{\partial \mathbf{G}} = 2\lambda(\mathbf{G} - \mathbf{F}_o) - \sum_{i=1}^{N_W} 2\mathbf{D}\widehat{\boldsymbol{a}}_i\mathbf{R}_i^T + 2\mathbf{G}\mathbf{R}_i\mathbf{R}_i^T \tag{15}$$

Setting the RHS of Equation (15) to be zero gives the following:

$$\widehat{G} = \left(\lambda\mathbf{F}_o + \sum_{i=1}^{N_W} \mathbf{D}\widehat{\boldsymbol{a}}_i\mathbf{R}_i^T\right)\left(\lambda\mathbf{I} + \sum_{i=1}^{N_W} \mathbf{R}_i\mathbf{R}_i^T\right)^{-1} \tag{16}$$

Note that initial inspection of the term $\left(\lambda\mathbf{I} + \sum_{i=1}^{N_W} \mathbf{R}_i\mathbf{R}_i^T\right)^{-1}$ might suggest that such a huge matrix inversion might lead to significant complexity increase, and might not be worth our while. However, note that $\mathbf{R}_i$ is a block extraction matrix, and thus $\mathbf{R}_i\mathbf{R}_i^T$ will essentially be block diagonal. Hence, the expression $\lambda\mathbf{I} + \sum_{i=1}^{N_W} \mathbf{R}_i\mathbf{R}_i^T$ will be block diagonal as well and there are efficient ways to invert such a matrix. Moreover, all the prior steps to estimate $\boldsymbol{a}_i$ in Equation (12) are essentially repeated applications of the Elastic Net, and we see that our approach still retains the complexity advantages of the Elastic Net.

Let us denote the final estimate for a particular window (after reshaping back to the dimensions $N_B \times T$) by $\mathbf{W}_{\text{estimate}}$.

We then proceed to solve the following optimization problem to reconcile all predictions from the individual windows:

$$\min_{\mathbf{G}} \lambda \|\mathbf{G} - \mathbf{F}_o\|_2^2 + \sum_{i=1}^{N_W} \|\mathbf{W}_{\text{estimate}} - \mathbf{G}\mathbf{R}_i\|_2^2 \tag{17}$$

whose solution is given by the following:

$$\widehat{G} = \left(\lambda\mathbf{F}_o + \sum_{i=1}^{N_W} \mathbf{W}_{\text{estimate}}\mathbf{R}_i^T\right)\left(\lambda\mathbf{I} + \sum_{i=1}^{N_W} \mathbf{R}_i\mathbf{R}_i^T\right)^{-1} \tag{18}$$

4. EXPERIMENTAL SETUP AND RESULTS

4.1. Description of Database and Algorithm implementation

For our ASR system, we use 8040 clean training files (containing single and continuous digit utterances) provided in the Aurora 2.0 database training set to train a continuous digit recognizer in HTK [9].

For the continuous digit recognition task, we take a random subset of 4000 digit utterances from Test A, B and C giving us subway, babble, car, exhibition, restaurant, street, airport, train station, subway (MIRS), and street (MIRS) noise.

We train the ASR on MFCCs with the delta and delta-delta coefficients. We use 23 frequency bands ($N_B = 23$), a hamming window size of 25 ms, and a frame shift of 10 ms. For the delta and delta-delta coefficients, we set the respective parameters in HTK to be equal to 2.

All algorithms are implemented in MATLAB.

4.2. Experimental Results

Table 1. *Results for various levels of corruption."CMN" refers to Cepstral Mean Normalization. "EN Averaging" refers to the procedure where the Elastic Net is applied to each window and contributions from the windows are subsequently averaged. "EN Coupled" refers to the new procedure we described where a second optimization formulation is employed to reconcile the predictions. Runtimes are measured in seconds per optimization. Significance testing is done at 95% confidence interval with the difference of proportions test.*

Algorithm	Accuracies (%)	Runtimes	Significant?
SNR 0 dB			
Unimputed	9.63	NA	NA
CMN	27.78	NA	NA
EN Averaging	26.64	0.0158	NA
EN Coupled	40.39	0.0197	Yes
SNR 5 dB			
Unimputed	36.78	NA	NA
CMN	55.91	NA	NA
EN Averaging	64.38	0.0209	NA
EN Coupled	72.57	0.0215	Yes
SNR 10 dB			
Unimputed	61.41	NA	NA
CMN	87.82	NA	NA
EN Averaging	83.85	0.0296	NA
EN Coupled	89.99	0.0364	Yes
SNR 15 dB			
Unimputed	81.99	NA	NA
CMN	95.67	NA	NA
EN Averaging	93.25	0.0409	NA
EN Coupled	95.84	0.0356	Yes

As shown in Table 1, we ran our experiments on a variety of SNR levels, namely SNR 0 dB, 5 dB, 10 dB and 15 dB. We present recognition results of the original noisy signal, the denoised version using Elastic Net and simple averaging, and our newly proposed coupled strategy (in the table as "EN Coupled").

As evident from our results, the proposed strategy performs consistently better than the simple averaging strategy. Moreover, from our experiments, we see that the new strategy has generally comparable runtimes relative to the simple averaging method. As mentioned before, the main latency involved in our proposed method is the time needed to invert the large square matrix $\lambda\mathbf{I}+\sum_{i=1}^{N_W}\mathbf{R}_i\mathbf{R}_i^T$. Due to the fact that this matrix is block diagonal, the inversion is efficient compared to inverting a non block diagonal square matrix, thus contributing to speedups needed to be comparable with simple averaging.

5. CONCLUSION AND FUTURE WORK

We showed that a more tightly coupled optimization that integrates the local optimization per window and the reconciliation step yields improved results in general compared to the commonly adopted simple averaging strategy (7.67% improvement on average in recognition accuracies). Our formulation also retains the complexity savings of the Least Angle Regression implementation of the Elastic Net, while speeding up the execution at the reconciliation step.

An immediate extension to the proposed scheme in this paper will be to incorporate the global structure of the entire spectral image into the optimization process. Currently, the sliding window framework, while to a small extent is already doing so by utilizing the overlapping portions, better targeted strategies could be developed.

6. REFERENCES

[1] J. Gemmeke, L. ten Bosch, L. Boves, and B. Cranen, "Using sparse representations for exemplar based continuous digit recognition," EUSIPCO Glasgow, Scotland, 2009.

[2] B.J. Borgström and A. Alwan, "Missing feature imputation of log-spectral data for noise robust ASR," *Workshop on DSP in Mobile and Vehicular Systems*, 2009.

[3] D. Kanevsky, T.N. Sainath, B. Ramabhadran, and D. Nahamoo, "An analysis of sparseness and regularization in exemplar-based methods for speech classification," in *Proc. Interspeech*, 2010, pp. 2842–2845.

[4] Q.F. Tan, P.G. Georgiou, and S. Narayanan, "Enhanced sparse imputation techniques for a robust speech recognition front-end," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2418 –2429, Nov. 2011.

[5] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[6] Q. F. Tan and S. S. Narayanan, "Novel variations of group sparse regularization techniques with applications to noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1337 –1346, May 2012.

[7] J. Barker, PD Green, and M. Cooke, "Linking auditory scene analysis and robust ASR by missing data techniques," *Proceedings-Institute of Acoustics*, vol. 23, no. 3, pp. 295–308, 2001.

[8] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[9] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book," 2000.